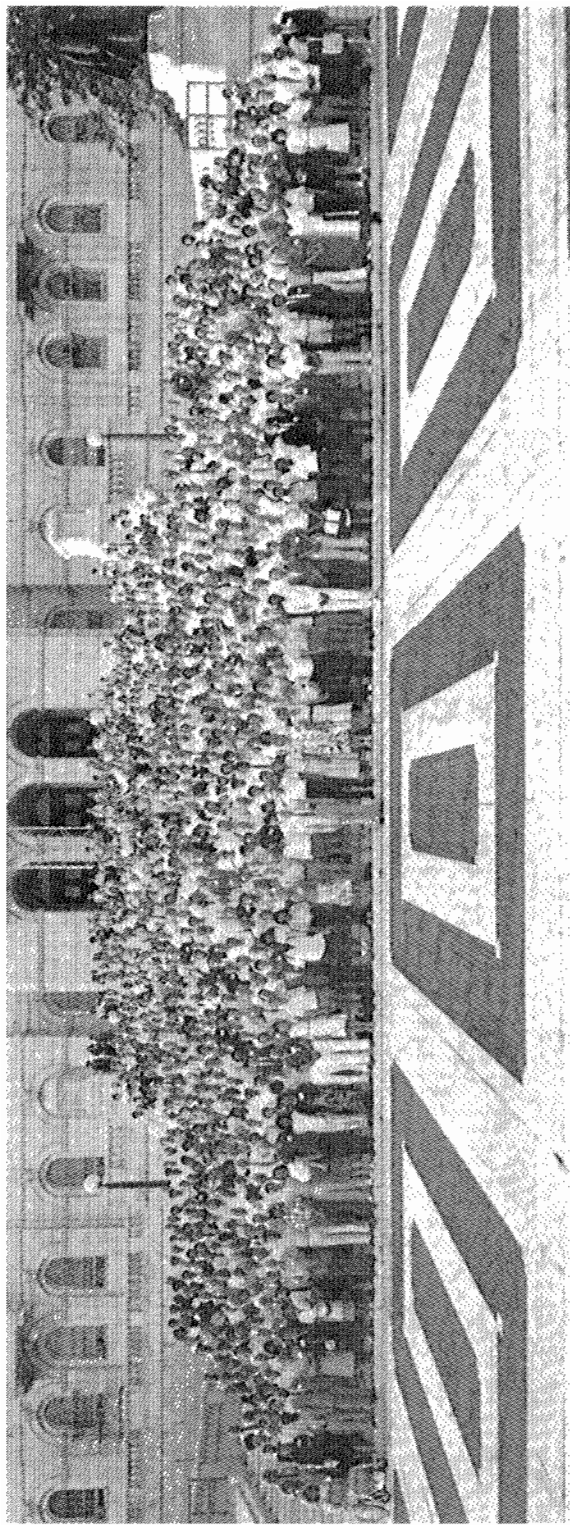


# Mathematics into the Twenty-first Century

1988 Centennial Symposium  
August 8–12



**Statehouse Reception**

AMERICAN MATHEMATICAL SOCIETY  
CENTENNIAL PUBLICATIONS

*Volume II*

**Mathematics into  
the Twenty-first Century**

**1988 Centennial Symposium  
August 8–12**

**Felix E. Browder**  
Editor



American Mathematical Society  
Providence, Rhode Island  
1992

MATHEMATICS INTO THE TWENTY-FIRST CENTURY  
1988 CENTENNIAL SYMPOSIUM  
AUGUST 8-12

1991 *Mathematics Subject Classification*. Primary 00-02, 00A69, 00B10, 00B20.

---

**Library of Congress Cataloging-in-Publication Data**

Mathematics into the twenty-first century; 1988 centennial symposium, August 8-12/[edited]  
by Felix E. Browder.

p. cm. —(American Mathematical Society centennial publications; v. 2)

Includes bibliographical references.

ISBN 0-8218-0167-8

1. Mathematics—Congresses. I. Browder, Felix E. II. American Mathematical Society.

III. Series.

QA1.A52693 1988 vol. 2

510'.6'073 s—dc20

[510]

91-22093

CIP

---

**COPYING AND REPRINTING.** Individual readers of this publication, and nonprofit libraries acting for them, are permitted to make fair use of the material, such as to copy an article for use in teaching or research. Permission is granted to quote brief passages from this publication in reviews, provided the customary acknowledgment of the source is given.

Republication, systematic copying, or multiple reproduction of any material in this publication (including abstracts) is permitted only under license from the American Mathematical Society. Requests for such permission should be addressed to the Manager of Editorial Services, American Mathematical Society, P.O. Box 6248, Providence, Rhode Island 02940-6248.

The appearance of the code on the first page of an article in this book indicates the copyright owner's consent for copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law, provided that the fee of \$1.00 plus \$.25 per page for each copy be paid directly to the Copyright Clearance Center, Inc., 27 Congress Street, Salem, Massachusetts 01970. This consent does not extend to other kinds of copying, such as copying for general distribution, for advertising or promotional purposes, for creating new collective works, or for resale.

Copyright ©1992 by the American Mathematical Society. All rights reserved.

Printed in the United States of America

The American Mathematical Society retains all rights  
except those granted to the United States Government.

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

This publication was typeset using  $\mathcal{A}\mathcal{M}\mathcal{S}\text{-}\mathcal{T}\mathcal{E}\mathcal{X}$ ,  
the American Mathematical Society's  $\mathcal{T}\mathcal{E}\mathcal{X}$  macro system.

10 9 8 7 6 5 4 3 2 1      97 96 95 94 93 92

# Contents

Preface	vii
Introduction	
FELIX E. BROWDER	ix
Symposium Speakers	xi
Representations of finite groups as permutation groups	
MICHAEL ASCHBACHER	1
Regularity of solutions and level surfaces of elliptic equations	
LUIS A. CAFFARELLI	7
Sufficiency as statistical symmetry	
PERSI DIACONIS	15
Atoms and analytic number theory	
C. FEFFERMAN	27
Working and playing with the 2-disk	
MICHAEL H. FREEDMAN	37
The incompleteness phenomena	
HARVEY FRIEDMAN	49
Elliptic curves and modular forms	
BENEDICT H. GROSS	85
Developments in algebraic geometry	
JOE HARRIS	89
A century of Lie theory	
ROGER HOWE	101
From quantum theory to knot theory and back: a von Neumann algebra excursion	
V. F. R. JONES	321
Modular invariance in mathematics and physics	
VICTOR G. KAC	337

<b>Mathematical fluid dynamics: the interaction of nonlinear analysis and modern applied mathematics</b>	
<b>ANDREW J. MAJDA</b>	<b>351</b>
<b>Two examples of mathematics and computing in the biological sciences:</b>	
<b>Blood flow in the heart and molecular dynamics</b>	
<b>CHARLES S. PESKIN</b>	<b>395</b>
<b>Bounds, quadratic differentials, and renormalization conjectures</b>	
<b>DENNIS SULLIVAN</b>	<b>417</b>
<b>Instantons and their relatives</b>	
<b>KAREN UHLENBECK</b>	<b>467</b>
<b>Geometry and quantum field theory</b>	
<b>EDWARD WITTEN</b>	<b>479</b>

## Preface

This volume contains the written versions of sixteen of the original eighteen addresses presented at the American Mathematical Society's Centennial Symposium *Mathematics into the Twenty-first Century* held from August 8–12, 1988. These talks, delivered at the Providence Performing Arts Center, were the principal component of the Scientific Program at the Centennial Celebration. Attendance at this meeting was unprecedented for AMS Summer Meetings with 1,949 members of the Society and a total of 2,502 in attendance including spouses, guests, etc.

The Centennial Celebration was organized by the Centennial Committee with the following members: Felix E. Browder, Rutgers University; Harold M. Edwards, Courant Institute of Mathematical Sciences, New York University; Andrew M. Gleason, Harvard University, a former President of the American Mathematical Society; George Daniel Mostow, Yale University, then current President of the American Mathematical Society; and Everett Pitcher, Chairman, Lehigh University.

The Symposium was organized by the Centennial Program Committee whose members were: Hyman Bass, Columbia University; Felix E. Browder, Chairman; Phillip A. Griffiths, Duke University; John W. Milnor, Institute for Advanced Study; Cathleen S. Morawetz, Courant Institute of Mathematical Sciences, New York University.

Dr. Edward E. David, Jr., Keynote Speaker, delivered the general address entitled *Renewing U.S. Mathematics: an Agenda to Begin the Second Century*. By invitation of the AMS–MAA Joint Program Committee, three retrospective talks were given by Raoul Bott, Peter Lax, and Saunders Mac Lane. The written versions of these talks have been published in *A Century of Mathematics in America*, Parts II and III.

The Centennial Committee thanks the National Science Foundation for its support of the symposium *Mathematics into the Twenty-first Century* (Grant #DMS8716887), and the Department of Energy (Grant #DE-FG02-88ER25056), the Office of Naval Research (Grant #N00014-88-J-1096), and the United States Army Research Office (Grant #DAAL03-88-G-0022) for grants supporting travel and subsistence for young mathematicians attending the Centennial.

## Introduction

The 1988 Symposium *Mathematics into the Twenty-first Century*, of which the present volume is the Proceedings, was organized to celebrate the hundredth anniversary of the founding of the American Mathematical Society (AMS). It was developed on a set of principles which differed in a number of respects from other commemorative celebrations, in particular from the 1938 Semi-centennial Celebration of the AMS. Though the Centennial gave rise to a series of three historical volumes, the Symposium was not historical in character nor did it itself contain speeches of reminiscence or celebration. Three talks of this sort (by Bott, Lax, and Mac Lane) were indeed delivered during the week of the Symposium, but the speakers and their commission were chosen by an entirely separate AMS-MAA organizing committee.

The main principles on which the Symposium was organized can be summarized as follows:

I. The talks should cover as many as possible of the most important central directions of contemporary mathematical research.

II. As far as we could choose, the speakers should be individuals of stature in these directions who have done their principal work in the United States and who are likely to be principal contributors after the year 2000.

III. The central topics of the talks should include not only pure mathematics in its classical forms but also the development of the rapidly developing interaction of sophisticated mathematics with front line areas in science and engineering—in physics, fluid dynamics, computational science, biology, statistics, and computer science.

IV. The talks ought to have an expository intent to make it possible for a broad audience of mathematicians and mathematical practitioners to understand as much as possible of the spirit of what mathematics has accomplished in the last fifty years and of what it hopes to accomplish in the next fifty years.

The Symposium took place with eighteen speakers (culled from an original list of twenty-four) and in the judgment of most people to whom we have spoken, it was an overwhelming success. In the middle of the summer in Providence, 2000 mathematicians came and actually listened to the talks. The speakers tried to be understood and often succeeded. Morale was high, indeed exceptionally so, and the intent of the celebration was vigorously fulfilled.



What we can say of the present volume of Proceedings in terms of the principles outlined above?

Sixteen of the eighteen speakers contributed to the Proceedings, albeit with a lot of coaxing in the process. In the case of Thurston and Tarjan, the two who did not, we got what we would reasonably expect. Thurston is almost a mythical figure in terms of his erratic record in publication, and Tarjan (as he remarked from the beginning) is clearly overcommitted by several hundred percent. There is one shortcoming in the volume that arises from Tarjan's defection: the lack of any contribution involving theoretical computer science and its combinatorial substructure.

The remaining fourteen speakers fulfilled their commitments. In one case, that of Roger Howe, some might say that by writing an "article" which is an expository book in itself on Lie theory and its applications, he has overfulfilled his commitment. My answer to objectors on grounds of uniformity, is: What is the harm? If every speaker had devoted as much effort and mental energy to writing up his talk, we might have used up enormously more paper for this volume (or rather volumes) with an even more useful result. Certainly the expository intent was well fulfilled by Howe's contribution.

The subjects treated in this volume are a reasonable selection of what should be in it. The speakers appeared in alphabetical order and are presented in that order followed by their general areas of study: Michael Asbacher, finite group theory; Luis Caffarelli, nonlinear elliptic partial differential equations; Persi Diaconis, statistics and group invariance; Charles Fefferman, analysis in mathematical physics; Michael Freedman, low-dimensional topology; Harvey Friedman, mathematical logic; Benedict Gross, algebraic number theory; Joseph Harris, algebraic geometry; Roger Howe, Lie theory; Vaughn Jones, knot theory and von Neumann algebras; Victor Kac, Kac-Moody algebras; Andrew Majda, computational fluid mechanics and nonlinear analysis; Charles Peskin, computational methods in biological models; Dennis Sullivan, dynamics and Riemann surfaces; Karen Uhlenbeck, differential geometry, nonlinear elliptic equations and gauge theory; Edward Witten, geometry and quantum field theory.

Aside from the regrettable omission of computer science, the papers presented here form a significant panorama of most of the most vital mathematics of the present epoch. In 1990 Jones and Witten, two of the speakers, received Fields medals at the International Congress of Mathematicians at Kyoto.

We look upon this volume as an obvious symbol of the central role of American mathematics on the world mathematical scene, including the many vital contributors to American mathematics who have come to the U.S. from other lands, especially since the 1930s. The intense vitality of mathematics to which we can all testify owes a great deal to this dynamic synergy between America and the rest of the mathematical world. The American Mathematical Society has been one of the principal agents of this synergy and is one of its most prominent symbols.

## Symposium Speakers



**Michael Aschbacher**

Professor of Mathematics

California Institute of Technology

Ph.D., University of Wisconsin, 1969

2:00 p.m.

Monday, August 8

### **Representations of finite groups as permutation groups**

The classification of the finite simple groups in 1981 changed the landscape of finite group theory and led to an increased effort to describe the structure and representations of the simple groups. Together with the classification, this effort has made possible unexpected applications of finite group theory in other branches of mathematics.

*Introduced by Daniel Gorenstein.*



**Luis A. Caffarelli**

Professor of Mathematics

Institute for Advanced Study

Ph.D., University of Buenos Aires, 1972

3:15 p.m.

Monday, August 8

### **The geometry of solutions to nonlinear problems**

This talk will discuss geometric techniques to study the shape and regularity of solutions to nonlinear elliptic equations and their level surfaces.

*Introduced by Louis Nirenberg.*



**Persi Diaconis**  
Professor of Mathematics  
Harvard University  
Ph.D., Harvard University, 1974

8:30 a.m.  
Tuesday, August 9

### **Sufficiency as statistical symmetry**

To judge what parts of a data set are worth saving, statisticians have developed a useful tool called *sufficiency*, which can be seen as an extension of the invariants of a group. Sufficiency allows a unified construction of statistical models, sheds light on the factorization of generating functions in combinatorics, and provides the underpinnings for recent work in statistical mechanics. This talk will explain the concept of sufficiency and survey these applications.

*Introduced by Gian-Carlo Rota.*



**Charles L. Fefferman**  
Professor of Mathematics  
Princeton University  
Ph.D., Princeton University, 1969

9:45 a.m.  
Tuesday, August 9

### **Problems from mathematical physics**

This talk will cover two problems in mathematical physics. The first is from quantum mechanics and concerns the question of how large numbers of electrons combine with large numbers of protons to form large numbers of atoms. The second is from general relativity and concerns a proof that some small initial disturbance will not concentrate and become a black hole.

*Introduced by Felix E. Browder.*



**Michael H. Freedman**  
Charles Lee Powell Chair Professor  
University of California, San Diego  
Ph.D., Princeton University, 1973

2:00 p.m.  
Tuesday, August 9

### **Working and playing with the two-dimensional disk**

The conformal structure of the disk is useful in studying the topology of (real) surfaces. A more combinatorial-topological study of maps of a disk has illuminated the study of three-dimensional manifolds. This talk will briefly survey the role of the disk in the theory of high-dimensional manifolds, and go on to address the special problems of a disk mapped into a four-dimensional manifold. This is the point at which the topological and smooth theories diverge, and some discussion of the disparities between them will be given.

*Introduced by William Browder, President-Elect of the AMS.*

**Harvey M. Friedman**

Professor of Mathematics

Ohio State University

Ph.D., Massachusetts Institute of Technology, 1967

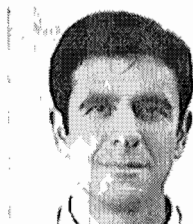
3:15 p.m.

Tuesday, August 9

**The incompleteness phenomena**

By 1922, the formalization of mathematics in terms of axiomatic set theory had emerged. The axioms and rules of inference of this formalism are collectively known as Zermelo-Frankel set theory with the axiom of choice (ZFC). The incompleteness phenomena—assertions which cannot be proved or refuted with ZFC—have not yet necessitated a reassessment of ZFC, but the twenty-first century may see debate on which axioms and rules of inference should be allowed. This talk will provide a historical perspective on the incompleteness phenomena.

*Introduced by Saunders Mac Lane, former President of the AMS.*

**Benedict H. Gross**

Professor of Mathematics

Harvard University

Ph.D., Harvard University, 1978

8:30 a.m.

Wednesday, August 10

**Modular forms and elliptic curves**

This talk will survey some major developments in the theory of elliptic curves. The theory of elliptic functions and modular forms, created in the nineteenth century, concerns the real and complex solutions of cubic equations and their moduli. In the last fifty years, the original arithmetic viewpoint has once again emerged. The problem of counting the number of solutions (mod  $p$ ) to equations with integral coefficients is related to certain Fourier expansions in the classical theory of modular forms. This relationship has led to some progress on the problem of constructing rational points.

*Introduced by John T. Tate.*

**Joseph Harris**

Visiting Scholar in Mathematics

Harvard University

Ph.D., Harvard University, 1977

9:45 a.m.

Wednesday, August 10

**Developments in algebraic geometry**

One of the oldest branches of mathematics, algebraic geometry is concerned with the geometry of curves, surfaces, and higher-dimensional objects defined by polynomial equations—conic sections, quadric surfaces, and so on. Over the last two centuries, algebraic geometry has undergone a series of transformations in which its basic objects of study were redefined, the most recent being the introduction of the concept of “schemes.” This talk will describe these stages in the evolution of the subject and indicate how they arose as outgrowths of classical problems.

*Introduced by Phillip A. Griffiths.*

**Roger E. Howe**

Professor of Mathematics

Yale University

Ph.D., University of California, Berkeley, 1969

2:00 p.m.

Wednesday, August 10

**A century of Lie theory**

The subject called Lie theory (the study of Lie groups, Lie algebras, algebraic groups, and their applications) is, like the AMS, just about one hundred years old. In that century, Lie theory has established itself as a central area of mathematics, using tools from many sources and having implications for many other fields. This talk will attempt to give a feeling for the diversity of applications of Lie theory and for the rich internal structure that supports the applications.

*Introduced by George Mackey.*

**Vaughan F. R. Jones**

Professor of Mathematics

University of California, Berkeley

Ph.D., Université de Genève, Switzerland, 1979

3:15 p.m.

Wednesday, August 10

**A von Neumann algebra excursion: From quantum theory to knot theory and back**

A surprising result in von Neumann algebras suggested representations of the braid group into an abstract algebra discovered in statistical mechanics. The result allows one to associate to each braid a number which turns out to depend only on the knot obtained by closing the braid. The resulting new knot invariant stimulated the discovery of many more such invariants. These invariants are being used to study the way enzymes “untie” knotted strands of DNA in the process of replication.

*Introduced by Joan S. Birman.*

**Victor G. Kac**

Professor of Mathematics

Massachusetts Institute of Technology

Ph.D., Moscow State University, 1968

4:30 p.m.

Wednesday, August 10

**Modular invariance in mathematics and physics**

This talk will focus on some beautiful, recently discovered connections between the representation theory of infinite dimensional Lie algebras and the theory of modular functions, and on related progress in theoretical physics. The basic examples covered will be: affine Kac-Moody algebra, the central extension of the loop group of a compact Lie group; and Virasoro algebra, the central extension of the Lie algebra of vector fields on the circle. The “modular invariant” representations of these algebras have been playing a fundamental role in recent developments of conformally invariant quantum field theories and in string theory.

*Introduced by Nathan Jacobson, former President of the AMS.*



**Andrew J. Majda**  
 Professor of Mathematics  
 Princeton University  
 Ph.D., Stanford University, 1973

8:30 a.m.  
 Thursday, August 11

**Mathematical fluid dynamics: The interaction of nonlinear analysis and modern applied mathematics**

The rapid evolution of applied mathematics through large-scale computation reveals new fluid flow phenomena that are far beyond the capability of experimental measures. To explain and control these complex phenomena, new mathematical ideas from nonlinear analysis, differential equations, probability theory, and geometry must interact with computational methods and more traditional tools of applied mathematics. This talk will present a survey of several examples of this new mode of interdisciplinary research in mathematical fluid mechanics.

*Introduced by Peter D. Lax, former President of the AMS.*



**Charles S. Peskin**  
 Professor of Mathematics  
 Courant Institute of Mathematical Sciences,  
 New York University,  
 Ph.D., Albert Einstein College of Medicine, 1972

9:45 a.m.  
 Thursday, August 11

**Mathematics and computing in physiology and medicine: Examples from the past, present, and future**

The examples considered are the Hodgkin-Huxley equations for the nerve impulse, computed tomography, a mathematical model for blood flow in the heart, and the robotics of large biological molecules. Computation is a key ingredient in all of these examples, and future success is tied to the development of large-scale computers and efficient numerical algorithms.

*Introduced by Cathleen S. Morawetz.*



**Dennis P. Sullivan**  
 Professor of Mathematics  
 Graduate School and University Center,  
 City University of New York,  
 Ph.D., Princeton University, 1966

2:00 p.m.  
 Thursday, August 11

**Progress on the renormalization conjectures in dynamical systems**

Computation has led theoretical physicists to the discovery that, in certain dynamical systems, the geometrical structure at successively smaller scales is asymptotically constant. Moreover, the structure is universal in the sense that inequivalent systems have the same limiting structure. This talk will summarize the progress in the theoretical understanding of this numerical discovery.

*Introduced by Stephen Smale.*


**Robert E. Tarjan**

James S. McDonnell

Distinguished University Professor  
of Computer Science

Princeton University and Distinguished Member  
of Technical Staff AT&T Bell Laboratories

Ph.D., Stanford University, 1972

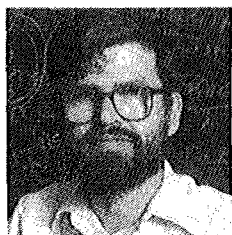
8:30 a.m.

Friday, August 12

**Mathematics in computer science**

This talk will explore the interdependencies between mathematics and computer science as illustrated in the variety of mathematical ideas used to derive results in computer science theory and the use of computation in the proof of mathematical theorems.

*Introduced by Ronald L. Graham.*


**William P. Thurston**

Professor of Mathematics

Princeton University

Ph.D., University of California, Berkeley, 1972

9:45 a.m.

Friday, August 12

**Three-dimensional geometry and topology**

Three dimensions is the crossroad for geometry and topology. In dimensions higher than 3, topology becomes much more arbitrary, while geometry becomes much more restricted and rigid. In dimensions lower than 3, topology is more limited, while geometric constructions are more flexible. This talk will describe several instances of the close match between the geometry and topology of 3-dimensional objects, including the theory of polyhedra, the theory of knots, and the theory of 3-dimensional manifolds.

*Introduced by Lipman Bers, former President of the AMS.*


**Karen K. Uhlenbeck**

Professor of Mathematics

University of Texas at Austin

Ph.D., Brandeis University, 1968

11:00 a.m.

Friday, August 12

**Instantons and their relatives**

Instantons are geometric objects which were discovered by theoretical high energy physicists as a result of failed attempts to understand strong interactions. The instanton equation—of which instantons are solutions—derives from the nonlinear version Maxwell's equations formulated by Yang and Mills in 1954. The importance of the instanton equation in mathematics was recognized only in the past decade. Vortices and monopoles are only two of the many related geometric objects having elegant, interesting, and useful mathematical properties. This talk will attempt to describe some of the more colorful properties and uses of instantons and some conjectures for the future.

*Introduced by Shiing S. Chern.*



**Edward Witten**  
Professor of Physics  
Institute for Advanced Study  
Ph.D., Princeton University, 1976

2:00 p.m.  
Friday, August 12

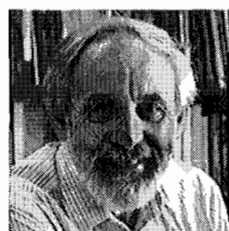
**Quantum field theory and Donaldson polynomials**

When Simon Donaldson initiated a program of using the self-dual Yang-Mills equations to study smooth four-manifolds, the relationship of his work to physical ideas was something of an enigma. Since then, it has become clear that relativistic quantum field theory provides a very natural setting for understanding Donaldson theory and its relationship to Floer theory, elliptic cohomology, conformal field theory, and possibly to other subjects, including string theory and the Jones polynomial. This talk will survey some of these developments.

*Introduced by Clifford Taubes.*

**AMS-MAA Invited Addresses**

By invitation of the AMS-MAA Joint Program Committee, the following speakers will speak on the history and development of mathematics.



**Raoul H. Bott**  
William Caspar Graustein Professor of Mathematics  
Harvard University  
D.Sc., Carnegie Institute of Technology, 1949

11:00 a.m.  
Tuesday, August 9

**The topological constraints on analysis**

This topic has been at the center of one of the two great American schools of topology. Some of its achievements during this century will be discussed.

*Introduced by Andrew M. Gleason.*



**Peter D. Lax**  
Professor of Mathematics  
Ph.D., New York University, 1949

11:00 a.m.  
Wednesday, August 10

**Mathematics: Applied and pure**

In this century, some have viewed mathematics as separated into pure and applied. Today more and more mathematicians realize that mathematics does not "trickle down" to application areas, but is an equal partner with other sciences. Modern computers have linked mathematics with other sciences.

*Introduced by George Daniel Mostow.*



**Saunders Mac Lane**

Professor Emeritus, University of Chicago

Ph.D., University of Göttingen, 1934

11:00 a.m.

Thursday, August 11

**Some major research departments of mathematics**

In the last century, the development of mathematics has been led by a number of outstanding research departments. The tradition was developed in the U.S. by Moore, Birkhoff, Veblen, Stone, and others. This talk will describe several mathematics research departments.

*Introduced by Leonard Gillman, President of the MAA.*

## Representations of Finite Groups as Permutation Groups

MICHAEL ASCHBACHER

In 1860 the Paris Academy offered its Grand Prix des Mathematiques for a contribution to the solution of the following problem:<sup>1</sup>

*For given  $n$ , what are the possible indices  $m$  of subgroups of the symmetric group of degree  $n$ , and given  $m$ , what are the subgroups of index  $m$ ?*

Three manuscripts were submitted to the Academy in the prize competition; the contributors were Kirkman, Jordan, and Mathieu. None of the contributions were judged worthy of the prize.

I believe it is fair to say that there was little significant progress on this problem until about 1955, when dramatic developments in the study of finite simple groups began to make the possibility of a solution more realistic. The classification of the finite simple groups in 1981 and the continued expansion of our knowledge of the finite simple groups themselves have now brought at least a weak solution to the problem within reach.

The effort to solve the problem is one of the current active areas of research in finite group theory and touches most of the other active areas of the subject. I propose to discuss this effort and to use that discussion as a focus for a more general discussion of the major developments in finite group theory of the last few decades and for speculation on the future of the subject.

Let us begin by restating our problem in modern language. A *representation* of a group  $G$  on an object  $X$  is a group homomorphism  $\pi : G \rightarrow \text{Aut}(X)$  of  $G$  into the group of automorphisms or symmetries of  $X$ . Most mathematicians are familiar with *linear representations*, where  $X$  is a vector space over a field. But for finite groups a more basic class of representations are the *permutation representations*, where  $X$  is a set. Thus a permutation representation of  $G$  is a group homomorphism  $\pi : G \rightarrow \text{Sym}(X)$  of  $G$  into the symmetric group on a set  $X$ .

---

1980 *Mathematics Subject Classification* (1985 Revision). Primary 20B05, 20D05.

<sup>1</sup>My (limited) knowledge of the early history of finite groups comes from a set of lectures given by Peter Neumann at Oxford in 1983 and from an expository article by Walter Feit [6].

In the mid-nineteenth century the term “group” meant “permutation group” or “group of transformations”. The notion of an abstract group did not yet exist and hence one could not speak of group representations. Each group came equipped with a permutation representation. However today we can restate our problem in the following form:

*Describe up to equivalence all permutation representations of finite groups.*

When stated in this form we see that we have not really posed the right question. For one thing the problem is not realistic: We cannot hope to completely describe all finite groups, much less their permutation representations. What we can do is decompose our representation and our group into indecomposables and irreducibles and attempt to describe the irreducibles.

A permutation representation  $\pi : G \rightarrow \text{Sym}(X)$  is indecomposable if it is *transitive*; that is, for all  $x, y$  in  $X$  there is a permutation in  $G\pi$  mapping  $x$  to  $y$ . It is a fact that any transitive permutation representation of  $G$  is equivalent to a representation of  $G$  by right multiplication on the set  $G/H$  of cosets of the subgroup  $H$  of  $G$  fixing  $x$  in  $X$ . Thus the study of permutation representations is equivalent to the study of subgroup structure.

A transitive representation is irreducible if it is *primitive*; that is,  $G$  preserves no nontrivial equivalence relation on  $X$ . This is equivalent to requiring that  $H$  be a maximal subgroup of  $G$ . Many problems on finite permutation groups can be reduced to a problem about primitive groups. Thus we are lead to reformulate our problem as follows:

*Determine up to equivalence all injective primitive permutation representations of finite groups.*

I contend this is the right formulation of our problem. It is right because the hypotheses are on the one hand sufficient for most applications and on the other hand restrictive enough to admit a solution, at least in a weak sense, which is strong enough for our applications. For example the theory of groups began in the nineteenth century, where permutation groups were used to study the solutions to polynomial equations. Today the classification of the finite simple groups and our knowledge of the subgroup structure of the simple groups has made possible the solution to problems in areas of mathematics as diverse as model theory, number theory, topology, and combinatorics. Most of these applications arise by reducing the model-theoretic or number-theoretic problem to a problem on primitive permutation groups. A more complete description of primitive permutation representations of the finite groups should lead to even more applications.

The reason that our new problem admits a solution is that most finite groups do not admit an injective primitive permutation representation. Indeed in [2] it is shown that each such group has one of five general structures. The most interesting structure occurs when  $G$  is *almost simple*; that is,  $G$  has a unique minimal normal subgroup  $L$  and  $L$  is a nonabelian simple group. Equivalently,  $\text{Inn}(L) \leq G \leq \text{Aut}(L)$ . Thus we have our final formulation of

our problem:

*Determine up to conjugation the maximal subgroups of each almost simple finite group.*

This formulation focuses attention on the simple groups and their subgroups. Thus I will interrupt our discussion of primitive groups to recall the statement of the Classification, to discuss the simple groups, and to make a few brief remarks about the history of the subject.

**CLASSIFICATION THEOREM.** *Each finite simple group is isomorphic to one of the following:*

- (1) *A group of prime order.*
- (2) *An alternating group.*
- (3) *A group of Lie type.*
- (4) *One of 26 sporadic simple groups.*

Of course there is a unique group of order  $p$  for each prime  $p$ . The alternating group  $\mathcal{A}_n$  of degree  $n$  is the normal subgroup of all even permutations in the symmetric group of degree  $n$ . The groups of Lie type are analogues of the simple Lie groups. Finally we have the twenty-six sporadic groups, which fall into no known naturally defined infinite family.

Roger Howe will be discussing Lie theory in more detail in a later talk in this series. Lie theory plays an important role in the study of finite simple groups. The simple Lie groups were classified by Killing and Cartan in the late nineteenth century; associated to each is a simple Lie algebra. In 1955, Chevalley [3] showed that each simple Lie algebra  $X$  over  $\mathbb{C}$  possesses a Chevalley basis with respect to which the structure constants of  $X$  are integers. Then the basis elements can be exponentiated and reduced modulo  $p$  for each prime  $p$  to produce a *Chevalley group*  $X(F)$  over any field  $F$ . When  $F$  is finite  $X(F)$  is finite and essentially simple. Chevalley's work was extended to produce other groups of Lie type: the *twisted Chevalley groups* analogous to real forms of Lie groups. The Lie theory also gives important information about these groups such as their automorphism groups and certain subgroups. Borel, Ree, Springer, Steinberg, and Tits made important contributions here. The finite simple groups of Lie type are divided into two classes: the *classical groups* and the *exceptional groups*. The classical groups are the special linear group plus the isometry groups of nondegenerate bilinear and hermitian symmetric sesquilinear forms. The exceptional groups correspond to the exceptional simple Lie algebras.

The sporadic groups are fascinating discrete objects. Each group, by the nature of its existence, corresponds to a number of pathological group-theoretic, combinatorial, and number-theoretic phenomena. I will say more about the sporadic groups in a moment.

The appearance of the Chevalley groups and twisted Chevalley groups was one of the important group theoretic events occurring in the midfifties. The other was the beginning of modern local group theory. *Local group theory*

studies a finite group  $G$  via its  $p$ -subgroups and the normalizers in  $G$  of these  $p$ -subgroups. Sylow's Theorem is perhaps the earliest result in local group theory. Philip Hall proved his extended Sylow theorem for solvable groups in 1937 and Brauer introduced his program for characterizing simple groups by the centralizers of involutions in the fifties. However the first spectacular success of the local theory was Thompson's verification of the Frobenius conjecture in his thesis in the late fifties, followed several years later by the verification by Feit and Thompson of the old conjecture of Burnside that groups of odd order are solvable. The local theory was the principal tool used to establish the Classification. While many mathematicians made major contributions to the local theory, I believe it is fair to say that Thompson had the largest role in its creation.

The next major event in finite group theory was the appearance of the sporadic groups. The first five sporadic groups were discovered by Mathieu (remember he made one of the contributions to the 1860 Paris Prize) in the nineteenth century as multiply transitive permutation groups. The next sporadic group was not discovered until 1965 by Janko, using the local theory. After that sporadic groups were discovered at the rate of about two or three a year until Janko also discovered the last of the groups in 1976.

The largest sporadic group (known as the *Monster*) was discovered independently by Fischer and Griess in 1974. There are a number of mysteries involving the Monster. For example it is conjectured [4] that there is a series  $\chi_i$ ,  $1 \leq i \leq \infty$ , of characters (*Thompson series*) of the Monster such that  $1/q + \sum_i \chi_i(1)q^i$  is the elliptic modular function and  $1/q + \sum_i \chi_i(g_p)q^i$  is a generator for the function field of genus 0 of a congruence subgroup for the prime  $p$ , as  $g_p$  ranges over elements of prime order  $p$  in the Monster. Moreover Frankel, Lepowsky, and Meurman [7] have shown that the Monster is a symmetry group of a holomorphic two-dimensional quantum field theory.

In finite group theory, the seventies was the decade of the push toward the Classification. As a new Ph.D. entering the field at the beginning of the seventies, I can vouch for the excitement created by the regular appearance of sporadic groups and the stream of wonderful theorems that appeared at that time. Many finite group theorists participated in the effort, but the most influential figure in the movement, both through his mathematical contributions and his orchestration of the program, was Danny Gorenstein.

I would like to say a few words about the complexity of the proof of the Classification and its implications for mathematics. The existing proof of the Classification is very long (Gorenstein estimates 10,000 pages), complicated, and messy. There are efforts to shorten and clean up the proof, but in the absence of some totally new idea, such efforts will still leave us with a complicated proof. I personally do not believe the proof will ever be simple. For one thing, the existence of the sporadic groups insures that the set of examples is rather complex. The groups of Lie type of small rank over small fields also exhibit sporadic behavior.

Many mathematicians seem to be uncomfortable with complicated proofs and pathological mathematical objects. I feel the sporadic groups are beautiful; without them, finite group theory would be less interesting. I also feel the Classification is a wonderful theorem. In discrete mathematics, assumptions of symmetry provide the structure which distinguishes interesting objects from the mundane and takes the place of the analytic or algebraic structure of classical mathematics. The Classification is a means for compactly encoding this structure. I believe it will come to be viewed as one of the most important results in discrete mathematics and as indispensable. If such a result requires a difficult proof, so be it.

After this long digression on simple groups, it is time to return to our problem. Recall we seek to describe the maximal subgroups of each finite simple group  $G$ . To do so we realize  $G$  as the group of automorphisms of a suitable mathematical object  $X(G)$ . We then seek to prove:

**STRUCTURE THEOREM FOR  $G$ .** *A proper subgroup  $H$  of  $G$  either stabilizes some member of a set  $\mathcal{C}(G)$  of natural structures on  $X(G)$ , or is almost simple and irreducible on  $X(G)$ .*

Such a structure theorem reduces our problem to the study of structures on  $X(G)$  and to the irreducible representation theory of simple groups in the category of  $X(G)$ . Our structures include substructures, coproduct structures, and product structures; I will give an example soon.

If  $G$  is a classical group of Lie type over a field  $F$  then  $X(G)$  is the pair  $(V, f)$ , where  $V$  is an  $FG$ -module,  $f$  is a bilinear or sesquilinear form on  $V$ , and  $G$  is the isometry group of  $f$ . A Structure Theorem exists for  $G$  [1] and it is conjectured that, with a short explicit list of exceptions, if  $H$  is almost simple and absolutely irreducible on  $V$  with the representation writable over no proper subfield of  $F$ , preserving no bilinear form other than  $f$ , and preserving no tensor product structure, then the normalizer in  $G$  of  $H$  is maximal in  $G$ . If this conjecture is established then in a weak sense we have determined the maximal subgroups of  $G$ . To do more would require an enumeration of the irreducible linear representations of finite simple groups over all finite fields.

Extending work of Dynkin on Lie groups [5], Seitz [10] has established the conjecture for algebraic groups and used his theorem to establish the conjecture when  $H$  is of Lie type with the same characteristic as  $G$ .

Clearly the study of the maximal subgroups of the classical groups impinges on another active area of finite group theory: the study of linear representations of finite groups. I do not have time to discuss this activity.

I believe the correct object  $X(G)$  for an exceptional group  $G$  over  $F$  is a minimal dimensional  $FG$ -module together with a three- or four-linear form on  $F$ . This approach has been successful with groups of type  $G_2$  and  $E_6$ , but much work remains to be done.

The maximal subgroups of twenty-three of the twenty-six sporadic groups

have been enumerated. However the treatments are ad hoc and often involve extensive machine calculation, so the situation is not entirely satisfactory.

I will close by considering the alternating group  $G = \mathcal{A}_n$  on a set  $X$  of order  $n$  as an example. We take  $X(G)$  to be  $X$ . Let  $n = |X|$  and  $S = \text{Sym}(X)$ . Except when  $n = 6$ ,  $S$  is  $\text{Aut}(G)$ . We have:

**STRUCTURE THEOREM FOR  $\mathcal{A}_n$  (O'Nan-SCOTT [9]).** *Let  $H$  be a proper subgroup of  $S$ . Then one of the following holds:*

- (1)  $H$  preserves a proper nonempty subset of  $X$ . (Substructure)
- (2)  $H$  preserves a nontrivial partition of  $X$ . (Coproduct structure)
- (3)  $H$  preserves a nontrivial realization of  $X$  as a set product. (Product structure)
- (4)  $H$  preserves an affine space structure on  $X$ .
- (5) The socle of  $H$  is the direct product of  $k$  copies of some nonabelian simple group  $L$  with  $n = |L|^{k-1}$ . (Diagonal structure)
- (6)  $H$  is almost simple and primitive on  $X$ .

Moreover it has been shown that, with known exceptions, the stabilizers of the structures listed in (1)–(5) and the normalizers of primitive almost simple subgroups are indeed maximal [8]. Thus, in a weak sense, we know the maximal subgroups of the alternating and symmetric groups.

As I have tried to show, it seems possible that within this century we will be able to completely describe in a weak sense all primitive finite permutation groups. Our present knowledge of such groups has already been applied effectively in various areas of mathematics. As the theory becomes more complete and as mathematicians become aware of its potential, I believe many more applications will be discovered.

## REFERENCES

1. M. Aschbacher, *On the maximal subgroups of the classical groups*, Invent. Math. **76** (1984), 469–514.
2. M. Aschbacher and L. Scott, *Maximal subgroups of finite groups*, J. Algebra **92** (1985), 44–80.
3. C. Chevalley, *Sur certains groupes simples*, Tôhoku Math. J. **7** (1955), 14–66.
4. J. Conway and S. Norton, *Monstrous moonshine*, Bull. London Math. Soc. **11** (1979), 308–339.
5. E. Dynkin, *Maximal subgroups of the classical groups*, Amer. Math. Soc. Transl. **6** (1957), 245–378.
6. W. Feit, *Theory of finite groups in the twentieth century*, Amer. Math. Heritage: Algebra and Applied Math. **13** (1981), 37–60.
7. I. Frenkel, J. Lepowsky, and A. Meurman, *Vertex operator algebras and the Monster*, Academic Press, San Diego, 1988.
8. P. Kleidman and M. Liebeck, *The subgroup structure of the finite classical groups*, Cambridge Univ. Press, Cambridge, 1990.
9. L. Scott, *Representations in characteristic  $p$* , Proc. Sympos. Pure Math., vol. 37, Amer. Math. Soc., Providence, RI, 1980, pp. 319–331.
10. G. Seitz, *The maximal subgroups of the classical algebraic groups*, Mem. Amer. Math. Soc., no. 365, Amer. Math. Soc., Providence, RI, 1987, pp. 1–286.

## Regularity of Solutions and Level Surfaces of Elliptic Equations

LUIS A. CAFFARELLI

In many instances, the regularity theory of solutions to second-order equations may be thought of as a stability question; that is, as how a perturbation propagates along a solution surface.

For instance, it is well known that if one slightly perturbs a solution of the wave equation, for instance by changing the data in part of the boundary, the perturbation propagates only on certain directions or regions and therefore one may not expect local regularization effects. That is, regularity, in the few instances in which it can be proven, has to come from somewhere else, i.e., from a regular data.

On the other hand, for uniformly elliptic linear equations, small perturbations propagate all over the surface, in fact in a quantitative fashion, and that implies regularity and stability of such surfaces.

I would like to discuss today a series of nonlinear problems where degeneracies and discontinuities make the question of regularity of solutions and level surfaces (and this associated idea of propagation of perturbations) a very challenging one.

**Elliptic equations and interior regularity.** We start by discussing the notion of ellipticity, and the circle of ideas surrounding Harnack type inequalities.

In the nonvariational context one may loosely say that a continuous function  $u$  or surface is an elliptic equation if one may control the smallest (more negative) eigenvalue of its Hessian  $D_y u$  by its largest one. For instance, in the sense that

$$|\lambda_{\min}| \leq F(\lambda_{\max}, x).$$

Of course, a continuous function has not necessarily a Hessian but one may avoid that by using the “viscosity method, i.e., requiring that such control exists for any  $C^2$  function  $\varphi$ , whose graph manages to touch” the graph of  $u$  by above at  $x_0$ .



Of course,  $u$  is a solution if both  $u$  and  $-u$  are subsolutions.

A remarkable theorem of Krylov-Safonov (the Harnack inequality) states that if the equation is "uniformly elliptic with a right-hand side in  $L^n$ ," i.e.,

$$f(x) = \sup_{\lambda > 0} \frac{F(\lambda, x)}{\lambda + 1}$$

belongs to  $L^n$ , then, for any ball  $B_{2R}$  in the domain of definition of  $u$ ,  $\sup_{B_R} u \leq C \inf_{B_R} u + R^2 (f_{B_{2R}} |f|^n)^{1/n}$ .

This is a very powerful theorem that implies, for instance, the Holder continuity of  $u$  at a point given the appropriate controlled growth of  $\int f^n$ .

In fact, this theorem is in turn very much inspired in its character and proof by DeGiorgi's work on the regularity of solutions of variational problems which is one of the great papers on partial differential equations. There, the ellipticity condition is given in "energy" terms, i.e., instead of considering functions  $u$  for which the eigenvalues of the Hessian are somewhat comparable, one looks at functions  $u$  whose energy is locally under control for the function and its truncations, i.e., for all  $\lambda$ ,

$$\int (\nabla(u - \lambda)^+)^2 \leq C R^{-2} \int [(u - \lambda)^+]^2 B_R + R^{\alpha-2}.$$

One may wonder at this point what is the relation between the first and second family of functions.

In the first case one is trying to say that at each point  $x$

$$|\lambda_{\min}| \leq C(\lambda_{\max} + f(x)) \quad \text{and} \quad \lambda_{\max} \leq C(\lambda_{\min} + f(x)),$$

or fixing coordinates

$$|a_{ij}(x) D_{ij} u| \leq C f(x)$$

for  $a_{ij}(x)$  a positive definite matrix changing discontinuously from point to point.

In the second case, one is saying that

$$|D_i(a_{ij} D_j u)| \leq C f(x)$$

with  $f$  of controlled growth.

The power of these regularity results can be understood when one applies them to the study of nonlinear equations of respectively nondivergence or divergence type, i.e., equations of the form

$$F(D^2 u) = 0 \quad \text{or} \quad D_i(F_i(\nabla u)) = 0.$$

In the first case, we say that the relation is elliptic if  $F(M)$  is monotone in the space of symmetric matrices, and strictly elliptic if we have the further quantitative estimate (for  $N$  positive definite)

$$F(M) + C_1 \|N\| \leq F(M + N) \leq F(M) + C_2 \|N\|.$$

In the second case, if the vector field is coercitive, then

$$\langle \vec{F}(p) - \vec{F}(q), p - q \rangle \geq 0,$$

or if it is strictly coercitive, then

$$C_1 \|p - q\|^2 \geq \langle \vec{F}(p) - \vec{F}(q), p - q \rangle \geq C_2 \|p - q\|^2.$$

If one is allowed to derivate  $u$ , both definitions correspond to classical nonlinear equations  $F^{ij} D_{ij}(u) = 0$ , where, in the first case  $F^{ij} = F^{ij}(D^2 u)$  and in the second case  $F^{ij} = F^{ij}(Du)$  are strictly positive definite bounded matrices.

Here, the DeGiorgi and Krylov theorems become interesting when applied to first derivatives of the functions under consideration.

Indeed, the definitions of both divergence and nondivergence nonlinear equations embody a comparison principle, i.e., two solutions,  $u_1$  and  $u_2$ , of the equation  $F(D^2 u) = 0$  cannot "touch," i.e., if  $u_1 \leq u_2$  for  $X \neq X_0$  and  $u_1(X_0) = u_2(X_0)$ , then at such a point  $D_{ij} u_1 < D_{ij} u_2$ , contradicting the strict monotony of  $F$ .

Of course, this is not entirely correct, but it is so if for instance

$$F(D^2 u_1) = 0, \quad F(D^2 u_2) \leq -\varepsilon$$

for some positive  $\varepsilon$  (so one may perform the old trick of looking at  $u_2 - \varepsilon|X|^2$ ).

Now, it has been noted many times that a comparison principle for solutions of a translation invariant operator is related to a maximum principle for the first derivatives.

In our example of  $F(D^2 u)$ ,

$$u_{\lambda, h} = u(X + he) + \lambda h$$

is again a solution of  $F(D^2 u) = 0$  and therefore if  $u_{\lambda_0, h} \geq u$  along the boundary of  $B_1$ , a ball in the domain of definition of both functions, then  $u_{\lambda_0, h} \geq u$  in the interior of  $B_1$ .

Indeed, this is true for  $\lambda$  very large, and the comparison principle tells us that there is no "first  $\lambda$ " ( $\lambda > \lambda_0$ ) for which  $u_{\lambda, h}$  may touch  $u$ .

It follows that the supremum of the incremental quotient

$$\frac{-u(X + he) + u(X)}{h} = \Delta_h u$$

is attained at the boundary of  $B_1$ . One may think of this as the fact that

$$F(D^2 u_{\lambda, h}) = F(D^2 u) = F^{ij}(D^2 u(\xi)) D_{ij}(\Delta_h u).$$

That is,  $\Delta_h u$  satisfies an elliptic equation with discontinuous coefficients.

But then Krylov's theorem says much more. It not only says that  $\Delta_{\lambda, h} u$  is positive but that it is comparable at any two points of any smaller ball  $B_{1/2} \ll B_1$ .

That is, Harnack's inequality is a very strong, *quantitative form* of the maximum principle.

It tells us not only that the solution surfaces  $u$  and  $u_{\lambda,h}$  separate, but that they do so *uniformly* and hence one can further translate  $u$  (to  $u_{\lambda,h+\dots}$ ) and still this translation will remain above the graph of  $u$ .

An iteration of this argument implies the Holder continuity of the first derivatives of  $u$ , i.e., bounded weak solutions of  $F(D^2u) = 0$  are locally  $C^{1,\alpha}$ .

Let me stress the perturbation view of this result. We have "modified" the boundary data of  $u$  (to those of  $u_{\lambda,h}$ , slightly larger), and this perturbation propagated all over the domain in some uniform way. Since the operator under consideration is translation invariant, this implies  $C^{1,\alpha}$  regularity of solutions.

If one wants to push this idea further, to second derivatives, a structural condition (concavity of  $F$ ) is necessary to ensure that pure second-order incremental quotients are formally subsolutions of the "linearized" differential equations. Then (Evans, Krylov) one combines the fact that  $D^2u$  lies in a Lipschitz, elliptic hypersurface with this fact to control its oscillation.

One may view this approach geometrically the following way: If  $F$  is concave and (possibly degenerate) elliptic, and one envelopes the solution surface  $u$  by above by paraboloids of fixed quadratic part (or spheres of fixed radius), then the new surface  $\bar{u}$  is a subsolution of the same equation.

Hence, if  $D_{ij}u$  are bounded above at the boundary, then  $\bar{u} \equiv u$  for a narrow enough choice of paraboloids, i.e.,  $D_{ij}$  are bounded above in the interior. This is the maximum principle part of the argument.

The Harnack inequality may then be thought of as taking envelopes of the variable quadratic part, so as to improve control of the second derivatives. We will come back to this point later.

**Free boundary problems and harmonic analysis in Lipschitz domains.** Let us now look at problems where the solution completely degenerates past a certain value of  $u$ . For instance, the simplest example is that of minimizers of

$$J(u) = \int (\nabla u)^2 + X_{u>0} dx.$$

Such a minimizer is harmonic when positive, or negative, i.e.,  $F(D^2u) = 0$  with  $F = \text{Trace}$ , and therefore perturbations propagate "elliptically" in regions where  $u$  keeps a "strict sign."

But in view of the previous discussion, the interesting phenomena to study is how a perturbation crosses the surface of discontinuity (for  $\nabla u$ )  $\{u = 0\}$ , i.e., how would a perturbation of order  $\varepsilon$  displace this surface? Would the new surface  $\{u_\varepsilon = 0\}$  separate uniformly from  $\{u = 0\}$  in the interior of the domain of definition? If so, does this "ellipticity" property of free boundaries imply its regularity?

In thinking about such a problem, we may naturally divide it into two parts.

The first part asks: How does this perturbation reach the boundary? At first the sets  $\{u > 0\}$  and  $\{u < 0\}$  are completely amorphous. Even for a  $C^\infty$  function  $u$  there is not much you can say about how narrow or cuspidal a level set may become—the most elementary geometric obstructions one may find for our perturbation to effectively reach the free boundary.

At this point a beautiful link between the basic geometric properties of minimizers to these variational problems and the theory of harmonic measure in Lipschitz domains occurs.

In terms of perturbations of solutions, this theory says that if you have locally a domain that is, say, the intersection of a Lipschitz surface  $S$  (or more generally, a surface with a Harnack chain property (Jerison and Kenig)), then if you have a function  $u$ , harmonic and nonnegative, vanishing on  $S$ , and you perturb it, this perturbation arrives to the boundary in full. That is, if we have two harmonic functions and  $u \leq u_\varepsilon$ ,  $u_\varepsilon(X_0)/u(X_0) \geq 1 + \varepsilon$ , then  $u_\varepsilon/u \geq 1 + C\varepsilon$  uniformly along (any compact subset of)  $S$ . That is,  $(u_\varepsilon)_\nu \geq (1 + (\varepsilon))u_\nu$ .

From the free boundary context, the nondegeneracy properties of  $u^+$  and a curious monotonicity formula (Alt, Friedman, and myself) allow you to assert that the above domain satisfies exactly the Harnack chain condition.

Since the variational term  $\chi_{u>0}$  translates into a jump relation between  $u_\nu^+$  and  $u_\nu^-$ , this makes  $u_\varepsilon$  a strict subsolution of the free boundary problem.

The second part of the problem answers the question: How is this perturbation, whose influence is felt fully along the free boundary, forcing it to drift away.

That is, we are thinking of the perturbation as occurring in two steps. First we lift  $u^+$  somewhere, but force  $S$  to stay fixed, and then we let  $S$  drift to  $S_\varepsilon$ , so that the energy attains equilibrium.

Since the free boundary still has almost no shape, it appears very difficult to construct such a perturbation. (As a parallel, a general strict comparison theorem for generalized minimal surfaces, due to L. Simon, is recent and delicate.)

Here we return to the question of variable parallel surface perturbation, to which we hinted at the end of the previous section.

Given a solution  $u$  to a general translation invariant equation

$$F(D^2u, Du, u) = 0,$$

then

$$u^h(X) = \sup_{B_h(X)} u$$

is a subsolution to the same equation, and it is (heuristically) correct that  $u^h$  is also a subsolution to an “elliptic” free boundary jump condition since it increases  $u_\nu^+$  and decreases  $u_\nu^-$ .

Therefore if  $u, v$  are solutions of a free boundary problem

$$F(D^2u, Du, u) = 0 \quad \text{for } |u| > 0$$

and  $u_\nu^+ = G(u_\nu^-, \nu)$  with  $F$  elliptic (monotone in  $D^2u$ ),  $G$  elliptic (monotone in  $u_\nu^-$ ), and  $u^h \leq v$  on  $\partial B_1$ , then  $u^h \leq v$  all over  $B_1$ .

But this is only a maximum principle, with no quantitative separation among  $u^h$  and  $v$ .

Suppose further that  $(1+\varepsilon)u^h \leq v$  somewhere away from the free boundary. Can we now assert that the surfaces  $\{u^h = 0\}$  and  $\{v = 0\}$  are  $\varepsilon$ -away?

The answer is (in very loose terms) yes, provided that  $F$  has a Harnack inequality "up to boundary," and  $G$  is strictly monotone. This is done by what we could call variable level surface perturbations; that is, defining  $u^\varphi = \sup_{\varphi(X)} u$  and asking when it is true that  $u^\varphi$  is again a subsolution of  $F(D^2u, Du, u)$ .

If  $F$  is uniformly elliptic with a Harnack inequality, one can see that it is enough for  $\varphi$  to satisfy an inequality of the type ( $L$  a Pucci extremal operator)

$$\varphi L\varphi > C|\nabla\varphi|^2$$

for  $u^\varphi$  to be a solution on the region  $u > 0$ .

This allows us to choose a  $\varphi = h$  near  $\partial\Omega$  (where we have only our original information) and a  $\varphi > h$  (where we know that  $u_h$  is strictly less than  $(1-\varepsilon)v$ ), and solve the inequality

$$\varphi L\varphi \geq C|\nabla|^2$$

in between, allowing the perturbation to travel *across* the free boundary. ( $\varphi$  variable distorts the free boundary relation, and  $u^\varphi$  has to be corrected using the "up to the boundary" Harnack inequality.)

**The common setting.** What is, then, the common setting for these problems? It is easy to approximate free boundary problems and problems of generalized surfaces of prescribed curvature relations by one-parameter families of solutions of operators,

$$F_\lambda(D^2u, Du, u) = 0,$$

that degenerate along a level surface  $u = 0$ . (For instance, solutions of  $\Delta u = \beta_\varepsilon(u)$  with  $\beta_\varepsilon$  properly chosen, converge for  $\varepsilon$  going to zero to solutions of the obstacle problem, i.e., minimizers of  $\int(\nabla u)^2 + u^+$ , or the cavitation flow problem, i.e., minimizers of

$$\int(\nabla u)^2 + \chi_{u>0},$$

or of sets of minimal perimeter, i.e., characteristic functions  $\chi_\Omega$  that locally minimize " $\int|\nabla\chi_\Omega|dx$ ".

Further, the operation  $u^h = \sup_{B_h} u$  constructs a new function  $u^h$ , whose level surfaces are parallel surfaces to those of  $u$ , i.e., the level surface  $u_h = t$  is the surface of those points in  $\{u < t\}$  whose distance to  $\{u = t\}$  is exactly  $h$ , i.e., is the  $h$ -level surface of the distance function to  $\{u = t\}$ , and it is well known that the tangential Hessian increases along level surfaces of the distance function. In fact, it does so dramatically (recall the formula  $x_i^h = x_i/(1 - x_i h)$  for the curvatures of the level surfaces of the distance function) if the curvatures of the original surface are large. Can we then, by looking at variable supremums  $u^\varphi$  or, what is related, looking at variable normal perturbations  $d(X, S) = g(X)$ , study how a smoothing effect propagates uniformly along level surfaces of  $u_\lambda$  (solution of  $F_\lambda$ ) independently of  $\lambda$ ?

Is it possible to infer regularity for level surfaces of  $F_\lambda$  independently of  $\lambda$ ?

How elliptic (or hyperbolic) is a problem in  $\lambda$ ; which type of perturbations travel in which direction?

How does a transient problem behave: Do perturbations travel fully in finite time to a free boundary. And many other questions related to Liouville type problems of elliptic or parabolic equations that would answer, after appropriate scaling, the fine structure of free boundaries, surfaces of prescribed curvature relations, and conservation laws.

To close, many efforts are under way that seem to indicate that there is indeed some substance behind these general comments.

SCHOOL OF MATHEMATICS, INSTITUTE FOR ADVANCED STUDY, PRINCETON, NEW JERSEY 08540

# Sufficiency as Statistical Symmetry

PERSI DIACONIS

**Abstract.** Sufficiency is a theoretical tool that has grown up in mathematical statistics. It may be described crudely as the theory of how much data can be thrown away. This paper reviews the basic achievements of the theory in statistical problems and sketches applications in other areas of mathematics. It is shown how the idea gives a suitable framework for exchangeability (an important piece of the Bayesian theory of statistics) and Gibbs states (the rigorous theory of phase transitions in statistical mechanics). In these last settings, sufficiency may be seen as a sweeping generalization of group invariance.

**1. Introduction to sufficiency.** One of the basic problems of statistics is this: one begins with a space  $\mathcal{X}$  and a family of probability measures  $\mathcal{P}$  on  $\mathcal{X}$ . It is assumed that an observation  $x \in \mathcal{X}$  is drawn from a fixed, unknown  $P \in \mathcal{P}$ . We are shown  $x$  and required to guess  $P$ . For example, the usual formulation for  $n$  flips of a coin takes  $x$  as the space of binary  $n$ -tuples. For each  $\theta \in [0, 1]$ , a probability  $P_\theta$  is defined on  $x$  by  $P_\theta(x) = \theta^t(1 - \theta)^{n-t}$  where  $t = t(x) = x_1 + \cdots + x_n$ . The family  $\mathcal{P}$  is taken as  $\{P_\theta\}_{\theta \in [0, 1]}$ . We are shown  $x$  and required to guess  $\theta$ .

In the example, the observation consists of the binary  $n$ -tuple  $x$ . It is natural to ask if all of this is required or if  $x$  can be compressed to  $t = x_1 + \cdots + x_n$  without essential loss. This is the subject matter of sufficiency.

In the general set-up a function  $T : \mathcal{X} \rightarrow \mathcal{Y}$  is called *sufficient for the family  $\mathcal{P}$*  if the conditional probability

$$(1.1) \quad P(x|T(x) = t)$$

is the same for each  $P \in \mathcal{P}$ . In (1.1) the definition of conditional probability is the natural extension of the elementary notion  $P(A|B) = P(A \cap B)/P(B)$ . Thus,  $P(x|T(x) = t)$  is defined as zero unless  $T(x) = t$ . It is taken as proportional to  $P(x)$  if  $T(x) = t$  with normalizing constant making it a probability distribution.

This leaves aside technical fine points which can be found in any standard graduate text in probability (e.g., Billingsley [5]).

EXAMPLE 1: COIN TOSSING. For coin tossing, the sum  $T(x) = x_1 + \cdots + x_n$  is a sufficient statistic. Indeed,

$$P_\theta\{x|T(x) = t\} = \frac{P_\theta\{x \text{ and } T(x) = t\}}{P_\theta\{T(x) = t\}} = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}.$$

The right side does not depend on  $\theta$ . This can also be seen from the following symmetry argument:  $P_\theta(x|T(x) = t)$  is the chance of observing the sequence  $x = (x_1 \cdots x_n)$  given  $T(x) = t$ . Imagine someone flipping a weird, biased coin. They announce that there have been two heads out of the first ten tosses. Whatever the bias, those two heads are equally likely to have appeared in any of  $\binom{10}{2}$  possible places.

Here is a different interpretation of sufficiency for coin tossing as a fact about symmetric functions. Let  $e_i(x_1, x_2, \dots, x_n)$  be the  $i$ th elementary symmetric function in variables  $x_1, x_2, \dots, x_n$ . Thus  $e_1 = \sum x_i$ ,  $e_2 = \sum_{i < j} x_i x_j$ , etc. The generating function for  $e_i$  is

$$\sum_{i=0}^n e_i t^i = \prod_{i=1}^n (1 + x_i t).$$

The factorization of this generating function is equivalent to the sum being sufficient for coin tossing. To see this, divide both sides of the identity above by  $(1 + \theta)^n$ , and multiply and divide  $e_i$  by  $\binom{n}{i}$ :

$$\sum_{i=0}^n \frac{1}{\binom{n}{i}} e_i \binom{n}{i} \frac{\theta^i}{(1 + \theta)^n} = \prod_{i=1}^n \frac{(1 + x_i \theta)}{(1 + \theta)}.$$

On the right is the generating function for  $n$  flips of a coin with probability of heads  $\theta/(1 + \theta)$ . On the left,  $\binom{n}{i}\theta^i/(1 + \theta)^n$  is the chance that  $n$  flips of such a coin lead to  $i$  heads. The term  $e_i/\binom{n}{i}$  is the generating function for  $n$  flips given that  $i$  of them are heads. In the language of random variables the identity appears

$$E_t \prod x_j^{x_j} = EE \left( \prod x_j^{x_j} \mid \sum x_j = t \right).$$

The inner expectation is free of  $\theta$  because  $\sum x_j$  is sufficient for  $\theta$ .

Many of the identities of symmetric function theory can be put into similar language. There is much of interest to do in fitting Schur functions into this picture. See, e.g., Macdonald [39].

Often, sufficiency is clear via symmetry. The point is that the notion is useful without an underlying group. As an example, consider  $n$  binary outcomes in which the chance of 1 increases over time. If the chance of a 1 in place  $i$  is taken as  $e^{\eta_i}/(1 + e^{\eta_i})$  with  $\eta \in [0, \infty)$ , this gives a family of probabilities  $\mathcal{P} = \{p_\eta\}_{\eta \in [0, \infty)}$  on binary  $n$ -tuples. The statistic  $T(x) = \sum_{i=1}^n i x_i$  is easily seen to be sufficient for  $\mathcal{P}$ .



The next example shows sufficiency in a continuous setting.

EXAMPLE 2. Take  $X = \mathbb{R}^n$  and  $\mathcal{P}$  the family of all probability measures on  $\mathbb{R}^n$  invariant under the orthogonal group  $O_n$ . Thus  $P \in \mathcal{P}$  satisfies

$$P(A) = P(\Gamma A)$$

for every Borel set  $A$  and orthogonal matrix  $\Gamma$ .

The sum of squares  $T(x) = x_1^2 + \cdots + x_n^2$  is sufficient for  $\mathcal{P}$ . Indeed  $P\{x|T(x) = t\}$  is uniform on the sphere of radius  $\sqrt{t}$  for every  $P \in \mathcal{P}$ . This example will reappear several times in later sections. The final example shows sufficiency in a less standard setting.

EXAMPLE 3: CONVEX SETS. Let  $\mathcal{C}$  be the class of compact convex subsets in  $\mathbb{R}^d$ . For  $c \in \mathcal{C}$ , define a probability  $P_c$  as the uniform measure inside  $c$ . Define  $P_c^n$  as  $n$ -fold product measure. Take

$$x = \mathbb{R}^{nd}, \quad \mathcal{P} = \{P_c^n\}_{c \in \mathcal{C}}.$$

This is a mathematical model for: "pick  $n$  points at random from inside an unknown convex, compact subset." This problem arises in estimating volumes of convex polyhedra. See, e.g., Deyer, Freize, and Kannen [10]. It is natural to ask what aspects of the data  $x_1 \cdots x_n$  are required to learn about  $c$ . It is not hard to see that only the extreme points  $T(x)$  of the convex hull are required. Indeed, given  $T(x)$ , the rest of the data is uniformly distributed inside the convex hull, no matter what convex set  $c$  underlies the selection process. It follows that  $T(x)$  is sufficient.

The next section reviews the history and main mathematical results of sufficiency. Section 3 introduces exchangeability as part of the Bayesian view of statistics. Section 4 shows how sufficiency ideas give a natural foundation for exchangeability, allowing a theory where there is no natural symmetry. The final section contains pointers to open problems and related subjects.

**2. Basic results of sufficiency.** Sufficiency began, as with so much else in mathematical statistics, with a paper of R. A. Fisher [18]. Fisher was comparing two different estimates for the scale parameter of the normal curve. The estimators were appropriate multiples of

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n |X_i - \bar{X}|.$$

Here the observation consists of  $X = (x_1, \dots, x_n)$  and  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ . Fisher showed that the first estimator was 15% more accurate and indeed that any estimate based on the sum of the absolute deviations would lose some of the information in the full observation. Fisher's argument introduced the ideas of sufficiency which were evident due to the invariance of the normal distribution under the orthogonal group. Later, Fisher [19] abstracted the idea away from invariance and outlined a general theory. This

history is discussed by Stigler [48] who also reports how an earlier giant, Laplace, missed the idea of sufficiency in his work on a very similar problem.

Fisher and Jerzy Neyman [42] developed techniques for finding sufficient statistics and quantifying in what sense a sufficient statistic contains all of the information in a sample. Basically, given  $T(x) = t$ , with no knowledge of which  $P \in \mathcal{P}$  generated  $x$ , a new observation  $x^*$  distributed just like the original  $x$  can be created by independent randomization. Of course, the distribution of  $T$  depends on the underlying  $P$ , but that is all.

Another sense in which a sufficient statistic captures the information is given by the Rao-Blackwell theorem. This considers an estimator  $\hat{P}(x)$  of the measure  $P$ . If  $\hat{P}(x)$  does not depend on  $x$  through a sufficient statistic, then a more accurate estimator can be found, no matter what notion of accuracy is being used. This necessarily vague statement is made precise in any of the standard graduate texts on mathematical statistics of which Lehmann [38] is recommended.

Modern work on the mathematics of sufficiency began with Halmos-Savage [25] and Bahadur [2]. They developed a rigorous general framework using  $\sigma$ -algebras and the Radon-Nikodým theorem. They began the love affair that mathematical statistics has had with refined measure theory. This continues to the present day.

Group theory was also being employed to reduce the dimensionality of statistical problems. If a problem is invariant under a group, the data can be reduced to a so-called maximal invariant (a report of which orbit of the group contains the data point). It might also be possible to reduce by sufficiency and the question of whether these reduction operations commute is natural. Charles Stein gave natural conditions for commutation which were expanded in Hall, Wijsman, and Ghosh [24].

Sufficient statistics arise easily in connection with so-called exponential families of measures. These have densities proportional to  $e^{\theta T(x)}$  with respect to a dominating measure which does not depend on  $\theta$ . For such a family, given a sample of size  $n$ ,  $T(x_1) + T(x_2) + \cdots + T(x_n)$  is a sufficient statistic. Conversely, if a family of measures admits a lower-dimensional sufficient statistic B. O. Koopman, E. J. G. Pitman, and G. Darrois gave conditions under which the family is exponential. To appreciate the problem, consider  $\mathcal{P}$  as the set of all measures on  $\mathbb{R} \times \mathbb{R}$ . There are 1-1 continuous functions from  $\mathbb{R} \times \mathbb{R}$  into  $\mathbb{R}$ . Any of these gives a sufficient statistic for  $P$ , which is not any sort of exponential form. To rule out such behavior, some notion of smoothness must be assumed. The best modern version due to Hipp [26] proves a theorem assuming  $T$  is locally Lipschitz.

Exponential families constitute convenient families which include most of the classically studied examples. A unified theory is summarized in Lehman [37, 38], Barndorff-Neilson [3], or Johanson [28].

Exponential families are quite a restricted family of measures. Modern statistics deals with far richer classes of probabilities. This suggests a kind of

paradox. If statistics is to be of any real use it must provide ways of boiling down great masses of data to a few humanly interpretable numbers. The Koopman-Pitman-Darmois theorem suggests this is impossible unless nature follows highly specialized laws which no one really believes.

There are two ways out of this conundrum. First, the Koopman-Pitman-Darmois theorem depends on reduction to fixed dimension. If the dimension of the reduction is allowed to grow with  $n$  a theory may be possible. As an illustration, in the convex set example of §1, the extremal points of the sample were a sufficient statistic. As the sample size grows, a polyhedral convex set has order  $(\log n)$  extremal points. See Gröenboom [23] for recent work. I do not know of a theory that uses these ideas.

The second way around the conundrum uses the idea of approximate sufficiency. This idea has been developed in a comprehensive fashion by Lucian Le Cam. As an example, a statistic  $T$  is approximately sufficient for a family  $\mathcal{P}$  if

$$\sup_{P, Q \in \mathcal{P}} d(P(\cdot|T=t), Q(\cdot|T=t))$$

is small, where  $d$  is a metric on measures such as Hellinger's distance or total variation. Le Cam has shown that if a family admits an approximately sufficient statistic, then the best one can do using all of the data is only a small bit better than what is achievable using only the statistic. This is a small part of a dazzling body of work. Le Cam and Yang [36] is an accessible introduction.

There are several interesting aspects of sufficiency not described in this brief review. The elegant theory of completeness and sufficiency connects the analytic properties of a family of measure with the distribution of "what's left over after a sufficient reduction." See Lehmann [38] for a recent review. The theory of minimal sufficiency asks about the existence of smallest reductions. There are still fascinating open problems here. See Landers and Rogge [30].

Of course, one need not throw away what is left over. These "ancillary statistics" can be used to investigate if the family of measures under consideration is really a reasonable match to the data being considered. This is apparent in Fisher's early work. Diaconis and Smith [15] give examples and a review of the literature.

### 3. Introduction to exchangeability and equivalence of ensembles.

*A. de Finetti's theorem.* Let  $\mathbb{Z}_2 = \{0, 1\}$ . Let  $\mathbb{Z}_2^\infty$  be the infinite product space. A probability  $P$  on  $\mathbb{Z}_2^\infty$  is *exchangeable* if it is permutation invariant:  $P(0, 1, **\cdots) = P(1, 0, **\cdots)$ , etc. An example is coin tossing measure with parameter  $\theta$ :  $P_\theta(t) = \theta^t(1-\theta)^{n-t}$ ,  $t = x_1 + \cdots + x_n$ . Here and above  $\{x_1, x_2 \cdots x_n, **\cdots\}$  denotes the cylinder set in  $\mathbb{Z}_2^\infty$  which begins  $x_1, x_2 \cdots x_n$ , where  $x_i$  are binary digits.

One version of de Finetti's basic result is the following theorem.

**THEOREM (de Finetti).** *The set of all exchangeable probabilities on  $\mathbb{Z}_2^\infty$  is a convex simplex with extreme points the coin tossing measures  $\{P_\theta\}_{\theta \in [0,1]}$ .*

The theorem says that for each exchangeable  $P$  there is a unique probability  $\mu$  on  $[0, 1]$  such that the following integral representation holds:

$$(3.1) \quad P\{x_1, x_2 \cdots x_n\} = \int \theta^t (1 - \theta)^{n-t} \mu(d\theta), \quad t = x_1 + \cdots + x_n.$$

This holds for every  $n$  and binary sequence  $x_1 \cdots x_n$  with the same  $\mu$ .

de Finetti's motivation was philosophical. Statisticians have used expressions like the right-hand side of (3.1) since Bayes and Laplace. The term  $\theta^t (1 - \theta)^{n-t}$  is the likelihood of observing  $x_1 \cdots x_n$ . The measure  $\mu(d\theta)$  is the prior distribution. The integral represents the probability of observing  $x_1 \cdots x_n$  averaging over different values of  $\theta$ .

Subjective Bayesians like de Finetti prefer not to focus on unobservable parameters like  $\theta$ . They are perfectly willing to assign probabilities to observable outcomes like the next  $n$  flips of a coin. de Finetti's theorem shows that a simple invariance condition characterizes the classical assignments. The theorem does more: starting from an exchangeable measure on observables, the theorem builds a "parameter space"  $[0, 1]$ , and the likelihood and prior as part of its representation.

A clear, readable introduction to de Finetti's point of view appears in de Finetti [9]. Exchangeability is of interest in many areas of probability. de Finetti's theorem can be shown to be easily equivalent to Hausdorff's moment problem. See Feller [17]. The survey by Aldous [1] gives a splendid treatment with many other applications.

It is natural to try to develop parallel characterizations of the classical parametric models of statistics. As will be seen, symmetry can only go part of the way. The next section uses sufficiency to build a satisfactory general theory. We begin by changing the space and group.

**B. Freedman's theorem.** In 1962, David Freedman gave a version of de Finetti's theorem suitable for the normal distribution. Call a probability  $P$  on  $\mathbb{R}^\infty$  *orthogonally invariant* if

$$(3.2) \quad P(A * * \cdots) = P(\Gamma A * * \cdots)$$

for every cylinder set  $A * \cdots *$  with  $A \subset \mathbb{R}^n$  for some  $n$  and  $\Gamma$  in the orthogonal group  $O(n)$ .

**THEOREM (Freedman).** *The orthogonally invariant probabilities on  $\mathbb{R}^\infty$  are a convex simplex with extreme points  $\{P_\sigma\}_{\sigma \in [0, \infty)}$ , where  $P_\sigma$  is the product measure on  $\mathbb{R}^\infty$  of a mean 0, variance  $\sigma^2$  Gaussian measure.*

The theorem says for every orthogonally invariant  $P$  on  $\mathbb{R}^\infty$  there is a unique probability  $\mu$  on  $[0, \infty)$  such that

$$P(A * * \cdots) = \int_A \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-(x_1^2 + \cdots + x_n^2)/2\sigma} \mu(d\sigma).$$

The present version of the theorem arose in Bayesian statistics. Earlier, equivalent versions arose in Schoenberg's [47] answer to a question in functional analysis: When can a metric space be isometrically imbedded in  $\mathcal{L}^2$ ? Berg, Christensen, and Ressel [4] and Graham [22] give recent surveys of this line of work. The theorem can also be phrased as a description of all natural measures on  $\ell^2$ —this space is too big to have translation invariant measures but orthogonally invariant measures are widely used as a substitute. Choquet [7] contains an extensive discussion.

Perhaps the oldest version in widespread use is a theorem in geometry. This result goes back at least to Mehler [41]:

Let  $S_{n-1} = \{(x_1 \cdots x_n) \in \mathbb{R}^n : x_1^2 + \cdots + x_n^2 = n\}$ . Pick a point from the uniform distribution  $U$  on  $S_{n-1}$ . The theorem says that the first coordinate of such a point has an approximate Gaussian distribution: for every real  $a < b$ , as  $n$  tends to infinity

$$U\{x \in S_{n-1} : a < x_1 < b\} \sim \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

A proof is easy by calculus. One is required to calculate the surface area of a sphere between a pair of parallel planes. Mehler derived the result while looking at orthogonal expansions on high-dimensional spheres.

An extension of the result implies Freedman's theorem: Indeed, the orthogonally invariant probabilities on  $\mathbb{R}^n$  form a convex set. The extreme points are the uniform distribution on spheres.

$$\begin{aligned} & U\{a_1 \leq x_1 \leq b_1 \cdots a_k \leq x_k \leq b_k\} \\ & \sim \int_{a_1}^{b_1} \cdots \int_{a_k}^{b_k} \frac{1}{(\sqrt{2\pi})^k} e^{-(x_1^2 + \cdots + x_k^2)/2} dx_1 \cdots dx_k. \end{aligned}$$

This shows that the extreme points are approximately products of Gaussian measures and, for measures arising from orthogonally invariant probabilities  $P$  on  $\mathbb{R}^\infty$ , must be exactly product Gaussian.

A careful version of this argument with error estimates appears in Diaconis and Freedman [13] or in Diaconis, Eaton, and Lauritzen [11]. The latter authors discuss the following variant: pick  $\Gamma$  at random (Haar measure) in  $O(n)$ . The joint distribution of  $\Gamma_{ij}$ ,  $i, j \ll n^{1/3}$ , are approximately independent product normal variables.

As a final variant, the result appears in the statistical mechanics literature phrased as a simple example of the equivalence of ensembles. Here, a system is constrained to move on a constant energy hypersurface in  $6n$ -dimensional space. In the easiest case (with no interaction) this surface can be taken as the sphere:

$$x_1^2 + \cdots + x_{6n}^2 = c.$$

In statistical mechanics, the chance of finding the system in some portion of phase space is given by the uniform distribution (the microcanonical

ensemble). Physicists routinely calculate with a different measure (the macrocanonical ensemble) supported on all of  $\mathbb{R}^{6d}$ . In the simple example considered here, this has density proportional to  $e^{-(x_1^2 + x_2^2 + \dots + x_{6n}^2)/2\sigma^2}$  with  $\sigma^2$  chosen to make the average energy equal to  $c$ . The equivalence of ensembles says that for certain sets the calculation under the macrocanonical ensemble is approximately equal to the calculation under the microcanonical distribution. Usually the bounds are fairly crude—enough to show that sets which are small under one measure are small under the second. In this simple setting, the quantitative versions of Freedman's theorem give more precise results. The microcanonical ensemble is approximately product normal for sets which only depend on  $o(n)$  coordinates. See Diaconis and Freedman [13] for a precise statement.

The equivalence of ensembles holds for very general energy functions. Lanford [31] or Ruelle [46] give further details. The general set-up is closely related to the general versions of de Finetti's theorem explained in the next section.

**4. Sufficiency and exchangeability.** The work on de Finetti's theorem described in §3 can be summarized as the study of measures invariant under a group. In the examples, the extreme points were identified and parametrized by a nice set:  $[0, 1]$  for exchangeable binary sequences and  $[0, \infty)$  for orthogonally invariant processes. These are special situations. In contrast, the basic set-up of ergodic theory considers processes indexed by  $\mathbb{Z}$ , with  $\mathbb{Z}$  acting by translation. Now there is no neat description of the extreme points—instead they are dense in the space of all invariant measures.

The problem of finding a generalization of the examples which would handle the standard families of mathematical statistics was solved using the language of sufficiency. To explain, observe that the exchangeable processes can either be characterized as measures invariant under the permutation group or as measures for which the sum is a sufficient statistic. Thus a measure is exchangeable if and only if, for each  $n$ ,

$$P(x_1 \cdots x_n | x_1 + \cdots + x_n = t)$$

is uniform on all binary  $n$ -tuples with  $t$  ones.

Similarly, a measure is orthogonally invariant if and only if

$$P\{\cdot | x_1^2 + x_2^2 + \cdots + x_n^2 = t\}$$

is uniform on the  $\sqrt{t}$  sphere. The following abstraction covers most cases of interest in statistics.

For each  $i$ , there is a space  $\Omega_i$  (usually taken as a Polish space with its Borel  $\sigma$ -algebra). Let  $\Omega = \prod_{i=1}^{\infty} \Omega_i$ . For each  $n$ , there is a "sufficient statistic"  $T_n : \prod_{i=1}^n \Omega_i \rightarrow W_n$ , where  $W_n$  is some range space. The analog of the uniform distribution on the inverse image of  $T_n$  is played by a family of pre-specified measures  $Q_{n,t}$  on  $\prod_{i=1}^n \Omega_i$ .

Given  $T_n$  and  $Q_{n,t}$ , define the class of partially exchangeable processes  $M_{Q,T}$  as all  $P$  on  $\Omega$  such that

$$P\{\cdot | T_n(x_1 \cdots x_n) = t\} = Q_{n,t}(\cdot).$$

More technically, a regular conditional distribution for  $P$  on the first  $n$  coordinates given  $T_n = t$  is  $Q_{n,t}$ .

The  $Q$ 's and  $T$ 's are required to fit together as follows:

$$(1) \quad Q_{n,t}\{T_n^{-1}(t)\} = 1.$$

(2) If

$$T_n(x_1 \cdots x_n) = T_n(x'_1 \cdots x'_n),$$

then

$$T_{n+1}(x_1 \cdots x_n, y) = T_{n+1}(x'_1 \cdots x'_n, y).$$

(3) For each  $s \in W_n$ ,  $t \in W_{n+1}$ ,

$$Q_{n+1,t}(x_1 \cdots x_n | T_n(x_1 \cdots x_n) = s, x_{n+1}) = Q_{n,s}(x_1 \cdots x_n).$$

As an example, for coin tossing,  $\Omega_i = \{0, 1\}$ ,  $T_n(x_1, \dots, x_n) = x_1 + \cdots + x_n$ , and  $Q_{n,t}$  is taken as uniform over all  $x_1, \dots, x_n$  with  $x_1 + \cdots + x_n = t$ . Conditions (1)–(3) are easy to check. For example, (3) says that if one is told there are  $s$  ones in the first  $n$  places and told  $x_{n+1}$ , then  $Q_{n+1}$  assigns equal conditional probability to all compatible strings.

It is easy to see that the partially exchangeable processes  $M_{Q,T}$  form a convex set. The first problem is to find a description of the extreme points. This involves an excursion to infinity. Let  $\Sigma = \bigcap_{n=1}^{\infty} \Sigma_n$  with  $\Sigma_n$  the  $\sigma$ -algebra generated by  $T_n(X_1 \cdots X_n)$ ,  $X_{n+1}$ ,  $X_{n+2}$ ,  $\dots$ . This  $\Sigma$  is called the partially exchangeable  $\sigma$ -algebra. The first result is the following abstract version of de Finetti's theorem due to Diaconis and Freedman [12].

**THEOREM.** *If  $Q_n$  and  $T_n$  satisfy (1–3) above, then there is an  $E \in \Sigma$  such that  $P(E) = 1$  for each  $E \in M_{Q,T}$  and such that*

(a)  $Q_n, T_n(X_1 \cdots X_n)$  converges weak-star to a limit  $Q_{(\omega)}$  as  $n \rightarrow \infty$ , for each  $\omega \in E$ .

(b)  $\{Q_{\omega}\}_{\omega \in E}$  ranges over the extreme points of the convex set  $M_{Q,T}$ .

(c) For each  $P \in M_{Q,T}$ , there is a unique  $\mu$  on  $E$  such that

$$P(\cdot) = \int_E Q_{\omega}(\cdot) \mu(d\omega).$$

The theorem evolved over generations. It begins in the group invariant case with Krylov and Bogulyov. See Oxtoby [43] and Farrell [16]. Hunts' [27] axiomatic treatment of the Martin Boundary of a Markov chain is very close to giving the full result. The crucial conditions (2) and (3) were abstracted in early work of Freedman [20] and Bahadur [2].

A general version in rather different language was sketched by Martin-Löf [40] and Lauritzen [32–34] in Denmark. These authors worked in a more general setting of projective limits rather than with the product description of  $\Omega$ .

In developing the modern approach to statistical mechanics, Dobrushin, Lanford, and Ruelle developed a similar theory and conditions (1), (2), and (3) are known as the D-L-R conditions in statistical mechanics. Preston [44] or Georgii [21] contain recent presentations.

The theorem presents the extreme points in a rather abstracted form and further work is required to massage this presentation into a classical mold. Diaconis and Freedman [12] present dozens of examples which have occupied researchers in Bayesian statistics for the past thirty years. Aldous [1], Lauritzen [34] and Ressel [45] also present unified pictures from different points of view. The latter is interesting in presenting a large class of examples where the sufficient statistics are sums with values in a semigroup and the extreme points are indexed by the dual semigroup.

As one example of recent progress, here is a result of Küchler-Lauritzen [29] and Diaconis-Freedman [14]: Suppose one begins with an exponential family through a sufficient statistic  $T$ . One can then form the  $Q_{n,T}$  as the conditional laws determined by the family. This gives the ingredients of the general set-up and one can ask if the extreme points of  $M_{Q,T}$  correspond with the original exponential family. While it is easy to construct counterexamples, a natural sufficient condition has been found which gives the answer "yes" for any reasonable continuous or discrete family. The argument involves a delicate measure-theoretic extension of Cauchy's functional equation to partially defined functions. It gives infinitely many natural examples of  $Q$ 's and  $T$ 's where the extreme points have a simple description.

## BIBLIOGRAPHY

1. D. Aldous, *Exchangeability and related topics*, Lecture Notes in Math., vol. 1117, Springer, Berlin, 1984, pp. 1-198.
2. R. R. Bahadur, *Sufficiency and statistical decision functions*, Ann. Math. Statist. **25** (1954), 423-462.
3. O. Barndorff-Nielsen, *Information and exponential families in statistical theory*, Wiley, New York, 1978.
4. C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic analysis on semi-groups*, Springer, New York, 1984.
5. P. Billingsley, *Probability and measure*, Wiley, New York, 1979.
6. L. Brown, *Fundamentals of statistical exponential families*, Institute of Mathematical Statistics, Hayward, CA, 1987.
7. G. Choquet, *Lectures in analysis*, Vol. III, Benjamin, Reading, MA, 1969.
8. B. de Finetti, *Funzione caratteristica di un fenomeno aleatorio*, Atti Accad. Naz. Lincei Mem. Cl. Sci. Fis. Mat. Natur. (6) **4** (1931), 251-299.
9. —, *La prévision: ses lois logiques, ses sources subjectives*, Ann. Inst. H. Poincaré **7** (1937), 1-68; translated in *Studies in subjective probability* (H. Kyburg and H. Smokler, eds.), Wiley, New York, 1964.
10. M. E. Deyer, A. M. Frieze, and R. Kannan, *A random polynomial time algorithm for approximating the volume of convex bodies*, Proc. 21st ACM Sympos. on Theory of Computing, 1989, pp. 375-381.
11. P. Diaconis, J. Eaton, and S. Lauritzen, *Finite de Finetti theorems in linear models and multivariate analysis*, Scand. J. Statist. (to appear).



12. P. Diaconis and D. Freedman, *Partial exchangeability and sufficiency* (G. K. Ghosh and F. Roy, eds.), Statistics Applications and New Directions: Proc. Indian Stat. Inst. Golden Jubilee Internat. Conf. Stat., Indian Statistical Institute, Calcutta, India, 1984, pp. 205–236.
13. —, *A dozen de Finetti-style results in search of a theory*, Ann. Inst. H. Poincaré **23** (1987), 397–443.
14. —, *Cauchy's equation and de Finetti's theorem*, Scand. J. Statist. **17** (1990), 235–250.
15. P. Diaconis and L. Smith, *Residual analysis of discrete longitudinal data*, Technical Report, Department of Statistics, Stanford University, 1989.
16. R. H. Farrell, *Representation of invariant measures*, Illinois J. Math. **6** (1962), 447–467.
17. W. Feller, *An introduction to probability and its applications*, Vol. II, 2nd ed., Wiley, New York, 1971.
18. R. A. Fisher, *A mathematical examination of the methods of determining the accuracy of an observation by the mean error and by the mean squared error*, Monthly Notices Roy. Astronom. Soc. **80** (1920), 758–770.
19. —, *On the mathematical foundations of theoretical statistics*, Philos. Trans. Roy. Soc. London Ser. A **222** (1922), 309–368.
20. D. Freedman, *Invariants under mixing which generalize de Finetti's Theorem*, Ann. Math. Statist. **33** (1962), 916–923.
21. H. O. Georgii, *Gibbs measures and phase transitions*, de Gruyter, Berlin, 1988.
22. R. Graham, *Isometric embedding of graphs*, Graph Theory III (L. Beineke and R. Wilson, eds.), Academic Press, London, 1988.
23. P. Gröenboom, *Limit theorems for convex hulls*, Probab. Theory Related Fields **79** (1988), 327–368.
24. W. J. Hall, R. A. Wijsman, and J. K. Ghosh, *The relation between sufficiency and invariance with applications in sequential analysis*, Ann. Math. Statist. **36** (1965), 575–614.
25. P. R. Halmos and L. J. Savage, *Application of the Radon-Nikodym Theorem to the theory of sufficient statistics*, Ann. Math. Statist. **20** (1949), 225–241.
26. C. Hipp, *Sufficient statistics and exponential families*, Ann. Statist. **2** (1974), 1283–1292.
27. G. Hunt, *Markoff chains and Martin boundaries*, Illinois J. Math. **4** (1960), 313–340.
28. S. Johanson, *Introduction to the theory of regular exponential families*, Lecture Notes Series #3, Dept. of Statistics, Copenhagen University, 1979.
29. U. Küchler and S. Lauritzen, *Exponential families, extreme point models, and minimal space-time invariant functions for stochastic processes with stationary independent increments*, Scand. J. Statist. **15** (1989), 237–261.
30. D. Landers and L. Rogge, *Minimal sufficient  $\sigma$ -fields and minimal sufficient statistics. Two counterexamples*, Ann. Math. Statist. **43** (1972), 2045–2049.
31. O. Lanford, *Entropy and equilibrium states in classical statistical mechanics*, Lecture Notes in Physics, vol. 20, Springer, Berlin, 1973, pp. 1–113.
32. S. L. Lauritzen, *On the interrelationships among sufficiency and some related concepts*, Technical Report, Dept. of Statistics, University of Copenhagen, 1974.
33. —, *Sufficiency, prediction, and extreme models*, Scand. J. Statist. **1** (1974), 128–134.
34. —, *Extremal families and systems of sufficient statistics*, Lecture Notes in Statistics, vol. 49, Springer, Berlin, 1988.
35. L. Le Cam, *Sufficiency and approximate sufficiency*, Ann. Math. Statist. **35** (1964), 1419–1455.
36. L. Le Cam and G. Yang, *Asymptotic methods in statistical decision theory*, Springer, New York, 1990.
37. E. Lehmann, *Theory of point estimation*, 2nd ed., Wiley, New York, 1983.
38. —, *Testing statistical hypotheses*, 2nd ed., Wiley, New York, 1986.
39. I. G. MacDonald, *Symmetric functions and Hall polynomials*, Oxford Univ. Press, Oxford, 1979.
40. P. Martin-Löf, *Repetitive structures and the relation between canonical and micro-canonical distributions in statistics and statistical mechanics*, Proc. Conf. Foundational Questions in Statistical Inference (O. Barndorf-Nielsen, P. Plaesild, G. Schon, eds.), Aarhus Univ. Press, Aarhus, 1974.

41. F. G. Mehler, *Ueber die Entwicklung einer function von beliebig vielen variablen nach Laplaceschen functionen höheren Ordnung*, J. für Math. (Crelle's J.) **66** (1866), 161–176.
42. J. Neyman, *Sur un teorema concennente le cosidette statiche sufficienti*, Giorn. Ist. Ital. Att. G (1935), 320–334.
43. J. C. Oxtoby, *Ergodic sets*, Bull. Amer. Math. Soc. **58** (1952), 116–136.
44. C. Preston, *Canonical and microcanonical Gibbs states*, Z. Wahrsch. Verw. Gebiete **46** (1979), 125–158.
45. P. Ressel, *de Finetti-type theorems. An analytical approach*, Ann. Probab. **13** (1985), 898–922.
46. D. Ruelle, *Thermodynamic formalism*, Addison-Wesley, Reading, MA, 1978.
47. I. J. Schoenberg, *Metric spaces and positive definite functions*, Trans. Amer. Math. Soc. **44** (1938), 522–536.
48. S. M. Stigler, *Laplace, Fisher, and the discovery of sufficiency*, Biometrika **60** (1973), 439–445.

DEPARTMENT OF MATHEMATICS, HARVARD UNIVERSITY, CAMBRIDGE, MASSACHUSETTS  
02138

## Atoms and Analytic Number Theory

C. FEFFERMAN

In this brief, expository article I will explain some recent work on a mathematical problem arising from the elementary quantum mechanics of atoms and molecules. That problem is to compute the ground-state energy of a single atom with large atomic number  $Z$ . To account for chemical phenomena, one wants a very accurate determination of the ground-state energy  $E(Z)$  and its analogue for molecules. Unfortunately, such great accuracy is more than we can achieve today. Nevertheless, there has been interesting recent progress. In particular, L. Seco and I have rigorously computed  $E(Z)$  with a percentage error  $o(Z^{-2/3})$ ; and Ivrii and Sigal have rigorously computed the analogue of  $E(Z)$  for a molecule with a percentage error  $o(Z^{-1/3})$ . This article picks out one aspect of the recent work, namely the connection of the asymptotics of  $E(Z)$  to analytic number theory. More generally, number-theoretic questions arise whenever one asks for precise asymptotics for a quantum-mechanical problem with many particles. Let us see this in the simplest possible quantum system, namely  $N$  free particles in a box. Then we will return to atoms.

Imagine, then,  $N$  free particles in a box  $Q = [-\pi, \pi]^3 \subset \mathbb{R}^3$ . We assume the particles obey Fermi statistics. What is the lowest possible kinetic energy  $KE(N)$  for the  $N$  particles? The problem is an exercise in elementary mathematics. We are trying to minimize  $\int_{Q^N} (-\Delta\psi)\bar{\psi} dx_1 \cdots dx_N$  over all wave functions  $\psi(x_1, \dots, x_N) \in L^2(Q^N)$  of norm 1, which satisfy the antisymmetry condition  $\psi(x_{\sigma_1}, \dots, x_{\sigma_N}) = (\text{sgn } \sigma)\psi(x_1, \dots, x_N)$  for permutations  $\sigma$ . (Here we ignore spin. This simplifies notation and changes a few coefficients, but does not affect the essential ideas.)

The problem is easily solved by separation of variables and the lowest kinetic energy is simply the least possible sum of the form  $|\xi_1|^2 + \cdots + |\xi_N|^2$ , where  $\xi_1, \dots, \xi_N$  are distinct lattice points in  $\mathbb{R}^3$ . Moreover, it is geometrically obvious how to place  $N$  lattice points to minimize the sum of

the squares of their norms. We simply pick a ball  $B(0, R) \subset \mathbb{R}^3$ , containing at least  $N$  lattice points in its closure, but at most  $N$  in its interior. Our  $\xi_1, \dots, \xi_N$  merely consist of all the lattice points in the interior of the ball, together with just enough on the bounding sphere to make a total of  $N$  points. If we are satisfied with crude asymptotics for  $N$  large, then we need only remark that the number of lattice points is approximately equal to the volume, so that

$$\begin{aligned} N &\approx \text{number of lattice points in } B(0, R) \approx \frac{4}{3}\pi R^3, \text{ and} \\ \text{Min. Kinetic Energy} &\approx \text{Sum of the squares of the norms} \\ &\text{of the lattice points in } B(0, R) \approx \int_{\xi \in B(0, R)} |\xi|^2 d \text{vol}(\xi) = \\ &(\text{const})R^5. \end{aligned}$$

Consequently,

$$\text{Min. Kinetic Energy} \approx (\text{const}')N^{5/3},$$

as is well known to anyone interested in quantum mechanics. However, if we want to know more precise information on the minimum kinetic energy, then clearly we need to know the number of lattice points in a ball of radius  $R$ . This is a classical problem of analytic number theory. So, already for free particles in a box, we encounter number theory if we ask for precise information. In more complicated quantum-mechanical systems such as a large atom, number-theoretic issues are still present.

Let me recall very briefly what is known about lattice points in a ball. Trivially, a ball  $B(0, R)$  in  $\mathbb{R}^n$  contains  $\omega_n R^n + O(R^{n-1})$  lattice points, where  $\omega_n$  is the volume of the unit ball. The error term simply counts the number of lattice points within distance  $O(1)$  of the boundary. However, the trivial error term can be improved. Hardy showed in 1913 that the number of lattice points in a disc of radius  $R$  is  $\pi R^2 + O(R^{2/3})$ . The best error term in two dimensions is conjectured to be  $O(R^{1/2+\epsilon})$ , but only slight improvements on Hardy's result are known at present. All we need for our present work on quantum mechanics is to improve the trivial error term to  $O(R^{n-1-\epsilon})$ . Thus, Hardy's result is enough for the moment.

Hardy's work relies crucially on the fact that the circle is curved. If we try to count the lattice points in a large square of side  $R$ , then the obvious estimate  $R^2 + O(R)$  will be best possible. Hardy's result generalizes from the circle to any domain in  $\mathbb{R}^n$  with a smooth boundary whose principal curvatures never vanish. Improvements over Hardy's theorem are special for the circle.

Now I will return to atoms, and sketch the heuristic picture of a large atom, discovered by the physicists. We are trying to find the lowest eigenvalue  $E(Z)$  of the Hamiltonian

$$(1) \quad H = \sum_{k=1}^Z \left( -\Delta_{x_k} - \frac{Z}{|x_k|} \right) + \sum_{i \leq j < k \leq Z} \frac{1}{|x_j - x_k|},$$

acting on antisymmetric functions  $\psi(x_1, \dots, x_Z) \in L^2(\mathbb{R}^{3Z})$ . This problem is hard because of the term  $\sum 1/|x_j - x_k|$ , in which the electrons interact. If our Hamiltonian had the form

$$(2) \quad H_0 = \sum_{k=1}^Z (-\Delta_{x_k} + V(x_k))$$

for a one-electron potential  $V$ , then by separation of variables the problem would reduce from  $3Z$  to three dimensions. In fact, exploiting the spherical symmetry, we would really be dealing merely with one-dimensional problems (ODEs). The main idea used to study atoms in physics and chemistry is to approximate (1) by (2), with  $V(x)$  picked as follows: Pretend for a moment that we know the particle density  $\rho(x)$  for the electrons. Thus,  $\rho(x)$  is defined on  $\mathbb{R}^3$ , and the integral of  $\rho$  over a set  $\Omega \subset \mathbb{R}^3$  is equal to the expected number of electrons found in  $\Omega$ , according to the full probability density

$$d \text{ Prob} = |\psi(x_1, \dots, x_N)|^2 dx_1 \cdots dx_N.$$

Here  $\psi$  is the ground-state eigenfunction for  $H$ .

Once we know  $\rho$ , we can cook up a potential  $V(x)$  for which  $H_0$  is a good approximation for  $H$ . In fact, we just take

$$(3) \quad V(x) = -\frac{Z}{|x|} + \int_{\mathbb{R}^3} \frac{\rho(y) dy}{|x - y|}.$$

If the electrons act more or less independently, then each electron  $x_k$  feels the repulsion from all other electrons approximately as if it were being repelled by a fixed, continuous charge distribution  $\rho$ . Hence it is plausible that  $H$  should be well approximated by  $H_0$ . (Actually,  $H$  should be approximated not by  $H_0$  but by  $H_0 - (\text{Large Constant})$ . We suppress a careful discussion of this point.)

The *Hartree-Fock* approximation approximates the ground-state eigenfunction for  $H$  by that of  $H_0$ . The ground-state eigenfunction for  $H_0$  which we call  $\psi_{\text{hf}}(x_1, \dots, x_N)$ , can of course be written explicitly as an antisymmetrized product of the eigenfunctions  $\phi_1(x), \dots, \phi_N(x)$  for the three-dimensional Schrödinger operator  $-\Delta + V(x)$ .

Once we have a guess  $\psi_{\text{hf}}$  for the ground-state eigenfunction, we can then produce a guess for the lowest eigenvalue  $E(Z)$ , merely by taking

$$E_{\text{hf}}(Z) = \langle H \psi_{\text{hf}}, \psi_{\text{hf}} \rangle.$$

This is an immense simplification over the original problem, because we are dealing with a three-dimensional problem instead of a  $3Z$ -dimensional one. To make it work, we need to decide which density  $\rho(x)$  to start with. Fortunately, the Hartree-Fock approximation leads to a natural equation for  $\rho$ . In fact, the Hartree-Fock wave function  $\psi_{\text{hf}}(x_1, \dots, x_N)$  gives rise to its own particle density  $\rho_{\text{hf}}(x)$  on  $\mathbb{R}^3$ . In terms of the eigenfunctions  $\phi_k$  of

$-\Delta + V(x)$  on  $\mathbb{R}^3$ ,  $\rho_{\text{hf}}$  is simply

$$\rho_{\text{hf}}(x) = \sum_{k=1}^N |\varphi_k(x)|^2.$$

We call this the Hartree-Fock density arising from  $\rho$ . To get a self-consistent approximation, we demand that  $\rho_{\text{hf}}(x) = \rho(x)$ , i.e., the density we produce must equal the density we started with. This is the Hartree-Fock equation. To recapitulate,  $\rho(x)$  gives rise to  $V(x)$  by (3); from  $V(x)$  we produce the eigenfunctions  $\varphi_1(x), \dots, \varphi_N(x)$  of  $-\Delta + V(x)$ ; and we demand that  $\rho$  be picked so that  $\rho_{\text{hf}} = \sum_{k=1}^N |\varphi_k|^2$  will be equal to  $\rho$ . This is a rather strange equation. It certainly is not a partial differential equation or an integral equation in the usual sense. To solve it in practice, physicists and chemists use the following successive approximation scheme: Suppose we can produce a reasonable initial guess  $\rho_0$  for the unknown particle density. Then we can successively define  $\rho_1, \rho_2, \dots$  by taking  $\rho_{k+1}$  to be the Hartree-Fock density arising from  $\rho_k$ . This appears in practice to lead to satisfactory approximate solutions to the Hartree-Fock equation after a few steps. It is certainly not immediately obvious why it should work. It is also not immediately obvious how to produce a good initial guess  $\rho_0$ . One way to find a  $\rho_0$  is to use the Thomas-Fermi theory, which we now describe.

Thomas-Fermi theory is based on approximations to the density and eigenvalue sum of a three-dimensional Schrödinger operator  $-\Delta + V(x)$ . If  $E_k$  are the negative eigenvalues of  $-\Delta + V(x)$  and  $\varphi_k(x)$  are the corresponding eigenfunctions, then we want to approximate  $\text{sneg}(V) = \sum_k E_k$  and  $\rho_{\text{hf}}(x) = \sum_k |\varphi_k(x)|^2$ . These are the important quantities for us, since  $\text{sneg}(V)$  is the lowest possible energy for the Hamiltonian (2), and  $\rho_{\text{hf}}$  is one side of the Hartree-Fock equation that determines the density. The *semiclassical approximations* are

$$(4) \quad \text{sneg}(V) \approx -(\text{const}) \int_{V < 0} |V|^{5/2} dx,$$

$$(5) \quad \rho_{\text{hf}}(x) \approx (\text{const}) |V(x)|^{3/2} \chi_{V(x) < 0}.$$

These approximations work well when  $V(x)$  is large and rather slowly varying, e.g.,  $V(x) = \lambda^2 V_0(x)$ , with  $V_0(x)$  fixed and smooth and  $\lambda \gg 1$ . To prove (4) and (5), one can follow the same ideas introduced by H. Weyl to prove that the number of eigenvalues  $< \lambda$  for the Laplacian on a domain  $\Omega \subset \mathbb{R}^n$  is asymptotic to  $c_n \lambda^{n/2} \text{vol } \Omega$  when  $\lambda \rightarrow \infty$ .

Using the semiclassical approximations, we can rewrite the Hartree-Fock equation  $\rho_{\text{hf}} = \rho$  in the much more pleasant form

$$(6) \quad \rho(x) = (\text{const}) \cdot (-V(x))^{3/2}.$$

(We expect that  $V(x)$  will be negative on  $\mathbb{R}^3$ , so (5) leads to (6).) Equations (3) and (6) make up a coupled system of equations for the potential  $V$

and density  $\rho$ . It is trivial to eliminate either  $V$  or  $\rho$  from the equation using (6), so that we get a single integral equation for (say)  $\rho$ . Taking the Laplacian of both sides, we then get a (nonlinear) partial differential equation for  $\rho$ . Since  $\rho$  is expected to be spherically symmetric, equations (3) and (6) finally reduce to an ordinary differential equation. Thus, the density  $\rho$  and potential  $V$  may be read off from ordinary differential equations. These  $\rho$  and  $V$  represent the semiclassical approximations to the Hartree-Fock approximation to the real atom. They are called the Thomas-Fermi density  $\rho_{\text{TF}}$  and the Thomas-Fermi potential  $V_{\text{TF}}$ . The parameter  $Z$  in equations (3) and (6) may be removed by a trivial scaling, and therefore

$$(7) \quad \rho_{\text{TF}}(x) = Z^2 \rho_1(Z^{1/3}x),$$

$$(8) \quad V_{\text{TF}}(x) = Z^{4/3} V_1(Z^{1/3}x)$$

for universal functions  $\rho_1$  and  $V_1$ , which may be found by solving ordinary differential equations. Putting (8) into (4), we obtain  $\text{sneg}(V_{\text{TF}}) \approx -(\text{const})Z^{7/3}$ . Thus we have computed the ground-state energy for the Hamiltonian (2). After taking into account the additive constant mentioned just after (3), we obtain the Thomas-Fermi approximation to the ground-state energy of an atom, namely

$$(9) \quad E(Z) \approx -c_{\text{TF}}Z^{7/3}.$$

As noted in the introduction, we need highly accurate approximations to  $E(Z)$ . The Thomas-Fermi approximation (9) is very crude. Using heuristic methods, physicists found a closer approximation than (9) to the Hartree-Fock energy, namely

$$(10) \quad E(Z) \approx -c_{\text{TF}}Z^{7/3} + \frac{1}{8}Z^2 - c_{\text{DS}}Z^{5/3}.$$

Of course, the Hartree-Fock energy itself is only an approximation to the true ground-state energy  $E(Z)$ .

I will not take the time to give a complete account of the ideas that led to (10), but I want to single out one part of the argument, namely a proposed refinement of (4) and (5) due to Schwinger [6]. Schwinger's formulas are as follows:

$$(11) \quad \text{sneg}(V) \approx -(\text{const}) \int_{V < 0} |V|^{5/2} dx + (\text{const}') \int_{V < 0} (\Delta V) |V|^{1/2} dx,$$

$$(12) \quad \rho_{\text{hf}}(x) \approx \chi_{V(x) < 0} \cdot \{(\text{const})|V(x)|^{3/2} + (\text{const}')(\Delta V) \cdot |V(x)|^{-1/2} - (\text{const}'')|\nabla V|^2 |V(x)|^{-3/2}\}.$$

Schwinger found these formulas by trying to guess the general case from the example of the harmonic oscillator  $-d^2/dx^2 + \lambda^2 x^2$  in one dimension. Formula (11) has to be modified when  $V$  contains a Coulomb singularity. This concludes our sketch of the atom according to physicists.

Next we turn to rigorous results. An excellent survey of what was known up to about 1980 is contained in Lieb [3]. The main result then known on atoms was the theorem of Lieb-Simon [5] that  $E(Z) = -c_{\text{TF}}Z^{7/3} + O(Z^a)$  with an explicit  $a$  between  $\frac{7}{3}$  and 2. The proof was very much in the spirit of Weyl's work on eigenvalues. It applies also to molecules. An important unsolved problem was to prove the "Scott conjecture," i.e.,  $E(Z) = -c_{\text{TF}}Z^{7/3} + \frac{1}{8}Z^2 + O(Z^a)$  with  $a < 2$ . This was settled by Hughes [2] and Siedentop-Weikard [7] in the mid-1980s. Recently, there has been further progress. Specifically, Ivrii and Sigal have proven the analogue of the Scott conjecture for a molecule; and L. Seco and I have proven [1] that

$$(12a) \quad E(Z) = -c_{\text{TF}}Z^{7/3} + \frac{1}{8}Z^2 - c_{\text{DS}}Z^{5/3} + O(Z^a)$$

for an  $a < \frac{5}{3}$  for atoms, justifying (10). It would be very interesting to combine the two results into a rigorous computation of the energy of a molecule modulo  $o(Z^{5/3})$ . It would also be interesting to write down the next correction term beyond  $Z^{5/3}$  in the asymptotic behavior of  $E(Z)$ . Our work suggests the form of the next term, but is not strong enough to prove it. The next term in  $E(Z)$  is not  $Z^{4/3}$  as one might expect, but rather a series from analytic number theory. That series is closely related to lattice point problems. It fluctuates as a function of  $Z$ , and is not proportional to any power of  $Z$ . Similar fluctuations occur already for free particles in a box.

How can one produce rigorous upper and lower bounds for  $E(Z)$ ? In principle, upper bounds are easy. Starting from the Thomas-Fermi density  $\rho_{\text{TF}}$  and its potential  $V_{\text{TF}}$ , we form the Hartree-Fock wave function<sup>1</sup>  $\psi_{\text{hf}}(x_1, \dots, x_N)$  by taking the antisymmetrized product of the eigenfunctions  $\phi_1(x), \dots, \phi_N(x)$  of  $-\Delta + V_{\text{TF}}$  on  $\mathbb{R}^3$ . We think but do not know that  $\psi_{\text{hf}}$  is close to the true ground-state. To make an upper bound for  $E(Z)$ , we have only to calculate the inner product  $\langle H\psi_{\text{hf}}, \psi_{\text{hf}} \rangle$ , with  $H$  given by (1). Minimax tells us that this inner product is a rigorous upper bound for  $E(Z)$ , whether or not  $\psi_{\text{hf}}$  is close to the true ground-state. If our opinions are correct, our upper bound will turn out to be quite sharp.

Producing lower bounds for  $E(Z)$  is much harder. Instead of computing  $\langle H\psi, \psi \rangle$  for a single wave function  $\psi$ , we need to prove a lower bound for  $\langle H\psi, \psi \rangle$  with an arbitrary antisymmetric  $\psi$ . The usual starting point, which goes back to Lieb [4], is to prove an inequality of the form

$$(13) \quad \sum_{i < j} \frac{1}{|x_i - x_j|} \geq \sum_{i=1}^N W(x_i) - C_0$$

for a suitable one-electron potential  $W$  and constant  $C_0$ .

Once such an inequality is known, it follows that the true Hamiltonian  $H$

<sup>1</sup>Actually, it is enough to form  $\psi_{\text{hf}}$  by guessing the eigenfunctions of  $-\Delta + V$ .



in (1) satisfies

$$H \geq \sum_{k=1}^Z \left( -\Delta_{x_k} - \frac{Z}{|x_k|} + W(x_k) \right) - C_0.$$

The right-hand side contains no interactions, and hence may be understood by separation of variables. Thus,  $E(Z) \geq \text{sneg}(-Z/|x| + W) - C_0$ , so that we have a rigorous lower bound for the ground-state energy. Whether this lower bound is sharp or useless depends on our skill in picking and proving a good inequality of the form (13). I will explain one way to prove (13) that leads to reasonable results. The starting point is an elementary observation, namely

$$(14) \quad |x - x'|^{-1} = \frac{1}{\pi} \int_{\substack{y \in \mathbb{R}^3 \\ R > 0}} \chi_{x, x' \in B(y, R)} \frac{dy dR}{R^5} \quad \text{for } x, x' \in \mathbb{R}^3.$$

Except for the fact that the coefficient here is  $\frac{1}{\pi}$ , this equation follows merely by noting that both sides have the same invariance under translations, rotations, and dilations. Taking  $x = x_j$ ,  $x' = x_k$  in (14) and summing over all possible pairs of particles, we see that

$$(15) \quad \sum_{i < j} |x_i - x_j|^{-1} = \frac{1}{\pi} \int_{\substack{y \in \mathbb{R}^3 \\ R > 0}} \frac{N(y, R)(N(y, R) - 1)}{2} \frac{dy dR}{R^5}$$

for any  $x_1, \dots, x_N \in \mathbb{R}^3$ , with  $N(y, R) = \text{number of } x_j \in B(y, R) = \sum_j \chi_{B(y, R)}(x_j)$ .

The next step is to make a guess  $\bar{N}(y, R)$  for the number of electrons in  $B(y, R)$ . For instance, we may take

$$(16) \quad \bar{N}(y, R) = \int_{B(y, R)} \rho_{\text{TF}}.$$

We believe (but do not know) that this guess is a good one. Using  $\bar{N}(y, R)$ , we rewrite the integrand in (15) as

$$(17) \quad \begin{aligned} \frac{1}{2} N(y, R)(N(y, R) - 1) &= \frac{1}{2} [N(y, R) - \bar{N}(y, R)]^2 \\ &\quad + [\bar{N}(y, R) - \frac{1}{2} N(y, R) - \frac{1}{2} [\bar{N}(y, R)]^2]. \end{aligned}$$

When we substitute (17) into (15), the term  $[\bar{N}(y, R) - \frac{1}{2} N(y, R)]$  contributes something of the form  $\sum_k W(x_k)$  to the energy, because  $N(y, R) = \sum_k \chi_{B(y, R)}(x_k)$  and  $\bar{N}(y, R)$  are independent of the  $x_k$ . The term  $-\frac{1}{2} [\bar{N}(y, R)]^2$  in (17) is also independent of the  $x_k$ , and hence merely contributes an additive constant to the energy.

The part of (17) that is hard to understand is  $[N(y, R) - \bar{N}(y, R)]^2$ . We can get a cheap lower bound, simply by discarding this positive term and writing

$$(18) \quad \frac{1}{2} N(y, R)[N(y, R) - 1] \geq [\bar{N}(y, R) - \frac{1}{2} N(y, R) - \frac{1}{2} [\bar{N}(y, R)]^2].$$

If our guess  $\bar{N}(y, R)$  was intelligent and if  $B(y, R)$  is likely to contain many electrons, then  $[N(y, R) - \bar{N}(y, R)]^2$  should be negligibly small compared to  $\frac{1}{2}N(y, R)[N(y, R) - 1]$ . Hence (18) may provide useful information.

However, if  $B(y, R)$  is small enough, then it will probably contain either 0 or 1 electrons. In this case,  $\bar{N}(y, R) \ll 1$ ,  $N(y, R) = 0$  or 1, and  $[N(y, R) - \bar{N}(y, R)]^2$  is not negligible compared to  $\frac{1}{2}N(y, R)[N(y, R) - 1]$ , so that using (18) is a bad idea. For small  $B(y, R)$  it is better just to use the trivial lower bound  $\frac{1}{2}N(y, R)[N(y, R) - 1] \geq 0$ . Therefore, we set

$$E = \{(y, R) | \bar{N}(y, R) > 1\}$$

(say), and we conclude from (15) and (18) that

$$\begin{aligned} \sum_{i < j} |x_i - x_j|^{-1} &\geq \frac{1}{\pi} \int_{(y, R) \in E} \left[ \bar{N}(y, R) - \frac{1}{2} \right] N(y, R) \frac{dy dR}{R^5} \\ &\quad - \frac{1}{\pi} \int_{(y, R) \in E} \frac{1}{2} [\bar{N}(y, R)]^2 \frac{dy dR}{R^5}. \end{aligned}$$

The first integral on the right has the form  $\sum_k W(x_k)$ , and the second integral is an additive constant  $C_0$ . Hence we have succeeded in proving an inequality of the form (13). If we define  $\bar{N}(y, R)$  intelligently, e.g., by (16), then the resulting inequality (13) leads to a lower bound for  $E(Z)$  strong enough to prove the Scott conjecture.

Let us summarize the preceding discussion. Upper bounds for  $E(Z)$  are proved by using the Hartree-Fock approximate ground-state  $\psi_{\text{hf}}$  as a trial wave function. Lower bounds are proved by invoking (13) to obtain an inequality  $H \geq \sum_k (-\Delta_{x_k} + V(x_k)) - C_0$ . The right-hand side can be understood by separation of variables. Of course, it takes lots of hard work to understand  $-\Delta + V(x)$  on  $\mathbb{R}^3$  with enough precision to carry this out. For atoms, the hard work deals with ODE's since  $V(x)$  is spherically symmetric. To get the Scott conjecture for a molecule, one needs instead to understand a genuinely three-dimensional problem.

In a sense, we have been lucky so far. The original problem is  $3Z$ -dimensional with  $Z \rightarrow \infty$ , yet we have not had to look seriously at any quantum-mechanics problem in dimension greater than 3. The reason for the good luck is that we could drop the only difficult term,  $[N(y, R) - \bar{N}(y, R)]^2$ , from (17), and thus bound the true Hamiltonian from below by a noninteracting one. The contribution of that term to the total energy is of the order of magnitude  $Z^{5/3}$ , when  $\bar{N}(y, R)$  is picked optimally. Hence, we can reduce matters from  $3Z$  dimensions to three, provided we are willing to ignore errors  $O(Z^{5/3})$  in the energy. This is good enough for the Scott conjecture, but not for (12a). In proving (12a) we are forced for the first time to come to grips with the quantum mechanics of an interacting system. This is perhaps the main point in our proof of (12a), but we will not discuss it further here.

Instead, we turn to the Schwinger formulas (11) and (12), which must be rigorously understood before one can hope to derive (12a). It would be very

interesting to give a rigorous discussion of the Schwinger formulas for rather general  $V$  in three dimensions. What Seco and I did was much easier. We understood the spherically symmetric case by making a very precise analysis of the eigenvalues and eigenfunctions of ordinary differential equations. This is of course good enough for an atom, but not for a molecule. The details of our work on ODE's are long and complicated. In the end, we derive formulas for  $\text{sneg}(V)$  and  $\rho_{\text{hf}}$  analogous to (11) and (12). However, in addition to the terms displayed on the right in (11) and (12), there are number-theoretic series related to the lattice-point problem. We are not surprised to see such series, in view of the example of  $N$  free particles. Thus, (11) and (12) are correct, provided the number-theoretic terms are negligibly small. In (12), this is simply not the case. The number-theoretic terms actually dominate over Schwinger's correction terms, and (12) is wrong, at least for radial potentials. One important point in our proof of (12a) is therefore to use only relatively crude asymptotics for  $\rho_{\text{hf}}$ , in order to get away without (12).

The role of the number-theoretic series in (11) is less destructive. If we estimate it by trivial methods, analogous to the trivial  $\pi R^2 + O(R)$  for the number of lattice points in a disc, then we see that the series is at most of the same order of magnitude as Schwinger's correction term in (11). So to prove (11), we need to make a small improvement over the trivial estimate of the number-theoretic sum. The analogue of Hardy's  $O(R^{2/3})$  result will be enough for our purposes. However, to apply Hardy's theorem, we need nonzero curvature. The condition that plays the role of nonzero curvature here turns out to be the following.

**PERIODIC ORBIT CONDITION.** Form the classical Hamiltonian  $H_{\text{cl}} = |\vec{p}|^2 + V(\vec{q})$  for  $\vec{p}, \vec{q} \in \mathbb{R}^3$ . Then the set of periodic zero-energy orbits for  $H_{\text{cl}}$  has measure zero in the set of all zero-energy orbits.

The number-theoretic term in (11) can be dropped if and only if the periodic orbit condition is satisfied. Hence, Schwinger's formula for the eigenvalue sum holds for a radial potential  $V$  if and only if the periodic orbit condition is satisfied. The connection of sharp eigenvalue asymptotics to periodic orbits is of course an old story in the context of the Laplacian on a manifold. The connection to periodic orbits clearly points to wave equation methods in any future attempt to understand nonradial  $V$ .

The periodic orbit condition fails with a vengeance for the harmonic oscillator, where the whole Hamiltonian flow is periodic. Thus (11) fails there also, as the reader may check by elementary computation. It is remarkable that Schwinger correctly guessed (11) by thinking hard about one of the few examples in which it is false.

After repeatedly discussing unspecified number-theoretic sums, we close by writing down the series that we believe forms the next term in  $E(Z)$  beyond  $Z^{5/3}$ .

Let  $V_{\text{TF}}(r)$  be the Thomas-Fermi potential for atomic number  $Z$ . For

$l \geq 0$ , set  $V_l(r) = l(l+1)/r^2 + V_{\text{TF}}(r)$ . Then define

$$\eta_l = \int_{\{V_l < 0\}} |V_l(r)|^{-1/2} dr \quad \text{and} \quad \phi_l = \frac{1}{\pi} \int_{\{V_l < 0\}} |V_l(r)|^{+1/2} dr.$$

Let  $\beta(t) = |t - k|^2 - \frac{1}{12}$ , where  $k$  is the integer nearest to  $t$ . Our conjecture is that

$$E(Z) \approx -c_{\text{TF}} Z^{7/3} + \frac{1}{8} Z^2 - c_{\text{DS}} Z^{5/3} + (\text{const}) \sum_{l \geq 1} \frac{(2l+1)}{\eta_l} \beta(\phi_l).$$

## REFERENCES

1. C. Fefferman and L. Seco, *On the ground-state energy of a large atom*, Bull. Amer. Math. Soc. (N.S.) **23** (1990), 525–530.
2. W. Hughes, *An atomic energy lower bound that agrees with Scott's correction*, Adv. in Math. **79** (1990), 213–270.
3. E. Lieb, *Thomas-Fermi and related theories of atoms and molecules*, Rev. Modern Phys. **53** (1981), 603–641.
4. —, *A lower bound for Coulomb energies*, Phys. Lett. A **70** (1979), 444–446.
5. E. Lieb and B. Simon, *The Thomas-Fermi theory of atoms, molecules, and solids*, Adv. in Math. **23** (1977), 22–116.
6. J. Schwinger, *Thomas-Fermi model: The second correction*, Phys. Rev. A **24** (1981), 2353–2361.
7. H. Siedentop and R. Weikard, *On the leading energy correction for the statistical model of the atom: Interacting case*, Comm. Math. Phys. **112** (1987), 471–490.

DEPARTMENT OF MATHEMATICS, PRINCETON UNIVERSITY, PRINCETON, NEW JERSEY 08544

## Working and Playing with the 2-Disk

MICHAEL H. FREEDMAN

This article is simply a written lecture and what philosophy it contains should not necessarily be taken seriously. However, it is much easier to learn a whole story than a single theorem, so many of the latter are woven into the former. Our hero, for fun, is the two-dimensional disk which seems to intrude at many important junctures of geometric topology. Also, there is the theme that ideas of great importance can be enormously simple. As the Centennial recalls to each of us our small mortal places and seems to threaten even mathematics with a certain loss of youth—computer proofs, proofs too long to write (or think), the joint power and vacuity of abstraction—I enjoy recalling a few forceful but simple ideas in the subject I know best. I have no prediction for the next century but am content to express the hope that mathematics will still, from time to time, be extraordinarily easy—that the last simple idea is still far off.

By now, topologists have learned to watch developments in analysis with an opportunistic eye. In 1913, I do not know how much attention was given to the topological implications of:

**THEOREM** (C. Carathéodory [Car] and, independently, W. F. Osgood and E. M. Taylor [OT]). *If  $\mathcal{D}$  is a Jordan domain, then any Riemann mapping of the unit disk  $U \rightarrow \mathcal{D}$  extends to a homeomorphism of the closures  $\bar{U} \rightarrow \bar{\mathcal{D}}$ .*

It follows that every imbedding of the circle  $S^1$  into the plane extends to an imbedding of the disk:

$$\begin{array}{ccc} S^1 & \xhookrightarrow{i} & R^2 \\ \downarrow \theta & \searrow \text{dotted } j & \\ D^2 & \hookrightarrow & \end{array}$$

The hooked arrows are 1-1 maps—in general not supposed to be more than continuous. The dotted arrow is the conclusion, whereas the solid arrows are hypotheses. (The diagram is commutative.)

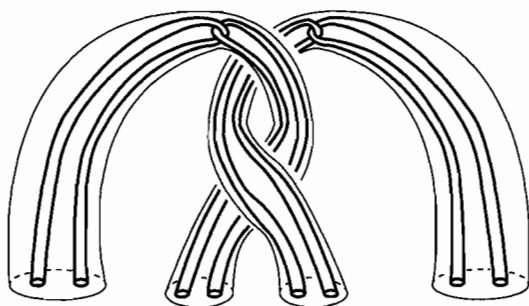


FIGURE 1.  $(S^3, B^3)/\text{Cantor set of arcs} \cong (S^3, \text{horned ball})$ .

Let  $\mathcal{D}$  be the interior domain of  $i(S^1)$ . The theorem finds a continuous (and, in fact, conformal on the interior) extension  $j'$  of some other parametrization  $i'$  of  $i(S^1)$ . The one-dimensional problem of isotoping  $i'$  to  $i$  is not hard and this leads to  $j$ .

Carathéodory's proof was an application of his recently developed theory of prime ends—a subject which is still a source of topological arguments (e.g., Sullivan's solution [S] of the Wandering Domain Problem). The other proof, while of less long-term importance, was discovered after W. F. Osgood had served (1905–1906) as President of the American Mathematical Society and is a striking example of life after bureaucratic service.

In 1922, J. Alexander, one of the founders of homology theory, announced (unpublished) a similar result regarding imbeddings of the two-dimensional spheres  $S^2$  in  $R^3$ . The argument was short-lived, for in 1924 Alexander published [A] the seminal counterexample, the *Alexander Horned Sphere*. Here we described it in a possibly unfamiliar way—but the usual image of infinitely interlocking horns can be retrieved with some scrutiny.

Imagine  $S^3 = R^3 \cup \infty$ . Attached to the horizontal plane  $P$  are a nested collection of solid cylinders as pictured in Figure 1.

At the “ $n$ th level” there are  $2^n$  solid cylinders and these are arranged so that the intersection of all levels is a Cantor set's worth of arcs which braid as they move upward. (The components of the intersection are arranged to be arcs by making each intersect horizontal planes in at most one point.) The braiding is increasingly rapid toward the upper end points and they are not topologically tame but wild.

Consider the quotient space (with the weak or *quotient* topology)  $S^3/\text{arcs}$  wherein each of these arcs is declared to be a point. The Alexander horned sphere is  $\pi(\bar{P})$  and the Alexander horned ball is  $\pi$  (upper half-space). By taking a limit of homeomorphisms,  $S^3 \rightarrow S^3$ , it is possible to find a map  $\theta: S^3 \rightarrow S^3$  whose nonpoint (usually called nontrivial) point preimages are exactly these arcs. The composition  $\pi \circ \theta^{-1}: S^3 \rightarrow S^3/\text{arcs}$  is a homeomorphism (in spite of the fact that  $\theta^{-1}$  is a relation!). In this way it is seen

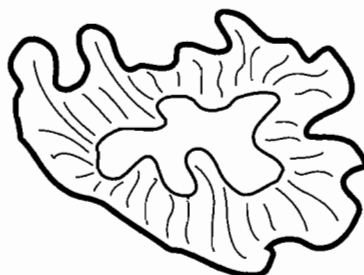


FIGURE 2

that these wild objects are actually subspaces of  $S^3$ . In fact, understanding this homeomorphism leads to the usual “horned” picture.

The horned sphere had intellectual descendants of two lineages. First it suggested two conjectures—repairs for the damage done by the counterexample—which became theorems during the following fifty-eight years. An imbedding of a space  $X$  is *collared* if it extends as  $X \times \frac{1}{2}$  to an imbedding of  $X \times [0, 1]$ . The theorems are:

**SCHOENFLIES THEOREM** (Proved by B. Mazur with finishing touches by M. Morse and slightly later by M. Brown, 1959; see [M], [Br]). *Any collared (topological) imbedding of  $S^{n-1} \rightarrow S^n$  extends to an imbedding of  $B^n$ .*

**ANNULUS THEOREM** (R. Kirby +  $\epsilon$  for  $n > 4$ , 1968, and F. Quinn for  $n = 4$ , 1982, see [K], [Q]). *Any collared imbedding of  $(S^{n-1} \amalg S^{n-1}) \rightarrow S^n$  extends to an imbedding of  $S^{n-1} \times [0, 1]$ .*

See Figure 2.

The other chain of descent attempted to explore rather than to define away the phenomenon. A key development came in 1952 when R. H. Bing [Bi] found that the *double* of the horned ball, DHB, is homeomorphic to the 3-sphere  $S^3$ . The double is defined by

$$DHB = HB \times \{0, 1\} / (x, 0) \sim (x', 1),$$

where  $(x, 0) \sim (x', 1)$  iff  $x = x'$  and  $x \in \text{frontier}(HB)$ . Bing's argument may be cast in the previous form by saying that he constructs a sequence of homeomorphisms  $S^3 \xrightarrow{\theta_i} S^3$  whose limit  $S^3 \xrightarrow{\theta} S^3$  has as its nontrivial point preimages the doubly wild Cantor set of arcs made by reflecting Figure 1 in the horizontal plane.

If  $\pi: S^3 \rightarrow (S^3 / \text{doubly wild Cantor set of arcs}) \cong DHB$  is the projection to the quotient space, the desired homeomorphism is  $\pi \circ \theta^{-1}$ . Unlike the earlier example, the shrinking homeomorphisms  $\theta_i$  are of extraordinary subtlety. They are generated by successive shears defined near the boundaries of the pictured tori (Figure 3 on next page). To get a feel, notice that rotation by roughly  $90^\circ$  in the angular coordinate of the large solid tori reduces the diameters of the smaller solid tori contained within them. Such

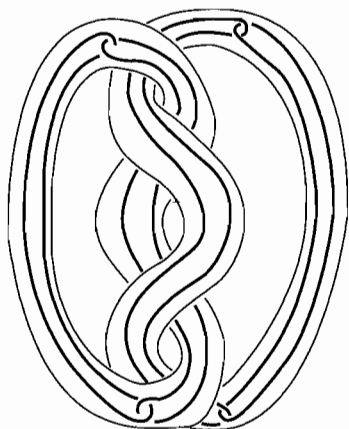


FIGURE 3

diameter reductions are painstakingly composed to reduce the diameter of each component of the  $\infty$ -stage—that is, each doubly wild arc—to zero.

The *pièce de résistance* of such shrinking arguments is the unpublished theorem of R. Edwards (1978; see [Da]). It gives a sharp criterion for when a quotient map of a high-dimensional topological manifold is approximable by homeomorphisms.

**THEOREM (Edwards).** *Let  $\pi: M^n \rightarrow X$ ,  $n \geq 5$ , be a C.E. map from a topological manifold onto a finite-dimensional ANR. Then  $\pi$  is approximable by homeomorphisms iff any map of the two-dimensional disk into the quotient  $f: D^2 \rightarrow X$  is approximable by an imbedding.*

A map is “C.E.” if every point inverse is null homotopic within any neighborhood of itself. All the hypotheses are now known to be indispensable. Roughly, we think of the theorem as saying that a quotient which might be a manifold is a manifold, provided it has manifold-like general position with respect to maps of the two-dimensional disk.

The proof is beautiful but too long to summarize here except to say that the two disks enter as the important parts of the dual to the  $(n-3)$ -skeleton of  $M$ . Experts know that  $(n-3)$  is the critical dimension for engulfing. I will not dwell on engulfing but later will spend a little time on its close cousin the  $h$ -cobordism theorem for which the 2-disk is also the key.

To redeem my promise that important results can be simple I now give a rather complete sketch of Brown’s proof of B. Mazur’s Schoenflies theorem [M]. It is that proof which, transfigured, reappears in the study of four-dimensional manifolds. First I should say that the stunning advances of algebraic, differential, and combinatorial topology in the forties and fifties together with the stunning stasis of the Schoenflies problem and its many relatives had led to a deep and well-informed pessimism on the prospects for naive geometric arguments in topology. It must have been a wonderful day



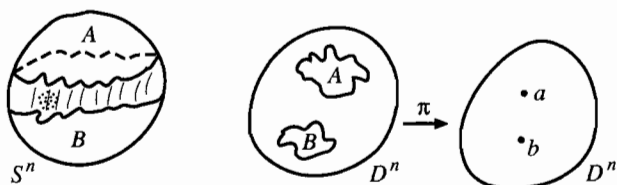


FIGURE 4

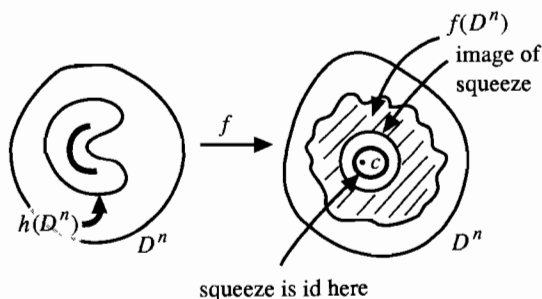


FIGURE 5

when Barry Mazur, then a graduate student at Princeton, cast the first bright light through the gloom.

**PROOF-SKETCH OF THE SCHOENFLIES THEOREM ACCORDING TO BROWN.**

Let  $A$  and  $B$  be the (closed) complementary pieces of  $S^n \setminus S^{n-1} \times (0, 1)$ . A subset on  $S^n$  is *cellular* if it can be written as a nested intersection of balls  $\bigcap_{i=1}^{\infty} B_i^n$ ,  $B_{i+1}^n \subset \text{int } B_i^n$ . The key is to show that  $A$  (or  $B$ ) is cellular, for then a simple limiting argument shrinks  $A$  to a point and in the process stretches the product collar lines of  $S^{n-1} \times (0, 1)$  into the radial lines of some polar coordinate system on one of the closed complementary components of  $S^{n-1} \times \frac{1}{2}$ —identifying it as a disk.

Remove a small open disk from  $S^{n-1} \times (0, 1)$  to obtain the picture shown in Figure 4.

The following lemma almost applies to the  $\pi$  in Figure 4.

**LEMMA.** *If  $f: D^n \rightarrow D^n$  has a single nontrivial point preimage  $f^{-1}(c) = C$  for  $c \in \text{int } D^n$ , then  $C$  is cellular. (Note that we do not assume  $f$  is onto.)*

**PROOF.** Consider  $h = "f^{-1} \circ \text{squeeze} \circ f"$  where squeeze is a "reimbedding" of  $D^n$  into a small neighborhood of  $c$ , which is the identity on a still smaller neighborhood. The quotient marks mean the imbedding which can be easily fashioned out of the relation that the notation literally describes. These imbeddings, associated to progressively stronger squeezes, show  $C$  is cellular. See Figure 5.  $\square$

To conclude the Brown proof, observe that " $f^{-1} \circ \text{squeeze}_b \circ f$ " :  $D^n \rightarrow D^n$  has  $A$  as its only nontrivial preimage. By the lemma,  $A$  is cellular.  $\square$

The Annulus Theorem could not be proved until manifold theory had reached maturity. It then required a brilliant device—the *torus trick*. The Schoenflies theorem can be used to replace the annulus conjecture with the conjecture that all homeomorphisms are *stable*. A homeomorphism  $h: R^n \rightarrow R^n$  is stable if it is a finite composition of homeomorphisms  $g_i$ , each of which is differentiable or piecewise linear on at least some open set  $U_i \subset R^n$ . This said, the pseudogroup of stable homeomorphisms and stable structures can be studied. It was known that all  $h: R^n \rightarrow R^n$  stable implies the annulus conjecture in  $S^n$ .

By a marvelous device which I cannot describe here, Kirby showed that any germ of  $h$  determines a potentially exotic triangulation of the  $n$ -torus  $T^n_\Gamma$ . The problem became the: "Hauptvermutung for Tori." That is, given  $T^n_\Gamma$  find a P.L. homeomorphism  $k: T^n_{\text{Standard}} \rightarrow T^n_\Gamma$ . If this could be done " $\text{id} \circ k = T^n_{\text{Standard}} \hookrightarrow$ " can be constructed. Any self-homeomorphism of  $T^n_{\text{Standard}}$  must be stable (by the controlled behavior of its lift to  $R^n$ ) and since  $k$  is P.L. it is a formality that " $\text{id}: T^n_\Gamma \rightarrow T^n_{\text{Standard}}$ " and, therefore,  $h: R^n \rightarrow R^n$  are stable.

Finding  $k$  involves deep manifold theory and actually cannot be done before a (harmless) passage to a  $2^n$ -fold covering space. The idea to pass to a cover was L. Siebenmann's; the construction of  $k$  (after covering) was carried out independently by T. Farrell, by W.-C. Hsiang and J. Shaneson, and by C.T.C. Wall.

It is in the depths of manifold theory that the 2-disk reenters the story. I began with the Riemann mapping theorem, skipped dimension = 3 permanently (the fundamental technical tool in three-manifold topology, Dehn's lemma—loop theorem—is a theorem for imbedding two-dimensional disks, however, an entire hour will be devoted to three-manifolds in a later lecture) and dimension = 4 temporarily, and we are now discussing the tools of high dimensional ( $n \geq 5$ ) smooth (or P.L.) manifold topology. This theory does more than help solve the annulus problem, but in this lecture we are oblivious to the rest.

We need to construct  $k$ . The method, rather odd at first glance, is to stick some P.L. manifold  $W$  in between  $T^n_\Gamma$  and  $T^n_{\text{Standard}}$  and then to try to simplify  $W$  so that the two inclusions of boundary components  $T^n_\Gamma \rightarrow W$  and  $T^n_{\text{Standard}} \rightarrow W$  are (simple) homotopy equivalences. Then one establishes a P.L. product structure on  $W$ . Following the product structure from bottom to top would give  $k$ . The process of simplification is called surgery. The construction of product structures is  $s$ -cobordism theorem. See Figure 6.

Surgery began with J. Milnor's discovery of new differentiable structures on the seven-sphere  $S^7$  and was extensively developed by the mid-1960s through the work of J. Milnor, M. Kervaire, W. Browder, S. P. Novikov,

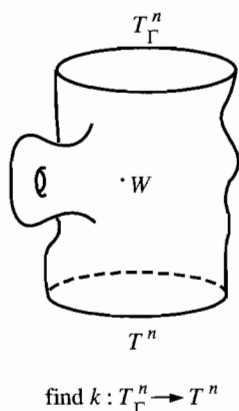


FIGURE 6

C.T.C. Wall, and others (for more details see the books of W. Browder and C.T.C. Wall [Br, Wa]). It is an obstruction to surgery on  $W$  that necessitates the passage to a finite cover.

The  $s$ -cobordism theorem was developed in the simply connected setting (where the letter  $h$  replaces  $s$ ) by S. Smale in 1959. It remains the most powerful method in topology for constructing isomorphisms between manifolds. D. Barden, B. Mazur, and J. Stallings worked out the obstructions which arise in the nonsimply connected setting. (These vanish for the fundamental group of an  $n$ -torus.)

In both of these major developments the 2-disk plays a key role in fitting geometry to algebra. The process is called the *Whitney trick* after H. Whitney's use of it [W] to construct imbeddings of  $n$ -manifolds in  $R^{2n}$ . See Figure 7 on next page.

In surgery theory manifolds are changed by manipulating spheres disjointly imbedded in them. The imbedding and disjointness information (when  $n$  is even and the spheres have dimension  $= n/2$ ) arrives in algebraic form: a total number of crossing points sums to zero. It must be converted into geometric information (disjointness and imbeddedness) by pushing portions of spheres across *Whitney disks* which pair crossings of opposite sign. In the  $h$ -cobordism theorem, the bubbling bouncing flow of a "gradient-like" vector field must be shifted and simplified to the greatest extent consistent with homology. This is also accomplished by standard moves guided by two-dimensional Whitney disks.

Most (but not all) of standard high-dimensional topological theory can now be brought down to dimensions  $n = 4$ . Quinn's proof of the annulus conjecture is a prime example of this. A crucial step was finding an imbedding theorem for 2-disks—Whitney 2-disks—to aid in simplifying  $s$ -cobordisms. Technically imbedded disks are not enough. The Whitney disk guides an isotopy and transverse coordinates are needed to write it down. So what is

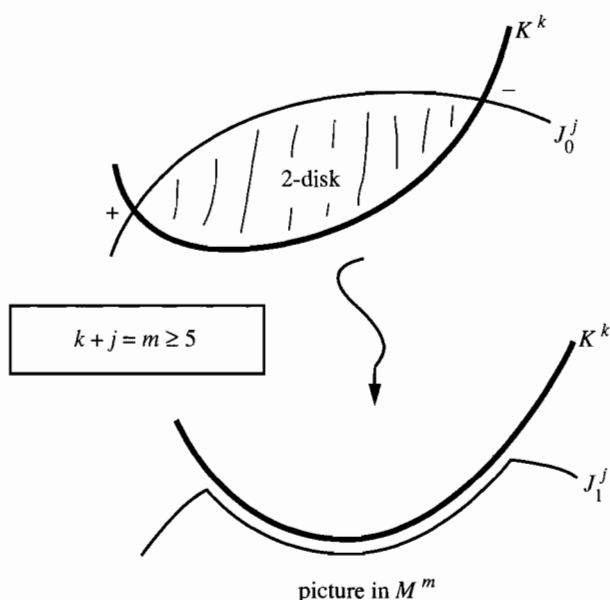
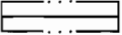


FIGURE 7



FIGURE 8

sought is a theorem for imbedding 2-handles  $H = (D^2 \times R^2, \partial D^2 \times R^2)$  when presented with a Whitney problem arising from surgery or in a five-dimensional  $s$ -cobordism.

In 1974 A. Casson found these handles [Cas] (in many important cases where  $\pi_1 \cong 0$ ) but they looked like Figure 8, and not like . His infinite construction gave smooth manifolds which we denote  $CH$  for "Casson Handle" which plausibly might be homeomorphic or diffeomorphic to  $H$ .

My contribution (in 1981) was to recognize (any)  $CH$  as homeomorphic to  $H$  by finding a common quotient:

$$H \xrightarrow{\alpha} CH/\mathcal{D} \xleftarrow{\beta} CH.$$

The projection  $\alpha$  is shown to be approximable by homeomorphisms by a difficult shrinking argument in the spirit of R. H. Bing with essential details supplied by R. Edwards, as explained in my paper [F]. The projection  $\beta$  is only known indirectly but, by the first step, its quotient is well in hand (i.e., homeomorphic to  $H$ ). Following the spirit of M. Brown's Schoenflies argu-

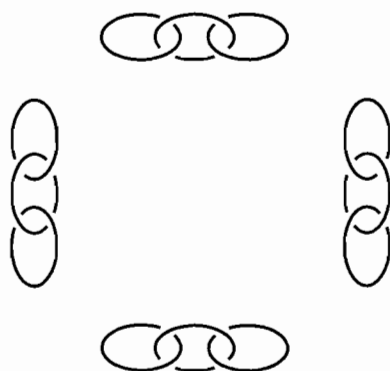


FIGURE 9. \$1 per opening-closing.

ment  $\beta$  is also shown to be approximable by homeomorphisms. At this point our story has come to a full circle, the two reactions to Alexander's horned sphere—one leading to  $\alpha$  and the other to  $\beta$ —are united in establishing  $H \cong CH$ , with the outcome joining usefully with the waiting machinery of manifold theory. But neither of these was to be the most surprising confluence.

The next year (1982) it became evident from S. Donaldson's work [Do] that at least many  $CH$ , though homeomorphic, were not diffeomorphic to  $H$ . From this came exotic structures on  $H$  and then  $R^4$  and a whole world of four-dimensional subtlety.

In the last few years, it has been necessary to look back to the most resistant settings where the starting material, Casson Handles, have not been found—and may not exist. An algebraic obstruction has been formulated in terms of Poincaré transversality [F'] and may be studied using the secondary theory of link invariants. This has been started by X.-S. Lin and me [FL] but talking about it is too much work. However, whether or not you can find disks where you want them, you can always play games on them. To set the mood, consider the necklace puzzle in Figure 9.

A topologist presents the above fragments to a jeweler and asks that they be hooked up into a necklace. He says: "I'll charge you one dollar for each link I must open and close so that will be four dollars, please." The topologist, of course, sees how to do it for three dollars.

In another game, suppose we consider a barber pole shear of infinite cylinder:

$$s_k: R \times S^1 \rightarrow R \times S^1 \\ (t, \theta) \rightarrow (t, \theta + kt).$$

The map  $s_k$  is linear and given by the matrix  $M = \begin{vmatrix} 1 & k \\ 0 & 1 \end{vmatrix}$ . The largest characteristic value of  $M$  (eigenvalue of  $\sqrt{M^T M}$ ) measures the factor by

which  $s_k$  can distort distance. For  $k$  large this eigenvalue is quite close to  $k$ .

Suppose (at some rather specialized place of business) that for a charge of \$1 any homeomorphism of distortion roughly 10 or less can be performed on  $R \times S^1$ . How much does it cost to make  $s_{10^4}$ ?

Well,

$$\begin{vmatrix} 1 & 10 \\ 0 & 1 \end{vmatrix}^{10^3} = \begin{vmatrix} 1 & 10^4 \\ 0 & 1 \end{vmatrix},$$

so maybe \$1,000. However,

$$\begin{vmatrix} 1 & 0 \\ 0 & 1/10 \end{vmatrix} \begin{vmatrix} 1 & 10 \\ 0 & 1 \end{vmatrix}^3 \begin{vmatrix} 1 & 0 \\ 0 & 10 \end{vmatrix} = \begin{vmatrix} 1 & 10^4 \\ 0 & 1 \end{vmatrix},$$

so  $S_{10^4}$  can be realized for five dollars.

The analytic theory (the Beltrami equation) for conformal distortion finds an even more graceful version of this factoring trick. The distortion discretely follows a geodesic in the Poincaré upper half-plane from  $(1, 10^4)$  to  $(1, 0)$ , the upper half-space being the Tiechmüller space for  $R \times S^1$  relative to its ideal boundary.

There is a conformal isomorphism

$$\begin{aligned} R \times S^1 &\xrightarrow{e^{-t}} U \setminus \{0\} \\ (t, \theta) &\mapsto (e^{-t}, \theta) \end{aligned}$$

and conjugating  $s_k$  by  $e^{-t}$  sends it to the logarithmic spiral  $\bar{s}_k(\rho, \theta) = (\rho, \theta + k \log \rho)$ . Thus we may see explicitly how logarithmic spirals can be quickly (in fact logarithmically) factored into compositions of quasiconformal maps of smaller conformal distortion. It is a joint result with Z.-X. He [FH] that no such rapid factoring exists for  $\bar{s}_k$  on  $D^2$  when small metrical distortion of the factors is required. Our result, in this example, says that if,  $s_{10^4}$  is to be written as a composition of  $n$  factors, each of which produces a metrical distortion of less than or equal to the distortion of  $s_{10}$ , then  $n \geq 996$ . One wonders if  $n$  must actually be  $\geq 1,000$ . The general problem, in which no real progress has yet been made, is to understand the behavior of metrical distortion on the 2-disk under composition and factoring. For example, it appears not to be known that a  $K$ -quasi-isometry of  $D^2$  can be factored into a composition of  $L$ -quasi isometries for any constant  $L$  which is smaller than  $K$ .

## REFERENCES

- [A] J. Alexander, *An example of a simply connected surface bounding a region which is not simply connected*, Proc. Nat. Acad. Sci. U.S.A. **10** (1924), 8–10.
- [B] William Browder, *Surgery on simply connected manifolds*, Springer-Verlag, New York, 1972.
- [BI] R. H. Bing, *A homeomorphism between the 3-sphere and the sum of two solid horned spheres*, Ann. of Math. (2) **56** (1962), 354–362.

- [Br] M. Brown, *A proof of the generalized Schoenflies theorem*, Bull. Amer. Math. Soc. **66** (1960), 74–76.
- [Car] C. Carathéodory, *Über die gegenseitige Beziehung der Ränder bei der konformen Abbildung der Inneren einer Jordanschen Kurve auf einen Kreis*, Math. Ann. **73** (1913), 305–320.
- [Cas] A. Casson, *Lectures on new infinite constructions in 4-dimensional manifolds*, Notes by Guillou, Orsay, 1974.
- [Da] R. Daverman, *Decomposition of manifolds*, Academic Press, 1986.
- [Do] S. Donaldson, *An application of Gauge theory to 4-dimensional topology*, J. Differential Geom. **18** (1983) 279–314.
- [F] M. H. Freedman, *The topology of four-dimensional manifolds*, J. Differential Geom. **17** (1982), 357–453.
- [F'] ———, *Poincaré transversality and 4-dimensional surgery*, Topology **27** (1988), 171–176.
- [FH] M. H. Freedman and Z.-X. He, *Factoring and logarithmic spiral*, Invent. Math. **92** (1988), 129–138.
- [FL] M. H. Freedman and X.-S. Lin, *On the  $(A, B)$ -slice problem*, Topology (to appear).
- [K] R. Kirby, *Stable homeomorphisms and the annulus conjecture*, Ann. of Math. (2) **89** (1969), 575–582.
- [M] B. Mazur, *On embedding of spheres*, Bull. Amer. Math. Soc. **65** (1959), 59–65; M. Morse, *A reduction of the Schoenflies extension problem*, Bull. Amer. Math. Soc. **66** (1960), 113–115.
- [OT] W. F. Osgood and E. M. Taylor, *Conformal transformations on the boundaries of their regions of definition*, Trans. Amer. Math. Soc. **14** (1913), 277–298.
- [Q] F. Quinn, *Ends of maps III: dimensions 4 and 5*, J. Differential Geom. **17** (1982), 503–521.
- [S] D. Sullivan, *Quasiconformal homeomorphisms and dynamics. I*, Ann. of Math. (2) **122** (1985), 401–418.
- [W] H. Whitney, *The self-intersections of a smooth  $n$ -manifold in  $2n$ -space*, Ann. of Math. (2) **45** (1944), 220–246.
- [Wa] C.T.C. Wall, *Surgery on compact manifolds*, Academic Press, New York, 1970.

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA AT SAN DIEGO, LA JOLLA, CALIFORNIA 92093-0112

## The Incompleteness Phenomena

HARVEY FRIEDMAN

The incompleteness phenomena have been a principal topic of research of the foundations of mathematics since the work of Kurt Gödel in the 1930's. Incompleteness refers to the following property of most, but not all, formal systems (i.e., set of axioms and rules of inference): that there remain sentences expressed within its language that are neither provable nor refutable within that formal system. Such a sentence is said to be *independent* of the given formal system. The incompleteness phenomena discussed here are distinguished by the variety of mathematical contexts and levels of abstraction represented by the independent sentences, as well as the scope or strength of the formal systems from which the sentences are independent.

To put the incompleteness phenomena in some historical perspective, note that two of the most celebrated revelations in the history of mathematics can be couched in its terms. The irrationality of  $\sqrt{2}$  corresponds to the fact that  $(\exists x)(x^2 = 2)$  is independent of the order field axioms, and the existence of non-Euclidean geometries corresponds to the independence of the parallel postulate from a suitable formal system for Euclidean geometry in which the parallel postulate is not present.

However, the incompleteness phenomena in the modern sense of the term, relates to formal systems surrounding those strictly mathematical concepts that are currently viewed as the basic notions from which all others are defined. Thus the focus has been on formal systems for natural numbers, and for sets, and also for restricted concepts of set.

The modern incompleteness phenomena obviously have the potential for forcing a reassessment of the foundations of mathematics. However, such a forced reassessment by the mathematics community has not occurred, despite the presently known incompleteness phenomena. We give a brief indication of why this is so.

The currently accepted foundation for mathematics is in terms of the formal system referred to as Zermelo Frankel set theory with the axiom of

---

1980 *Mathematics Subject Classification* (1985 Revision). Primary 03E35; Secondary 03C62, 03D35.

This research was partially supported by NSF.

©1992 American Mathematical Society  
0-8218-0167-8 \$1.00 + \$.25 per page



choice, abbreviated ZFC. This system seems to contain all of the easily identified and intuitive axioms about sets that stem from the usual explanation (story) about the universe of sets (as represented by the so-called cumulative hierarchy). The axiom of choice is a bit of a sore point in that, unlike the other axioms, it unavoidably asserts the existence of sets without naming them explicitly in terms of given sets. But the axiom of choice still has a pretty reasonable story.

It seems that there are no such additional axioms meeting these stringent criteria. Unfortunately, at present there is no theorem to this effect. However, as we discuss later, there certainly are additional proposed axioms, but the stories are comparatively far-fetched. Even for one of the most mild of all such proposed axioms—that of the existence of inaccessible cardinals—the story is already pretty strained.

All but one of the axioms of ZFC hold in the universe of (hereditarily) finite sets. The exception is of course the axiom of infinity. In fact, if one writes down all the obvious natural and intuitive principles about the universe of hereditarily finite sets that do not directly contradict the axiom of infinity, you wind up with a system (more or less) equivalent to ZFC without the axiom of infinity. Thus, in some profound sense that is not yet understood, the usual axioms for set theory (ZFC) are a straightforward adaptation of the usual axioms for finite set theory to an infinite context.

On a more general note, the entire issue of what constitutes an axiom about, say, natural numbers or sets, versus what constitutes merely a fact, is shrouded in mystery. It seems clear that induction is an axiom about the natural numbers, yet “ $x^3 + y^3 = z^3$  fails universally” is a fact and not an axiom.

Further evidence that mathematicians are complacently satisfied with ZFC as the foundation for mathematics (to the extent that they think about foundations) is that every proof put forth in mathematics to date by mainstream mathematicians is straightforwardly formalizable in even small portions of ZFC, with rather rare, minor, and easily removable exceptions. On those rare occasions when mathematicians use something outside of ZFC—most commonly the continuum hypothesis—they state the outside assertion as a *hypothesis* to the theorem in question, thus staying within ZFC.

On the other hand, the continuum hypothesis *is* independent of ZFC [Go1, Co]. Furthermore, the continuum hypothesis is just about the most basic and fundamental question that can be raised in the context of set theory. In fact, the question was raised by Georg Cantor early in his initiation of set theory, and appears as the first problem on Hilbert’s famous problem list.

So why has the independence of such a fundamental question not caused a crisis in the foundations of mathematics, and rendered ZFC obsolete?

We believe that the fundamental reason is the relative intellectual distance from the continuum hypothesis to finitary problems in mathematics.

To make the point in an extreme way, suppose that, instead of the continuum hypothesis, the twin prime conjecture (or some similar question about the infinitude of the prime pairs) was shown to be independent from ZFC. The mathematical community would be thrown into a foundational crisis. If, as is likely, such a result would be accompanied by a proof (of the conjecture about prime pairs) from some understandable extension of the ZFC axioms, then great interest would attach to the question of whether the additional axioms should be adopted. Having different answers or no answers to such questions about prime pairs according to which extensions of the ZFC axioms you postulate, would be regarded as wholly undesirable, and a uniform response would be widely sought.

It is not simply a matter of the independent sentence being about finite objects such as natural numbers. Thanks to Kurt Gödel, we already know that for any system such as ZFC there are sentences which are independent and which are, in a sense, even more finitary than twin prime conjectures. In fact, the consistency of ZFC itself is one such. (The latter result, known as the Gödel second incompleteness theorem, has terribly profound meaning for the foundations of mathematics in another direction.) Furthermore, it is also known that in systems such as ZFC there are always Diophantine equations over the integers such that the existence of solutions is independent of ZFC, using work of Gödel and Matijacevic [Mat], yet any such known example is truly gargantuan in size. *It is clearly also a matter of subject matter.*

To encapsulate: The continuum hypothesis is too infinitary. The consistency of ZFC is not a basic mathematical question (though it is a basic metamathematical or logical question). No remotely reasonable Diophantine equation is anywhere near being shown to display any kind of independence. And no twin prime conjecture is anywhere near being shown to be independent of ZFC.

The fundamental issue is this: Is there a basic mathematical problem about standard finite objects such as, say, natural numbers or rational numbers or polynomial rings over finite extensions of the rationals, etc., with a clear and intuitive meaning, conveying interesting mathematical information, that is readily graspable, and which is independent of ZFC?

We speculate that sometime during the twenty-first century, someone will answer the above question in the affirmative, and there will be nearly universal agreement in the mathematics community that this has been accomplished. Furthermore, there will be proofs of such mathematical problems accompanying such independence results using some of the extensions of ZFC that have already been explored in the set theory community. The current state of the art regarding this conjecture is discussed in the Appendix.

Before beginning the detailed discussion of various incompleteness phenomena, we give an informal description of the axioms of ZFC for the reader's convenience.

In ZFC, every object is a set, and the primitive relations between sets are that of equality and membership.

Informally, the axioms are as follows.

- (1) *Extensionality*. Two sets are equal if and only if they have the same elements.
- (2) *Pairing*. For any two sets  $x, y$ ,  $\{x, y\}$  exists.
- (3) *Union*. For any set  $x$ ,  $\bigcup x$  exists, which is the set of all elements of elements of  $x$ .
- (4) *Power set*. For any set  $x$ , the power set  $\mathcal{P}(x)$  exists, which is the set of all subsets of  $x$ .
- (5) *Separation (comprehension)*. For any set  $x$ ,  $\{y \in x : A(y)\}$  exists, where  $A$  is any set theoretically describable predicate;  $A$  is allowed to mention specific sets called *parameters*.
- (6) *Infinity*. There are many equivalent forms this axiom can take but the following is customary: There is a set  $\omega$  which contains the empty set  $\emptyset$  as an element, and for every  $x \in \omega$ , we have  $x \cup \{x\} \in \omega$ .
- (7) *Axiom of choice*. Again there are many equivalent forms this can take, and the following is customary: For every set of pairwise disjoint nonempty sets, there is a set which meets each of these nonempty sets at exactly one place.
- (8) *Replacement*. This asserts that for any set  $x$  and any function from  $x$  into sets that is set-theoretically described (with parameters as in 5 above), the range of this function exists as a set.
- (9) *Foundation*. Every nonempty set possesses an  $\varepsilon$ -minimal element, i.e., an element which is disjoint from the given set.

Mathematicians seldom use axioms (8) or (9).

Some good works on set theory include [Je2] and [Levy].

An important line of research that goes in a different direction than that emphasized here is the work on the projective hierarchy of sets (of real numbers). This hierarchy begins with the Borel sets and then grows upward in complexity through the operations of projection and complementation. The goal is to understand the properties and structure of the projective sets as thoroughly as we understand the Borel sets (in the sense of classical descriptive set theory). After a couple of levels or so in the hierarchy, we know that ZFC is not sufficient to do anything interesting along these lines. However, under the additional axiom of constructibility (or in the constructible universe of sets), all appropriate questions about the projective sets are answered. Alternatively, using axioms for large cardinals, again all appropriate questions are answered, but typically with different answers. See [Mar] and [MS].

**1. General incompleteness phenomena.** These are the incompleteness properties that apply to a very wide class of formal systems. The major such results are due to Kurt Gödel [Go3, Smo]:

**FIRST INCOMPLETENESS THEOREM.** *In any formal system such as ZFC, there always are sentences which are neither provable nor refutable.*

**SECOND INCOMPLETENESS THEOREM.** *In any formal system such as ZFC, the consistency of the system itself is not provable in the system.*

For the first incompleteness theorem, we need to know that the system is effectively axiomatized, is consistent (i.e., free of contradiction), and contains a small amount of basic integer arithmetic.

For the second incompleteness theorem, we additionally need to know that provability in the system can be adequately formalized within the system. In fact, the modern formulations of the theorem assert that no adequate formalization of the consistency of the system will itself be provable in the system, and give particular families of adequate formalizations, which include usual intuitively based ones.

As spectacular as these results are, they do not provide examples of mathematically motivated problems which cannot be proved or refuted in ZFC, as discussed in the introduction here. This came later.

But these results did put an end to Hilbert's program, one of whose goals was to secure the consistency of mathematics within weak principles of integer arithmetic.

However, one of the most interesting of all the *completeness* phenomena is the result of [Tarski] that the axioms of real closed fields are complete. These axioms augment the ordered field axioms by the axioms which assert that every single variable polynomial of odd degree (with leading coefficient 1) has a zero, and every positive element has a square root. The system is effectively axiomatized, and so appears to violate the first incompleteness theorem. However, note that the real closed field axioms do not contain basic integer arithmetic (only real number arithmetic).

Another general incompleteness phenomenon is obtained by combining work of Gödel and Turing with the result of [Mat] on the nonrecursiveness of the solvability of Diophantine equations over the integers (Hilbert's 10th problem). Also see [DMR]. The following result is from the folklore:

**THEOREM.** *In any formal system satisfying the usual hypotheses for the first incompleteness theorem such as ZFC, there always is a Diophantine equation over the integers which is unsolvable, yet cannot be proved to be unsolvable in the system.*

As mentioned in the introduction, this theorem does not give any reasonable example of such a Diophantine equation. It is an open question whether for any system such as ZFC, there is such a Diophantine equation that can be written down on a page or so with all coefficients and exponents written out in base 10.

**2. Independence results via forcing.** The forcing method was introduced in [Co] and provides a general method for obtaining new models of ZFC from given ones by adjoining new objects. The resulting new models will in

general satisfy different sentences than the given models, and, therefore, one can show that certain sentences are independent of ZFC in this way.

More specifically, in the modern treatment of forcing one begins with a countable model of ZFC, called the ground model. One chooses a partially ordered set from this ground model, called the "notion of forcing." Then one defines a family of sets called the generic sets with respect to the notion of forcing. If the notion of forcing satisfies some minimal conditions, then there are continuum many such generic sets. Then one proves that if any generic set is adjoined to the ground model, the resulting model is a model of ZFC. Models of ZFC obtained in this way are called generic extensions. Furthermore, there are methods which are useful in determining whether sentences hold in generic extensions in terms of whether related sentences hold in the ground model about the notion of forcing. To apply the method in order to show that some particular sentence of interest is consistent with ZFC, one chooses a suitable ground model and delicately adjusts the notion of forcing in order that the generic extensions satisfy the given sentence.

The method has been extensively developed, streamlined, and unified by set theorists since [Co], and has been quite successful in establishing independence results of a certain kind. However, the method has inherent limitations, particularly for establishing the independence of sentences of more than a certain level of concreteness as we presently indicate.

Firstly, every generic extension has the same ordinals as the ground model. In particular they have the same integers, and in fact the same basic arithmetical operations. From this it is clear that any sentence about the natural numbers must hold in the generic extension if and only if it holds in the ground model. The same assertion holds for any sentence about finite objects. Therefore we cannot hope to establish the independence of any sentence about finite objects through (at least a direct application of) the method of forcing.

Secondly, every generic extension also has the same functions on ordinals defined by transfinite recursion as the ground model. It is known that every sentence about natural numbers and (possibly infinite) sets of natural numbers that is not too complicated can be reduced in ZFC to a corresponding sentence about ordinals and functions on ordinals defined by transfinite recursion. Therefore, in the same way as we saw in the previous paragraph, we cannot hope to establish the independence of any sentence about finite objects and sets of finite objects at or below a certain level of complexity, through (at least a direct application of) the method of forcing. The precise result we are using is that every  $\Pi^1_3$  sentence that holds in the generic extension must also have held in the ground model, and so one cannot prove the consistency of a  $\Pi^1_3$  sentence directly through the method of forcing (see, e.g., [Je2, pp. 530–531]).

The method of forcing in its original form was also used to obtain models of ZF in which the axiom of choice fails. This aspect of the forcing method

has not been given as neat a streamlining and unification as it has in the context of ZFC models. Nevertheless, most of the important independence results about ZF without the axiom of choice can be proved by constructing appropriate generic extensions satisfying ZFC, and then taking a suitable submodel to obtain the desired model satisfying ZF in which the axiom of choice fails.

The first application of the forcing method was the independence of the continuum hypothesis which we state as follows:

**PROPOSITION 2.1.** *Every uncountable set of real numbers is in one-one correspondence with all of the real numbers.*

The consistency of the above with ZFC is from [Go1] and the consistency of the negation of the above with ZFC is from [Co].

Another application is to Souslin's hypothesis which we formulate as follows. The consistency is from [ST] and the consistency of the negation is from [Je1]:

**PROPOSITION 2.2.** *Every nonseparable linearly ordered set has an uncountable subset in which every element is isolated.*

Another application is to Whitehead's group conjecture, which has been proved in the case of countable groups (see [Fu]). The independence is from [Sh]:

**PROPOSITION 2.3.** *If  $\text{Ext}(G, \mathbb{Z}) = 0$  then  $G$  is free (for Abelian  $G$ ).*

Another application is to Kaplansky's conjecture which we state as follows.

(For the consistency of the negation, see [Dales], where the negation is actually proved from the continuum hypothesis. For the consistency, see [DW].)

**PROPOSITION 2.4.** *Any homomorphism from the Banach algebra  $C[0, 1]$  into any (separable) Banach algebra is continuous.*

Yet another application is to a generalization of Fubini's theorem, in which the hypothesis of two-dimensional measurability is relaxed. The consistency of its negation is clear since the generalization is refutable using the continuum hypothesis (folklore). For the consistency of this and other strengthenings see [Ship].

**PROPOSITION 2.5.** *If  $F : [0, 1]^2 \rightarrow [0, 1]$  has almost all  $F_x, F^y$  measurable, then  $\int(\int F(x, y) dx) dy = \int(\int F(x, y) dy) dx$ .*

All of the above examples should be regarded as set-theoretic in that they involve unrestricted selections from uncountable domains. For example, in the statement of the continuum hypothesis above, we refer to arbitrary sets of real numbers. We can be more specific about the kinds of sets of real numbers to be considered by imposing a regularity condition. The most

common regularity conditions on subsets of separable metric spaces are that of measurability, Borel measurability, and various strengthenings of Borel measurability. Measurability works well in contexts in which measure 0 sets are regarded as equivalent. If they are not regarded as such, then obviously one really does not have a regularity condition per se, since the measure 0 sets are as badly behaved as arbitrary sets.

From the logical point of view, Borel measurability makes sense as a kind of minimal regularity condition to impose. The Borel measurable subsets of complete separable metric spaces form a very wide class of objects for the vast majority of mathematical purposes and in an appropriate sense constitute (or at least include all) those subsets which are constructed via sequential processes.

Thus, in essence, the imposition of the regularity condition of Borel measurability removes mathematically undesirable and irrelevant pathology from the context. In §4, we explore the effect this “Borel point of view” has on the incompleteness phenomena.

We mention an independence result involving set-theoretically definable sets of real numbers. The consistency of the negation with ZFC follows from [Go1, Go3], and the consistency with ZFC is from [Sol]:

**PROPOSITION 2.6.** *Every cross section of every definable set of real numbers is measurable and has the property of Baire. If the cross section is uncountable then it has a perfect subset.*

We can consistently add measurability, the Baire property, and uncountability implies perfect subsets for *all* sets to ZF (i.e., ZFC without the axiom of choice). But this is not so interesting without also having some choice. Fortunately, in [Sol] dependent choice is added to ZF for this result, which is enough choice to prove the basic facts about measurability, the property of Baire, and uncountability.

**3. The constructible point of view.** Needless to say, the incompleteness phenomena involving ZFC are not desirable features of the commonly accepted foundation for mathematics. It is natural to explore possible remedies for the situation short of overhauling ZFC.

We have already hinted at one possible remedy which will be explored in the next section. That is the remedy of imposing the regularity of Borel measurability on the objects considered. Of course, this does not make the original sentences any less independent of ZFC than they were before, but it does give a general process for removing the offending pathology that might be responsible for the difficulties while preserving the essential mathematical content of the original sentences.

In this section we explore a different remedy. We consider the effect of placing a general regularity condition of a logical nature on the set concept itself.

The usual modern description of the universe of sets is in terms of the

cumulative hierarchy. This hierarchy associates a family of sets to every ordinal  $\alpha$ . The sets are just the sets that appear somewhere in the hierarchy. Of course there is a circularity here since it is also customary to define the ordinals as certain kinds of sets, but this is usually ignored since the hierarchy is used as an informal description to motivate the axioms of ZFC.

This cumulative hierarchy is given as follows:  $V_0 = \emptyset$ ,  $V_{\alpha+1} = \wp(V_\alpha)$ , and  $V_\lambda = \bigcup_{\beta < \lambda} V_\beta$  for limit ordinals  $\lambda$ . Here  $\wp$  stands for the power set operation—the family of all subsets of the set to which it is being applied. It is provable in ZFC that every set appears somewhere in this hierarchy. The class of all sets is denoted by  $V$ .

Now observe that there are really two quite different operations that drive this cumulative hierarchy. One is the ordinals and the process of transfinite recursion, and the other is the power set operation.

From the constructible point of view it is the power set operation that is suspect. All objects should be constructed on the basis of some general form of transfinite recursion, where “events take place on the basis of earlier events.” In this sense, the power set operation must be derived from something more fundamental; every set that exists must be constructed from earlier constructed sets in some way.

The constructible point of view originated with Kurt Gödel in his proof of the consistency of the continuum hypothesis, where he introduces the so-called constructible hierarchy. Although he did briefly hold at least some variant of the constructible point of view, he quickly renounced it in favor of a strongly Platonist point of view now common among specialists in set theory (see [Go2]).

The constructible hierarchy is given as follows:  $L_0 = \emptyset$ ,  $L_{\alpha+1}$  = the set of all subsets of  $L_\alpha$  that are explicitly definable over  $L_\alpha$  (allowing parameters for elements of  $L_\alpha$ ), and  $L_\lambda = \bigcup_{\beta < \lambda} L_\beta$ . The class of all constructible sets (i.e., sets that appear somewhere in this hierarchy) is denoted by  $L$ .

$L$  has many desirable properties. Within ZFC, we can prove that  $L$  obeys all of the axioms of ZFC. We can prove this even within ZF. This latter fact is what allowed Gödel to conclude that the axiom of choice was consistent with ZF. Put somewhat differently:  $L$  obeys the axiom of choice for a good reason, whereas  $V$  obeys it by conventional wisdom.

If the constructible hierarchy is modified in small ways, then we still provably get the same class of sets.

If we start with the constructible hierarchy as the point of departure in motivating the axioms of ZFC, we can use the same story that we use for motivating ZFC from the cumulative hierarchy (in fact, the story for the axiom of choice is much improved), except for the power set axiom. This is not surprising since the power set axiom is explicitly part of the mechanism of the cumulative hierarchy. However, reasonable extensions of the story for the replacement axiom in the constructible hierarchy can be given which will motivate the power set axiom in the constructible hierarchy. To more



fully clear up the philosophical issues here we need to develop an appropriate general theory of transfinite iteration which applies to contexts much more general than set theory. It is likely that this can be done.

Regardless of these philosophical niceties, we can now effectively regard constructibility (i.e., membership in  $L$ ) as a kind of regularity condition on sets.

But what happens to our mathematical proofs if we restrict all mathematical objects to constructible objects?

From what we have said above, it is clear that if we start with a proof in ZFC, then the result of this uniform restriction to  $L$  is still a proof in ZFC. We just have to attach proofs of the  $L$ -restricted forms of the axioms of ZFC that are used in the original proof; these attached proofs can be themselves given in ZFC. And since these attached proofs have already been given by Gödel in [Go1], there is no need for anyone to do anything other than what they are doing now.

Now that it is clear that restricting to constructible sets is not of any real operational consequence for mathematicians (other than some set theorists who operate outside of ZFC), conceptually speaking how much of a restriction is constructibility?

It follows from what has been said above that it is consistent with ZFC that  $V = L$ , i.e., all sets are constructible. Thus there is no way to construct a nonconstructible set within ZFC. This effectively reduces the level of restriction for mathematicians other than some set theorists to nil.

On the other hand, what is the advantage of everybody simply declaring that they are using only constructible sets, functions, numbers, etc.?

The advantage is that if, e.g., Propositions 2.1–2.6 are reinterpreted as being about constructible sets (functions, and numbers, etc.), then the independence results associated with them disappear.

More specifically, the following is proved in [Go1]:

**THEOREM 3.1.** *Proposition 2.1 holds in the constructible universe.*

The following is due to [Jensen] (and see [Je2, pp. 226–229]):

**THEOREM 3.2.** *Proposition 2.2 fails in the constructible universe.*

The following is proved in [Sh]:

**THEOREM 3.3.** *Proposition 2.3 holds in the constructible universe.*

For the following see [Dales] since the continuum hypothesis holds in  $L$ :

**THEOREM 3.4.** *Proposition 2.4 fails in  $L$ .*

The following is a consequence of the continuum hypothesis holding in  $L$ :

**THEOREM 3.5.** *Proposition 2.5 fails in the constructible universe.*

The following is proved in [Go1, Go3]:

**THEOREM 3.6.** *Proposition 2.6 fails in the constructible universe.*

In fact, virtually all sentences that have been proved to be independent from ZFC by a direct application of the forcing method have now been decided when restricted to the constructible universe.

The axiom of constructibility asserts that  $V = L$ , i.e., all sets are constructible. Although it is not as obvious as it sounds, Gödel proved that the axiom of constructibility holds in the constructible universe.

From a purely operational point of view, there is no functional difference between assuming the axiom of constructibility and deciding to restrict oneself to constructible sets only.

However, it is *not* a tenant of the constructible point of view that the axiom of constructibility is somehow evidently true, or is even true at all. This would be like saying that a mathematician who imposes the regularity condition of differentiability of functions in his work somehow believes that all functions are differentiable. Constructibility is merely intended to be a regularity condition.

This author is quite sympathetic to the constructible point of view. We would like to go even further. We believe that the usual description of the set-theoretic universe is not sufficiently clear to "determine" an answer to even such a set-theoretically fundamental question such as the continuum hypothesis. The unrestricted power set of infinite (and especially uncountable) sets become a vague blur when examined too intensely.

We also believe that the constructible point of view is not going to prove to be sufficiently powerful to avoid all of the foundational difficulties that we suspect will arise. In particular, it is obviously helpless in dealing with the status of sentences about finite objects since they are already constructible. It is also of no use in dealing with the status of not too complex sentences about sets of natural numbers ( $\Pi_2^1$  sentences) since they are provably equivalent to their restrictions to the constructible universe. In fact, we later discuss examples of sentences about Borel functions on groups and graphs which remains independent of ZFC even when restricted to the constructible universe.

**4. The Borel measurable point of view.** The Borel measurable point of view is based on a quite natural mathematical regularity condition.

We start with a complete separable metric space. The Borel measurable sets constitute the least  $\sigma$ -algebra containing the open sets. (Henceforth we omit the word "measurable.") The Borel functions are those functions for which the inverse image of every open set is Borel.

The Borel functions can be arranged in a tower of length  $\omega_1$  where we start with the class of continuous functions, and at every nonzero ordinal we take the class of all everywhere defined sequential limits of functions from the earlier classes. These concepts and this construction have obvious generalizations to the case of Borel functions between two spaces, and also Borel functions of finite or even countably infinitely many arguments (using

product constructions for metric spaces). It is natural to also consider partially defined Borel functions, which are merely the restrictions of Borel functions to Borel subsets of the relevant space. Recall that there is a one-one onto Borel function with Borel inverse between any two uncountable complete separable metric spaces.

Throughout mathematics, one works with structures in the sense of a nonempty set endowed with distinguished elements, relations, and (partial) functions. A structure is said to be a Borel structure if its domain is a Borel set of real numbers, and its relations and functions are all Borel. (Sometimes it is convenient to allow equality to be represented by a Borel equivalence relation.) Most of the important structures in mathematics are naturally isomorphic to Borel structures. Separable Banach spaces form a natural family of such structures.

The Borel point of view takes the position that all mathematical structures to be considered are Borel structures (or naturally equivalent to such) and all sets and functions to be considered are Borel sets and Borel functions in and between Borel structures.

How severe is such a regularity condition?

Upon examination, it appears not to be very restrictive. Virtually all of the more important and intensively studied mathematical structures are Borel and the same is true of particular examples of sets and functions.

The typical case of where one goes beyond Borel in mainstream mathematics is where one is developing a general theory, say of groups or fields. A lot of useful facts simply can be proved without restricting the algebraic objects in some nonalgebraic way, such as being Borel representable. However, if one imposes the Borel regularity condition, then the theory is not generally any easier, and no mathematical content is lost in the theory. In particular, all of the examples one normally applies the theory to that are of central interest are generally relatively concrete in nature and meet the Borel regularity condition.

What is the advantage of imposing Borel regularity conditions?

In many cases, sentences independent of ZFC have straightforward reinterpretations using Borel regularity. Typically, the resulting sentence is no longer independent of ZFC. This is the case for Propositions 2.1–2.6 as follows:

The following is implicit in, e.g., [Luzin]:

**THEOREM 4.1.** *Every uncountable Borel set of real numbers is in Borel one-one correspondence with the set of all real numbers.*

The following is proved in [HMS]:

**THEOREM 4.2.** *Every nonseparable Borel linear ordering of the reals has an uncountable Borel subset in which every element is isolated.*

The following is proved in [Sp] (that  $G$  is not Borel free is in the folklore):

**THEOREM 4.3.**  *$G$  is free, but  $G$  is not Borel free, where  $G$  is the group of bounded infinite sequences of integers.*

In the above, Borel free means that there exists a Borel set of independent generators. The connection with Proposition 2.3 is that a free group  $G$  always has  $\text{Ext}(G, \mathbb{Z}) = 0$ . However, we can still ask whether Proposition 2.3 is true for Borel groups, with the usual (non-Borel) notion of free group.

The following is proved in [Ajtai], but goes back to Laurent Schwartz:

**THEOREM 4.4.** *Any Borel homomorphism from one separable Banach space to another is continuous.*

The following is classical:

**THEOREM 4.5.** *Fubini's theorem for Borel functions from the square into itself.*

The following is classical (see, e.g., [Luzin]):

**THEOREM 4.6.** *Borel sets are measurable, have the property of Baire, and if uncountable have perfect subsets.*

We are also sympathetic to the Borel point of view. It raises an interesting issue as to the proper role of generality in mathematics.

The rest of the discussion of the incompleteness phenomena will almost exclusively focus on the independence of sentences that are admissible from the Borel point of view.

Many of the independence results discussed are not independence results from the full ZFC axioms, but rather from significant fragments of ZFC. In fact, these independence results from fragments of ZFC that are discussed here are in fact theorems of ZFC.

It is natural to inquire as to the significance of such independence results since mathematicians generally accept all of ZFC. We give two replies.

Firstly, as noted above, virtually all of mathematics done outside of set theory is easily formalizable in surprisingly weak fragments of ZFC. This immediately raises the important and interesting question of whether and to what extent the axioms of ZFC (beyond such weak fragments) are useful or relevant to mathematics.

Secondly, we are still very far from a really convincing mathematically basic and interesting example of a theorem of ZFC about finite objects which uses more than, say, the part of ZFC that applies to countable sets; e.g., such a theorem of ZFC which cannot be proved in ZFC with the power set axiom deleted. It is only since 1977 that we have had a pretty convincing such example which cannot be proved in finite set theory (ZFC without the axiom of infinity), and since 1981 that we have gone beyond significant parts of countable set theory. Such fragments of ZFC form significant barriers to progress towards ZFC and beyond, and also have intrinsic interest.

**5. Cantor's theorem and the discrete topology.** We now discuss a theorem from [Fr1] about Borel functions which arises from an examination of the proof of Cantor's fundamental theorem that the reals are uncountable. This is an example of a basic theorem about Borel functions whose proof rather noticeably must take one quite far from the context in which it is expressed. In particular its proof cannot be given in what may be called separable mathematics. In separable mathematics, all of the objects one works with are countable, or at least can be described completely in countable terms. This allows for complete separable metric spaces, since they can be specified by the restriction of the metric space to any countable dense set. Elements in complete separable metric spaces are also admissible since they can be specified by any sequence from the countable dense subset that converges to it. Continuous functions between complete separable metric spaces can be specified by their restriction to any countable dense set. Borel functions can be specified by the countable process from which they are built (Borel codes).

From the axiomatic point of view, an appropriate system that reflects the above conception of separable mathematics is obtained by deleting the power set axiom from ZFC. The resulting system is written as  $\text{ZFC} \setminus \wp$ . The (hereditarily) countable sets form a model of this system.

Cantor's theorem can be stated as follows. Let  $x_1, x_2, \dots \in I$ , where  $I$  is the closed unit interval. Then there exists  $y \in I$ ,  $y \neq x_1, x_2, \dots$ .

Standard methods for constructing Borel functions establish rather easily that there is a Borel  $F : I^\infty \rightarrow I$  such that for all  $x \in I^\infty$ ,  $F(x)$  is not a coordinate of  $x$ . For example, the following function obeys this property and is Borel: Take  $F(x) = \bigcap_n J_n$ , where each  $J_{n+1}$  is the first closed dyadic rational interval of length at most  $2^{-n}$  contained in  $J_n$  which is disjoint from  $\{x_1, \dots, x_n\}$ , and  $J_1 = I$ . (Any listing of the closed dyadic rational intervals will do for this construction.)

However, note that the value of  $F$  at a sequence may depend on the order in which that sequence is given, not just on the image of the sequence (even if multiplicities are counted). This leads to the following question: Is there such a Borel function which is permutation invariant, i.e., obeying  $F(\sigma x) = F(x)$ , for all permutations  $\sigma$ ?

The answer is no. The following is proved in [Fr1]:

**THEOREM 5.1.** *Every permutation invariant Borel  $F : I^\infty \rightarrow I$  sends some point to a coordinate of itself.*

The proof uses the topology  $(\mathbf{I})^\infty$ , where  $\mathbf{I}$  is the closed unit interval endowed with the *discrete* topology, i.e., the product of countably infinitely many copies of  $\mathbf{I}$ . The Baire category theorem can be stated and proved in this context. One can also prove a 0, 1-law for Baire category which states that every permutation invariant Borel subset of  $\mathbf{I}$  is meager or comeager. Since every Borel subset of  $\mathbf{I}$  (with the usual separable topology) is also a Borel subset of  $\mathbf{I}$ , we see that every permutation invariant Borel subset of  $\mathbf{I}$

is meager or comeager in the sense of **I**. One can then prove by standard techniques that there is a  $c \in I$  such that the given function  $F$  is constantly  $c$  on a comeager set in the sense of **I**. (So far we have not really used the nonseparability of **I**.) But by heavy use of the nonseparability of **I**, we see that comeagerly many  $x$  contain  $c$  as a coordinate in the sense of **I**. (The latter is false for Baire category or measure on the usual  $I$ ). Hence for at least one  $x$ ,  $F(x)$  is a coordinate of  $x$ .

The following is also proved in [Fr1]:

**THEOREM 5.2.** *Theorem 5.1 cannot be proved in  $\text{ZFC} \setminus \wp$ . Hence, in the appropriate sense, the theorem cannot be proved within “separable mathematics.”*

We sketch some of the ideas in this proof.

It suffices to prove that there is a model of  $\text{ZFC} \setminus \wp$  from the axioms of  $\text{ZFC} \setminus \wp$  together with Theorem 5.1. For then, if  $\text{ZFC} \setminus \wp$  were to prove Theorem 5.1, then  $\text{ZFC} \setminus \wp$  would prove the existence of a model of  $\text{ZFC} \setminus \wp$ , and hence by Gödel’s second incompleteness theorem,  $\text{ZFC} \setminus \wp$  would be inconsistent, which it is not.

Next, we introduce a system called second-order arithmetic, and written as  $Z_2$ . Despite its name, it is an ordinary first-order formal system like all of the ones we have been discussing. It has variables over natural numbers and over sets of natural numbers, contains the usual arithmetic of addition and multiplication, the axiom scheme of induction, and most importantly, the comprehension scheme which asserts that each  $\{n \mid \varphi(n)\}$  exists, where  $\varphi$  may mention numbers and sets of numbers as parameters, and have quantifiers over all numbers and over all sets of numbers. It is known how to build a model of  $\text{ZFC} \setminus \wp$  from a model of  $Z_2$  directly, and in particular within  $\text{ZFC} \setminus \wp$ . Models in general do not have to have only standard integers (they may have nonstandard ones), but if the original model of  $Z_2$  has only standard integers then the resulting model of  $\text{ZFC} \setminus \wp$  also has only standard integers.

Combining the above two paragraphs, we now see that it suffices to construct a model of  $Z_2$  with only standard integers using only  $\text{ZFC} \setminus \wp$  and Theorem 5.1.

There are still real difficulties in obtaining such a model relating to the parameters that are allowed in the comprehension axiom scheme above. So the crucial next step, carried out in detail in [Fr1], is the consideration of  $p - Z_2$ , which is the same as  $Z_2$  except no parameters are allowed in the comprehension scheme. It is shown in [Fr1] how to go from a model of  $p - Z_2$  with only standard integers to a submodel with only standard integers obeying  $Z_2$ . Again this construction can be done explicitly within  $\text{ZFC} \setminus \wp$ .

Combining the above three paragraphs, it is clear that it suffices to construct a model of  $p - Z_2$  with only standard integers using only  $\text{ZFC} \setminus \wp$  and Theorem 5.1.

Now to every sequence  $x \in I^\infty$  we can associate a family of sets of natural numbers  $M(x)$ , which can be viewed as an attempted model of  $p - Z_2$  with only standard integers. We can use any Borel correspondence of  $I$  with  $\wp(\omega)$  for this purpose, taking  $M(x)$  to be the image of  $x$  under this correspondence. Of course we can assume that  $M(x)$  is viewed as being equipped with numbers and arithmetic. Thus the only possible reason that  $M(x)$  does not satisfy the desired  $p - Z_2$  is that the parameterless comprehension axiom scheme might fail.

We now let  $F(x)$  be obtained by looking up the first instance of parameterless comprehension that fails in  $M(x)$  and taking the image of the missing set under the above chosen Borel correspondence to be  $F(x)$ . (If parameterless comprehension holds, i.e., if  $M(x)$  satisfies  $Z_2$ , then we are done anyway, but in this case let  $F(x)$  be 0 by default.)

Careful consideration of the construction of  $F$  reveals that it is a permutation invariant Borel function.

Applying Theorem 5.1, there is an  $x$  such that  $F(x)$  is a coordinate of  $x$ . Tracing through the construction of  $F$ , we see that the only way this can happen is the default case above, and hence  $M(x)$  must satisfy  $p - Z_2$  as desired.

The following two related results are proved in [Fr2]: Let  $K$  be the Cantor space consisting of the infinite sequences of 0's and 1's. The important shift map is given by  $s(x) = (x_2, x_3, \dots)$ , where  $x = (x_1, x_2, x_3, \dots)$ , i.e., shift deletes the first term. We say that  $F : K \rightarrow K$  is shift invariant if it obeys  $Fsx = Fx$ . We also let  $x^{(2)} = (x_1, x_4, x_9, \dots)$ .

**THEOREM 5.3.** *Every shift invariant Borel function  $F : K \rightarrow K$  is somewhere its "square," i.e., for some  $x$ ,  $F(x) = x^{(2)}$ .*

**THEOREM 5.4.** *Theorem 5.3 cannot be proved in  $ZFC \setminus \wp$ . Hence in an appropriate sense, it cannot be proved within "separable mathematics."*

Theorems 5.1 and 5.3 can be proved just beyond  $ZFC \setminus \wp$ . For example, if we add the existence of  $\wp(\omega)$  to  $ZFC \setminus \wp$ , then the resulting system is powerful enough to prove these two theorems.

Theorems 5.1 and 5.3 are examples of what we call Borel diagonalization theorems. Such theorems assert that there are no Borel diagonalization functions with certain invariance properties.

Looking at such theorems conversely, they illustrate the following general principle which we do not know how to formulate in anything like full generality:

**GENERAL THEME.** Every "invariant" Borel function from one "space" into another sends some element to a "simpler" element.

**6. Borel diagonalization on equivalence relations, linear orders, groups, and graphs.** In this section we present some more powerful Borel diagonalization theorems than the basic Theorem 5.1 above. They very clearly illustrate the General Theme.

These diagonalization theorems fall into three basic categories:

**CLASS A.** These are the theorems of ZFC which, like Theorems 5.1 and 5.3, can be proved just beyond countable set theory (e.g., in  $ZFC \setminus \wp + \wp(\omega)$  exists), but not within  $ZFC \setminus \wp$  (or separable mathematics).

**CLASS B.** These are the theorems of ZFC which can be proved just beyond ZC but not within ZC itself. For instance, they can be proved within systems such as  $ZC + V(\omega + \omega)$  exists, or  $ZFC \setminus \wp + V(\omega + \omega)$  exists. Here ZC is Zermelo set theory with the axiom of choice, which is obtained from ZFC by the removal of the replacement axiom scheme (and optionally, removal also of the formulation axiom).

**CLASS C.** These are the theorems of ZFC which can be proved using uncountably many iterations of the power set operation, but not using any (explicitly given) countable number of such iterations. In particular they cannot be proved within Zermelo set theory with the axiom of choice, ZC.

The phrase "iterations of the power set operation" needs some explanation. Recall the cumulative hierarchy as presented in §3. The stages in the hierarchy represent iterations of the power set operation. The ordinal number of the stage represents the number of iterations. Thus when we say that we have uncountably many iterations of the power set operation, we mean that we have, for each countable ordinal  $\alpha$ , the stage  $V_\alpha$ . The system ZC is easily seen to correspond to having  $\omega + \omega$  iterations of the power set operation, i.e., having each  $V_{\omega+n}$ , where  $n$  is finite.

We first consider a direct generalization of Theorem 5.1. Let  $E$  be any equivalence relation on  $I$ . We use  $[ ]$  for the equivalence classes under  $E$ . For  $S \subseteq I$ , we write  $[S]$  for  $\{[x] : x \in S\}$ .

We say that the Borel diagonalization theorem holds for  $E$  if there is no Borel function  $F : I^\infty \rightarrow I$  such that (a) if  $[\text{rng}(\bar{x})] = [\text{rng}(\bar{y})]$  then  $[F(\bar{x})] = [F(\bar{y})]$ , and (b)  $[F(\bar{x})] \notin [\text{rng}(\bar{x})]$ , for all  $\bar{x}$ .

The following is proved in [Fr1]:

**THEOREM 6.1.** *The Borel diagonalization theorem holds for any Borel equivalence relation  $E$ . Furthermore, this theorem is in class C.*

A set  $E \subseteq I^n$  is called analytic if it is of the form  $\{x : \text{for some } y, (x, y) \in S\}$  for some Borel set  $S \subseteq I^{n+1}$ . Analytic sets go well beyond Borel sets from a conceptual point of view since they are obviously not constructed by any countable limit process. Analytic sets form the next natural step up in abstraction or complexity from Borel sets in what is called the projective hierarchy, which we discussed briefly at the end of the introduction.

The relevance here of analytic sets is that Theorem 6.1 was extended in [St] as follows:

**THEOREM 6.2.** *The Borel diagonalization theorem holds for any analytic equivalence relation  $E$ . Furthermore, this theorem is in class C.*



The coanalytic sets are just the complements of the analytic sets. On the other hand, the following can be proved:

**THEOREM 6.3.** *There is a coanalytic equivalence relation on  $I$  for which the Borel diagonalization theorem fails.*

There are many interesting equivalence relations (on Borel subsets of complete separable metric spaces) that are analytic. Thus Theorem 6.2 applies to them. Theorem 6.2 was used in [St] to obtain the following: Let  $S(Q)$  be the Cantor space of subsets of  $Q$ , where  $Q$  is the rational numbers. Clearly every set  $A \subseteq Q$  can be viewed as a linear ordering inherited from the linear ordering of  $Q$ . We say that two elements of  $S(Q)$  are isomorphic if they are isomorphic as linear orderings. We say that  $F : S(Q) \rightarrow S(Q)$  is isomorphically invariant if isomorphic arguments are sent to isomorphic values.

**THEOREM 6.4.** *Every isomorphically invariant Borel function on  $S(Q)$  sends some argument to an isomorphic copy of an interval in that argument. Furthermore, this theorem is in class  $C$ .*

Theorem 6.4 can be modified in many different minor ways while remaining in class  $C$ . For instance, we can insist that the interval in the argument have endpoints in the argument, or that the interval be bounded from above and below in the argument.

We now let  $\mathbf{G}$  be alternatively the space of all binary operations, semigroups, or groups on the natural numbers  $N$ . (This just means that the field of points is  $N$ .) These are Borel subspaces of the Baire space  $N^N$ . We also let  $\mathbf{G}_f$  be the subspace of, respectively, finitely generated operations, semigroups, or groups. We say that a subset of an operation on  $N$  is finitely equationally defined if it is the set of all solutions of some finite set of equations in one variable with parameters allowed from  $N$ . In the case of groups, we allow the inverse operation to be used in these equations.

The following is proved in [St]:

**THEOREM 6.5.** *Every isomorphically invariant Borel function on  $\mathbf{G}$  sends some group (semigroup, operation) to an isomorphic copy of a subgroup (sub-semigroup, suboperation). Furthermore, this theorem is in class  $A$  for each one of the three choices for  $\mathbf{G}$ .*

Mappings  $F : \mathbf{G}_f^\infty \rightarrow \mathbf{G}_f$  are also considered in [St]. The following is proved there:

**THEOREM 6.6.** *Every isomorphically invariant Borel function  $F : \mathbf{G}_f^\infty \rightarrow \mathbf{G}_f$  sends some sequence of finitely generated groups (semigroups, operations) to a finitely generated group (semigroup, operation) which is embeddable in one of its coordinates. Furthermore, this theorem is in class  $B$  for each one of the three choices for  $\mathbf{G}$ .*

For our purposes, a graph consists of a subset of  $N$  called vertices, and a set of unordered pairs of vertices called edges. Infinite graphs are allowed, but no multiple edges. The space of graphs is naturally a Cantor space.

The detached subgraphs of a graph are taken to be the unions of connected components of the graph.

We have been able to prove the following:

**THEOREM 6.7.** *Every isomorphically invariant Borel function on graphs sends some graph to an isomorphic copy of a detached subgraph. Furthermore, this theorem is in class C.*

All of the examples given thus far in this section clearly illustrate the general theme stated at the end of the previous section. We conclude this section with an example that does not really fit into the general theme, but which is closely tied up with the so-called axiom of determinacy, which figures so prominently in the work on the projective hierarchy discussed at the end of the introduction.

By way of background, the following is well known to be false:

**PSEUDOTHEOREM.** *Every Borel set  $E \subseteq I \times I$  contains or is disjoint from the graph of a Borel function on  $I$ .*

However, the following is proved in [Fr1] (we call a set  $E \subseteq I \times I$  symmetric if  $(x, y) \in E$  if and only if  $(y, x) \in E$ ):

**THEOREM 6.8.** *Every symmetric Borel set  $E \subseteq I \times I$  contains or is disjoint from the graph of a Borel (or even left continuous) function on  $I$ . Every symmetric Borel set  $E \subseteq K \times K$  contains or is disjoint from the graph of a continuous function on  $K$  ( $K$  is the Cantor set). Furthermore, both theorems are in class C.*

**7. Strong Borel diagonalization on groups and graphs.** In this section we present some extensions of Theorem 6.6 which are not provable in ZFC. They are, however, theorems of one of the most intensively studied extensions of ZFC by set theorists, i.e., ZFC + “there exists a measurable cardinal.” We abbreviate this system by ZFM.

This additional axiom is most simply stated as follows: There exists a countably additive measure on the class of all subsets of some set where the measure of every set is either 0 or 1, and the measure of points is zero.

So clearly these extensions of Theorem 6.6 are consistent with ZFC if ZFM is consistent. But is ZFM consistent?

Unfortunately, this question has a confusing answer. It seems to be consistent in the sense that the set theorist’s use of ZFM has not led to any inconsistencies. On the other hand, the number of man hours devoted to testing ZFM is insignificant compared to that devoted to general

mathematics, and set theorists have had a vested interest in ZFM being consistent for many years now.

Of course, it would be best if one could prove that ZFM is consistent if and only if ZFC is consistent, and carry out this relative consistency proof within ZFC.

Unfortunately the second incompleteness theorem creates an obstacle to this ever happening. The reason is that ZFM itself proves that ZFC is consistent. Hence if we could carry out this desired proof within ZFC (or even within ZFM) then we would have a proof within ZFM that ZFM is consistent. The second incompleteness theorem says this is impossible unless ZFM is inconsistent! Such is the legacy of Kurt Gödel.

Should we accept the consistency of ZFM on faith? Or should we regard this question as not meaningful? Or perhaps meaningful but perhaps forever beyond our grasp to decide?

These are deep questions about which there is no consensus among logicians. There is the background question which in this context is critical. Is it important whether or not ZFM is consistent?

The importance of an extension of ZFC such as ZFM is dependent on what you can do with it that you cannot do in ZFC. An ultimate illustration of the importance of the consistency of ZFM would be afforded by a dramatic result such as the following, which is by no means ruled out at this point (but of course could be ruled out at any time): I am *not* making this as a conjecture.

**POSSIBLE BUT WILDLY SPECULATIVE.** There is a specific simple variant of the twin prime conjecture which is true if and only if ZFM is consistent. This equivalence is provable well within ZFC.

If some result anywhere near this was obtained, then clearly questions about the status of systems like ZFM would assume central importance in the history of mathematics.

It has been our view for many years that a first step towards obtaining this kind of stunning result is to first obtain such a result for a statement that at least fits into the Borel point of view. This already proved to be a difficult obstacle and there is still the expectation of much better results along these lines that fit into the Borel point of view.

Recall the definition of graph and detached subgraph that we used in the previous section. We say that a graph is embeddable in another graph if there is a one-one map from the vertices of the first into the vertices of the second such that every edge in the first is sent to an edge in the second. We say that a graph is completely embeddable if the same holds with the additional requirement that two vertices are connected by an edge in the first graph if and only if their images are connected by an edge in the second graph. Also, we say that a graph is locally finite if every vertex is joined to at most finitely many vertices.

The following propositions are discussed in [St]:

**PROPOSITION 7.1.** *Every isomorphically invariant Borel  $F : \mathbf{G}_f^\infty \rightarrow \mathbf{G}_f$  sends all of the infinite subsequences of some sequence  $G$  to a group (semi-group, operation) which is embeddable in one of the coordinates of  $G$ .*

**PROPOSITION 7.2.** *Every isomorphically invariant Borel  $F : \mathbf{G}_f^\infty \rightarrow \mathbf{G}$  sends all of the infinite subsequences of some sequence  $G$  to a group (semi-group, operation) which is embeddable in some direct limit of  $G$ .*

The following is proved in [St]:

**THEOREM 7.3.** *Propositions 7.1 and 7.2 are provable in ZFM but not in ZFC. This is true for any of the three choices (groups, semigroups, operations) for  $G$ .*

Alternatively, graphs can be used in the following way instead of groups, semigroups, and operations:

**PROPOSITION 7.4.** *Every isomorphically invariant Borel function on the locally finite graphs sends all of the detached subgraphs of some  $G$  to graphs embeddable (completely embeddable) into  $G$ .*

**THEOREM 7.5.** *Proposition 7.4 is provable in ZFM but not in ZFC. This is true for both kinds of embeddability.*

**THEOREM 7.6.** *Propositions 7.1, 7.2, and 7.4 imply the consistency of ZFC. Furthermore this fact can be proved well within ZFC.*

We now discuss the implications that the results cited in this section have for the constructible point of view.

Recall that the axiom of constructibility is known to decide the set-theoretic propositions that have been shown to be independent of ZFC by direct use of the forcing method such as Propositions 2.1–2.6.

However, here the propositions in question are not decided by the axiom of constructibility. In fact, the axiom of constructibility has a clear meaning in the context of weaker systems than ZFC, and so the same point can be made with regard to the results cited in §§5 and 6. More specifically:

**THEOREM 7.7.** *Theorems 5.1, 5.3, 6.1, 6.2, 6.4–6.8 and Propositions 7.1, 7.2, 7.4 remain unprovable in the same systems in which they were originally stated to be unprovable, even if the axiom of constructibility is added to those respective systems.*

Also recall that the constructible point of view does not assert that the axiom of constructibility is true, but only proposes that all mathematical statements be relativized to the constructible sets, i.e., that the mathematical universe be taken to be the constructible sets in the sense of a regularity condition. What happens when the assertions cited in Theorem 7.7 are so relativized?

**THEOREM 7.8.** *If any of Theorems 5.1, 5.3, 6.1, 6.2, 6.4–6.8 and Propositions 7.1, 7.2, 7.4 are relativized to the constructible sets, then their metamathematical status as cited remains unchanged, i.e., the resulting statements are provable in the same systems in which they were stated to be provable, and remain unprovable in the same systems in which they were stated to be unprovable.*

This important point can be taken further. There are various natural short initial segments and fragments of the constructible hierarchy of sets that have been studied. One purpose of examining such fragments is that, to varying extents, they constitute more explicit universes of sets which do not depend on the acceptance of any concept of abstract ordinal which is necessary in the case of the full constructible hierarchy of sets. Aside from the smallest of these fragments, the sets of integers present are closed under the hyperjump operation. Most of Theorem 7.8 depends only on the closure of the constructible sets under this operation:

**THEOREM 7.9.** *If any of Theorems 5.1, 5.3, 6.1, 6.5, 6.6, and 6.7 and Propositions 7.1, 7.2, 7.4 are relativized to any given universe of sets closed under hyperjump, then their metamathematical status as cited remains unchanged.*

The original propositions about Borel functions that exhibit these strong metamathematical properties appeared in [Fr1]. The versions discussed here are more natural.

**8. The predicative point of view.** The comprehension axiom scheme in ZFC allows one to construct a set by writing down  $\{x \in a : A(x)\}$ , where  $A(x)$  is any set-theoretic property of sets  $x$  that is expressible in the language of ZFC. Of course,  $A(x)$  may have side parameters. Here we discuss some philosophical aspects of this set existence principle in case the set  $a$  is  $N$ , the set of all natural numbers. Thus we are concerned with proofs of the existence of sets of natural numbers.

The issue is this. Suppose we assert the existence of  $\{n \in N : A(n)\}$ . Suppose also that the property  $A$  refers to all sets of natural numbers in its expression in the language of set theory. Have we really constructed a set of natural numbers? Why do we accept the existence of such a set of natural numbers?

If we take the position that this set of natural numbers is constructed by writing down  $\{n \in N : A(n)\}$  in the sense that it did not exist before anybody wrote this down (unless it coincidentally happened to have the same members as some such set that was written down earlier), then there is the real question of the meaning of, say,  $A(1)$ . Do the references in  $A$  to all sets of natural numbers refer to the set allegedly under construction? How about sets that have not been so constructed, but will be so constructed in the future? If it is not clear what sets are being referred to in  $A$  then in what sense is this a

construction? In what sense is  $A$  meaningful?

The most natural position to take on such matters, assuming one wishes to accept this set existence principle, is that all sets of natural numbers exist independently of how humans construct them, view them, or understand them. They are just there, independently of our mental processes, and we use our mental processes to observe them, study them, and use them. Through our mental processes we have observed that  $\{n \in N : A(n)\}$  exists, and was there before any human thought about it or thought about  $A$ .

A problem with this so-called Platonistic approach is that it is unclear how far it can be reasonably taken. If a purely external objective reality of all sets of natural numbers exists for us to observe and study, then why not such a reality of all sets of sets of natural numbers? But then we seem stuck with accepting the point of view that the continuum hypothesis is a matter of objective reality that simply awaits additional observation and study. As discussed earlier, the continuum hypothesis is not only independent of ZFC, but at this point the discovery of any new fundamental principles about the cumulative hierarchy of sets that would settle it seems very remote. It seems hard to merely accept that what is needed is simply some hard work or clever idea, as has proved to be the case for so many hard open mathematical problems that eventually get solved. Most mathematicians are quite uncomfortable with the concept of objective external reality when pushed as far as to include sentences such as the continuum hypothesis. They are even more uncomfortable in the context of such sentences as "there are measurable cardinals."

As discussed earlier, many specialists in set theory wish to take this Platonistic approach to the extreme; that the entire cumulative hierarchy of sets has an objective external reality awaiting our observation and study, and that any well-formed assertion about this hierarchy is objectively true or false.

But for those who do not accept this extreme view, the question of where the objective external reality ends and human intervention begins is a real issue.

It seems to us that, ultimately, there is no such good dividing line, and that a certain kind of relativism is emerging: That there is no such thing as an objective external reality anywhere outside the most extreme basic context (such as the study of the integers from 1 to 100). Instead, there are degrees or levels of external objective reality, running the spectrum from  $\{1, 2, \dots, 100\}$  to the entire cumulative hierarchy of sets (or even maybe beyond). On this view, the really interesting thing to do is to analyze the relationships between these contexts and to obtain definite mathematical results which shed light on these degrees or levels. The incompleteness results discussed in this manuscript do just that. Other types of results, such as consistency proofs, which are not discussed here, also contribute to this general aim.

Let us return to our discussion of  $\{n \in N : A(n)\}$ . The predicative point of view accepts an objective external reality of the totality of natural numbers,

but rejects any objective external reality of the totality of sets of natural numbers. On this view, sets of natural numbers do not exist independently of their construction. All constructions of sets of natural numbers must in some sense be grounded in the natural numbers themselves. On the predicative point of view, all such constructions are admissible.

The most typical case of a construction of a set of natural numbers from the predicative point of view is that of  $\{n \in N : A(n)\}$ , where  $A$  is arithmetical. In other words, when all quantifiers in  $A$  range over the natural numbers. If side parameters exist for sets of natural numbers, then the construction is relative to those side parameters. If the side parameters have been constructed, i.e., given a predicative meaning, then the expression is then a predicatively meaningful construction. This amounts to what is called the arithmetical comprehension axiom scheme.

Life would be very simple if one could merely identify predicativity with the arithmetical comprehension axiom scheme. However, consider the following situation. One may have an explicit assignment to each natural number  $n$  of an arithmetical formula  $A_n(k)$ , say, with no side parameters. Then we may wish to construct, say,  $\{n \in N : (\exists k)(A_n(k))\}$ . This cannot be done within arithmetical comprehension, but seems to be arguably within the scope of predicativity.

There has been considerable effort devoted to codifying the predicative point of view into appropriate formal systems, with some theorems suggesting, in some way, that such formal systems completely capture predicativity. We do not believe that the point of view naturally lends itself to such characterization, although there clearly are constructions such as the ones cited above which obviously fall within the predicative, as well as constructions which obviously do not fall within the predicative. It is possible that one may be able to amplify on the usual description of the predicative point of view, maintaining its fundamental philosophical flavor, so that the view would naturally lend itself to such characterization. But even this has not been accomplished in any convincing way.

We think that, under these circumstances, the really fruitful investigation is to see what consequences the predicative point of view has on actual mathematics.

Fortunately, in nearly all known interesting mathematical situations, a given proof of a theorem is either obviously predicative or obviously impredicative. Usually, a given theorem either can be given a proof which is obviously predicative, or a recursion-theoretic result is known which implies that it obviously has no predicative proof. The typical case of the former is that the arithmetical comprehension scheme is enough, and the typical case of the latter is that the theorem is shown to be false in the universe of hyperarithmetical sets of natural numbers.

Typical cases of theorems which are known to not be predicatively provable by the above method are (1) the order comparability of well-orderings

of the natural numbers, (2) the presence of perfect sets within uncountable closed sets of real numbers, and (3) the least upper bound principle for (even arithmetically defined) sets of real numbers.

However, notice that all three examples assert the existence of some set of natural numbers (perhaps disguised as a real number, a function on the natural numbers, or a perfect set as the complement of the union of a sequence of rational open intervals). It is perhaps not too surprising that there would be such basic examples, since the predicative point of view severely restricts the set existence axioms allowed.

A crucial issue about the predicative point of view is whether there are such basic mathematical theorems that do not assert the existence of infinite sets of natural numbers, even under disguise, yet can only be proved impredicatively. It is to be expected that mathematicians advocating the predicative point of view are likely to believe that there are no such significant examples.

However, in the next section we present such examples, which have only been discovered in the 1980's.

Strong advocates of the predicative point of view include such great mathematicians as Hermann Weyl and Henri Poincaré. It would have been interesting to see how their advocacy of predicativity would have been affected by the discovery of these examples.

For more discussion on predicativity, see work of S. Feferman; e.g. [Fe1] and [Fe2]. In the next section we use the formal system  $ATR_0$  as a working model for the upper limit of predicativity. This is generally accepted in light of its connection with hyperarithmetical sets and the proof theoretic ordinal  $\Gamma_0$ .

**9. Finite trees and finite graphs.** In this section we present the examples mentioned at the end of §8 of theorems not involving the existence of infinite sets of natural numbers, yet which cannot be predicatively proved. Some of the examples have the stronger property that they do not even mention infinite sets of natural numbers, even in disguise.

The first such example was the celebrated theorem of J. B. Kruskal in 1960 concerning the embeddability of finite trees in infinite sequences of finite trees. It was not until 1981 that anyone observed that it cannot be predicatively proved, despite the fact that the original proof was blatantly predicative and Kruskal had called attention to the peculiar nature of the proof.

The proof was later greatly simplified and streamlined by Nash-Williams. This new proof spawned a whole new interesting field of combinatorics called *wqo theory*. The Nash-Williams proof is sufficiently simple and the crucial impredicative step is sufficiently easy to identify, that we give a sketch of it here.

A tree consists of a nonempty set  $V$  called vertices, together with a partial ordering  $\leq$  on  $V$  such that (a) there is a (unique) least element called the



root, and (b) the set of predecessors of every vertex under  $\leq$  is linearly ordered under  $\leq$ . We have the obvious sup and inf operations on sets of vertices provided that the tree is finite (i.e., has a finite number of vertices).

The crucial notion of embedding  $h$  from one finite tree  $T_1$  into another  $T_2$  is this:  $h$  is a one-one mapping from the vertices of the first into the vertices of the second, and  $h$  is inf preserving in the sense that  $h(a \inf b) = h(a) \inf h(b)$ . These conditions imply that  $h$  is order preserving in the strong sense that  $a \leq_1 b$  if and only if  $h(a) \leq_2 h(b)$ . We write  $T_1 \leq T_2$  if and only if there exists such an embedding from  $T_1$  into  $T_2$ .

The following is proved in [Kr]:

**THEOREM 9.1.** *In any infinite sequence  $T_1, T_2, \dots$  of finite trees, there are  $i < j$  such that  $T_i \leq T_j$ . In any infinite set of finite trees, one of the elements is embeddable into another.*

The following is proved in [Si1]:

**THEOREM 9.2.** *Theorem 9.1 cannot be proved in the formal system  $ATR_0$ , and hence cannot be given a predicative proof. This holds for either of the two forms given.*

Before we sketch the Nash-Williams proof of Theorem 9.1, we give some other variants which may be a little more natural from a graph theorist's viewpoint.

We can alternatively define a tree to be a connected graph with no cycles. Note there is no root in this treatment. The relevant concept of embedding is that of a one-one mapping  $h$  from vertices in the first tree into vertices in the second tree such that if  $ab$  and  $ac$  are edges in the first tree,  $a \neq b \neq c$ , then the unique simple path from  $h(a)$  to  $h(b)$  in the second tree does not cross the unique simple path from  $h(a)$  to  $h(c)$  in the second tree (except of course at  $h(a)$ ). Or, alternatively, we may view graphs as topological spaces (1-dimensional complexes), and we merely require that the embeddings be homeomorphic mappings (continuous and one-one). The latter does not require that vertices go to vertices. If we did require that, then it would be identical to the graph-theoretic definition we have just given.

We have looked into these alternative definitions and found that the differences are inessential from the metamathematical point of view:

**THEOREM 9.3.** *Theorems 9.1 and 9.2 hold for infinite sequences of graph-theoretic finite trees, under any of the notions of embedding discussed above.*

We now sketch the proof of Theorem 9.1 given in [Na]. The method is called the minimal bad sequence argument.

A quasiordering is merely a nonempty set under a transitive and reflexive relation (i.e., if  $a \leq b$  and  $b \leq c$  then  $a \leq c$ , and also  $a \leq a$ ). A well quasi ordering is a quasi ordering with the crucial property that for all infinite sequences  $a_1, a_2, \dots$ , there are  $i < j$  such that  $a_i \leq a_j$ . It is interesting

and well known that this is equivalent to the requirement that within any infinite set  $A$  there are  $a, b \in A$  such that  $a \leq b$ .

Note that Theorem 9.1 can be restated as asserting that the finite trees under embeddability constitute a well quasi ordering.

Let  $(Q, \leq)$  be a quasi ordering. Then we form the new quasi ordering  $FIN(Q)$  consisting of the finite subsets of  $Q$  under the following quasi order:  $A \leq^* B$  if and only if there is a one-one mapping  $h: A \rightarrow B$  such that for each  $a \in A$ ,  $a \leq h(a)$ .

Let us assume for the moment the following theorem from [Hi] known as Higman's lemma:

**THEOREM 9.4.** *If  $Q$  is a well quasi ordering then so is  $FIN(Q)$ .*

We continue the sketch of the proof of Theorem 9.1.

By way of contradiction we let  $T_1, T_2, \dots$  be a counterexample to (the first form of) Theorem 9.1. Such a counterexample is called an infinite bad sequence. We want to prove that there is no infinite bad sequence.

We first need to construct what is called a minimal bad sequence. Let  $S_1$  be any finite tree of minimal possible size (as measured by the number of vertices) such that  $S_1$  starts some infinite bad sequence. Let  $S_2$  be any finite tree of minimal possible size such that  $S_1, S_2$  starts some infinite bad sequence. Continue in this way to obtain the minimal bad sequence  $S_1, S_2, \dots$ .

There are two not very explicit aspects to this construction. Firstly, the axiom of choice is used since we did not specify which of the several possible finite trees is to be chosen at each stage. But this is truly a minor point. We can enumerate all the finite trees up to isomorphism in some reasonable order before the construction begins in order to avoid this problem (using canonical representations from the equivalence classes in a standard and routine way). Secondly, there is a blatant impredicativity in the construction since at each stage we refer to unrestricted infinite sets of natural numbers (finite trees), including the one being constructed. This is just the kind of construction that is criticized by predicativists. This second point is the crux of the matter. Theorem 9.2 explains why this aspect is unavoidable.

Now that we have our minimal bad sequence, we let  $Q$  be the set of all upwardly closed subtrees of the  $S$ 's whose roots lie right above the roots of the  $S$ 's. In other words, each  $S_i$  is the joining together of several disjoint subtrees by the root of  $S_i$ ; such subtrees are called the immediate subtrees.  $Q$  consists of all such immediate subtrees. We make  $Q$  into a quasi ordering by our notion of embeddability.

It is not difficult to see that because of the minimal badness of the  $S$ 's,  $Q$  must be a well quasi ordering. For, if  $Q$  had an infinite bad sequence, then that sequence could be used to obtain a new infinite bad sequence which agrees with  $S$  for a while, and then uses subtrees of the  $S$ 's; this would violate the minimality of the  $S$ 's at the spot where the subtrees of the  $S$ 's

start, and where the copying of the  $S$ 's themselves end.

Now by Higman's lemma, the finite sequences from  $Q$  are also well quasi ordered. From this we obtain  $i < j$  such that the set of all immediate subtrees of  $S_i$  is  $\leq^*$  the set of all immediate subtrees of  $S_j$ . But this immediately implies that  $S_i \leq S_j$ , which is the contradiction we have been seeking.

Higman's lemma itself (Theorem 9.4) can also be proved by a (simpler) minimal bad sequence argument. However, in contrast to Theorem 9.1, it can be given an alternative proof within the arithmetical comprehension scheme.

Note that although Theorem 9.1 does not state the existence of infinite mathematical objects, it does mention them (universally). We now discuss a finite reformulation of Theorem 9.1 which does not involve infinite mathematical objects at all.

The idea is simple and natural. We first weaken the statement by placing bounds on the number of vertices,  $|T|$ , of the trees  $T$ . Thus we can consider the following:

(\*) For all  $k$  and finite trees  $T_1, T_2, \dots$ , with each  $|T_i| \leq k + i$ , there are  $i < j$  such that  $T_i \leq T_j$ .

Note that the collection of infinite sequences of finite trees satisfying this growth condition (for a fixed  $k$ ) is a compact space. Hence as is standard in such situations, this is true for infinite sequences of such trees if and only if it is true for sufficiently long finite sequences of such trees. Thus we are naturally led to the following:

**THEOREM 9.5.** *For  $r \gg k$  and finite trees  $T_1, \dots, T_r$  obeying  $|T_i| \leq k + i$ , there are  $i < j$  such that  $T_i \leq T_j$ .*

The following is proved in [Si1] and [Smith]:

**THEOREM 9.6.** *Theorem 9.5 cannot be proved in  $ATR_0$  and hence cannot be given a predicative proof. This is true even if graph-theoretic trees are used.*

An obvious question is: how large must  $r$  be as a function of  $k$  in Theorem 9.5? It is clear that this is a recursive function, since we can just look for a big enough  $r$  and check to see that we have it. However, the following is proved in [Si1], which obviously strengthens Theorem 9.6:

**THEOREM 9.7.** *No provably recursive function of  $ATR_0$  is sufficient to bound the required size of  $r$  as a function of  $k$  in Theorem 9.5.*

We mention an alternative finite form of Theorem 9.5 that may be viewed as being even more natural (see [Smith]):

**THEOREM 9.8.** *If  $r \gg k$  then every sequence  $T_1, \dots, T_r$  of finite trees obeying  $|T_i| \leq i$  contains an increasing subsequence of length  $k$ . This is not provable in  $ATR_0$  and hence does not have a predicative proof. Furthermore, no provably recursive function of  $ATR_0$  is sufficient to bound the required size of  $r$  as a function of  $k$ . Again, graph-theoretic trees can be used.*

In [Smith] yet another finite form is considered which involves only the growth condition  $|T_i| \leq i$  as in Theorem 9.8. Only this time the  $k$  represents the number of labels. Kruskal also considered finite trees with labels from a finite set. The embeddability condition is strengthened to demand that the embedding be label preserving. The following is implicit in [Smith]:

**THEOREM 9.9.** *If  $r \gg k$  and  $T_1, \dots, T_r$  are finite trees with  $k$  labels obeying  $|T_i| \leq i$ , then there are  $i < j$  such that  $T_i \leq T_j$  (label preserving). This theorem has the same properties cited in Theorem 9.8.*

The independence results stated in Theorems 9.2–9.3 and 9.6–9.9 are understated in that they hold for systems somewhat stronger than  $ATR_0$ . The optimal system to use is the somewhat stronger system  $\Pi_2^1 - BI_0$ . However, certainly  $ATR_0$  is a more natural system, representing, in a sense, the border of the usual formalisms for predicativity, and being equivalent to basic mathematical facts such as the comparability of well-orderings, as in reverse mathematics (see [Si2]).

It is clear that all of the theorems about trees discussed from Theorem 9.5 and beyond in this section are of the form  $\forall \exists$  over the natural numbers. Actually, technically speaking, they are presented in form  $\forall \exists \forall$ , since they assert that for all  $k$  there is a  $t$  such that for all  $r > t$ , something holds. However, it is obvious for the statements under question that if any  $r$  works then trivially any larger  $r$  works. Thus the statements are actually of the form: for all  $k$  there is an  $r$ .

If we specialize the outermost quantifier  $k$  of an  $\forall \exists$  statement, then we get an  $\exists$  statement. Such a statement is always provable in any reasonable system if and only if it is true. But the interesting question, under these circumstances, is: how large is the least possible  $r$ ? And, how large is the least possible proof that there is an  $r$ ?

From the point of view of the incompleteness phenomena, the second question is what really is interesting. If one can show that the least possible proof that there is an  $r$  is ridiculously large, then one has exhibited an incompleteness phenomena that is different from what we have discussed up to this point.

Let  $2^{[n]}$  be a stack of  $n$  two's iteratively exponentiated; e.g.,  $2^{[4]} = 2^{16}$ .

The following is proved in [Smith]:

**THEOREM 9.10.** *In Theorem 9.9, if  $k$  is set to 6 then the resulting  $\exists$  statement cannot be proved within  $ATR_0$  without using at least  $2^{[1000]}$  symbols. Hence it cannot be proved predicatively without using a humanly unreasonable number of symbols: in this sense, it is unprovable predicatively.*

Similar results can be given for all of the  $\forall \exists$  statements considered in this section.

Kruskal's theorem with finitely many labels can be strengthened so as to obtain independence results such as the above from yet stronger systems. We

did discover such natural strengthenings by adding an additional condition on the embeddings. The added condition is called the gap condition.

The idea is as follows. Suppose  $h$  is an inf preserving embedding from  $S$  into  $T$ , and assume that the trees are labeled from the finite set  $\{1, \dots, n\}$  and are label preserving. If  $b$  is an immediate successor vertex to  $a$  in  $S$ , then  $h(b)$  may not be an immediate successor vertex to  $h(a)$  in  $T$ . Of course,  $h(b)$  does lie above  $h(a)$  in  $T$ . But there might be a gap of vertices strictly in between. The additional condition asserts that all of the labels of the vertices in this gap in  $T$  must be numerically at least as large as the label of  $b$  (or  $h(b)$ ). We write  $\leq_n$  for this quasi ordering.

The following is proved in [Si1]:

**THEOREM 9.11.** *Each  $\leq_n$  is a well quasi ordering. This theorem can be proved in  $\Pi_1^1 - CA$  but not in  $\Pi_1^1 - CA_0$ , or in what is called finitely iterated inductive definitions.*

The proof that each  $\leq_n$  is a well quasi ordering involves an iteration of the minimal bad sequence construction  $n$  times. One way of looking at the proof is as follows: assume that the result is false, and then construct an appropriate minimal bad sequence. From this sequence, construct a new quasi ordering and another minimal bad sequence through that. Iterate this procedure  $n$  times until finally one obtains a bad sequence through some quasi ordering which ostensibly is a well quasi ordering, obtaining the desired contradiction. Such a proof would involve (roughly)  $n$  iterated inductive definitions. The union of  $n$  iterated inductive definitions corresponds to  $\Pi_1^1 - CA_0$ .

We conjectured that  $\leq_\omega$  is a well quasi ordering, where the domain is the finite tree labeled from  $\omega$  and the embedding is required to be inf preserving, nowhere label decreasing, and the gap condition holds (in the gap, the labels are all numerically at least that of  $h(b)$ ). In fact, we made the more general conjecture that this was true for each  $\leq_\alpha$  for any ordinal  $\alpha$ . This conjecture has been recently proved in [Kriz]. It is interesting to observe that the proof is given in  $\Pi_2^1 - CA$ , even for  $\alpha = \omega$ . It is known that for each  $\alpha$  this must take at least about  $\alpha$  iterated inductive definitions to prove that  $\leq_\alpha$  is well quasi ordered. So the lower and upper bounds are wildly far apart at this time.

The fact that the  $\leq_n$  is a well quasi ordering was subsequently used several places in the very lengthy proof that the finite graphs are well quasi ordered under the relation of minor inclusion, written  $\leq_m$  (see [RS]). We say that  $G$  is *minor included* in  $H$  if  $G$  can be obtained from  $H$  by successive applications of the following operations: (1) removing an edge, (2) removing a vertex (and all edges coming out of it), and (3) contracting an edge to a vertex.

**THEOREM 9.12.** *The relation  $\leq_m$  on the finite graphs is a well quasi ordering, i.e., for all  $G_1, G_2, \dots$ , there are  $i < j$  such that  $G_i \leq_m G_j$ .*

The following is proved in [FRS] by showing that " $\leq_m$  is a well quasi

ordering" implies "each  $\leq_n$  is a well quasi ordering."

**THEOREM 9.13.** *Theorem 9.12 cannot be proved within  $\Pi_1^1 - CA_0$ . It can be proved in  $\Pi_1^1 - CA + BI$ . In particular, there is no predicative proof of Theorem 9.12.*

We can give a number of finite forms of this graph minor theorem which also cannot be proved in such systems just as we did for Kruskal's theorem. Let  $|G|$  be the sum of the number of vertices and edges of  $G$ . We give two such forms as discussed in [FRS]:

**THEOREM 9.14.** *If  $r \gg k$  and  $G_1, \dots, G_r$  are finite graphs with each  $|G_i| \leq k + i$ , then there are  $i < j$  such that  $G_i \leq_m G_j$ . If  $r \gg k$  then every sequence  $G_1, \dots, G_r$  of finite graphs obeying  $|G_i| \leq i$  contains an increasing subsequence of length  $k$  under minor inclusion. Neither of these theorems are provable in  $\Pi_1^1 - CA_0$  and hence do not have predicative proofs. Furthermore, no provably recursive function of  $\Pi_1^1 - CA_0$  is sufficient to bound the required size of  $r$  as a function of  $k$ .*

**10. Finite Ramsey theory.** Ramsey theory has become an established branch of combinatorics of extensive scope (see [GRS]). In this section we will be discussing the original Ramsey theorems that form the basis of the subject. In [PH] a modified form of the original finite Ramsey theorem was given and shown to be unprovable within Peano arithmetic (or finite set theory). It is provable from the original infinitary Ramsey theorem, which in turn is provable by, for instance, augmenting Peano arithmetic with functions defined by arithmetical recursion. Putting it more simply, the modified finite Ramsey theorem is not provable in finite set theory or Peano arithmetic, but can be proved just beyond them.

The modified finite Ramsey theorems were the first examples of interesting mathematical theorems about finite objects which were shown to have substantial independence properties. They predate the earliest results of this kind from §9 by four years (1977 versus 1981). There was considerable expectation that the examples would blossom into further examples which would exhibit much stronger independence properties such as having no predicative proof. However, for this purpose the direct approach via Ramsey theory turns out to be an apparant dead end.

Here is the original infinitary Ramsey theorem from [Ramsey]:

**THEOREM 10.1.** *If all of the  $k$ -element subsets of a countably infinite set are colored from a finite set, then there is an infinite subset all of whose  $k$ -element subsets are assigned the same color.*

The proof of this theorem is by induction on  $k$ . The case  $k = 1$  is obvious. The case  $k + 1$  is reduced to case  $k$  as follows. Observe that if we fix any element  $x$  then we obtain an induced coloring of the  $k$ -element subsets of the set without  $x$ . Thus we fix  $x_1$ . We choose  $A_1$  to be any infinite set excluding  $x_1$  such that all  $k$ -element subsets of  $A_1$  are assigned

the same color induced by  $x_1$ . Then choose  $x_2$  to be any element of  $A_1$  and  $A_2$  to be any infinite subset of  $A_1$  excluding  $x_2$  such that all  $n$ -element subsets of  $A_2$  are assigned the same color induced by  $x_1$ . Continue in this way indefinitely. This results in an infinite sequence of  $x$ 's. It is clear that the color assigned to any subset of the  $x$ 's of size  $k+1$  depends only on the identity of the earliest  $x$  in the subset. Since there are only finitely many colors, there is an infinite set  $E$  of  $x$ 's such that all subsets of the  $x$ 's whose first  $x$  is from  $E$  must be assigned the same color. In particular, clearly every subset of  $E$  of size  $k+1$  must be assigned the same color.

The original finite form of Theorem 10.1 is as follows [Ramsey]:

**THEOREM 10.2.** *If  $r \gg k, n, m$  and all  $k$ -element subsets of an  $r$ -element set are colored from an  $n$ -element set, then there is an  $m$ -element subset all of whose  $k$ -element subsets are assigned the same color.*

The easiest proof of this theorem is to derive it from Theorem 10.1. Fix  $k, n$ , and  $m$ , and assume Theorem 10.2 is false. Then for each  $r$  there is a counterexample coloring of the  $r$ -element set  $\{1, 2, \dots, r\}$ . One can now construct counterexample colorings  $C_r$  for each  $r$ , such that each coloring is extended by the next coloring. The union of the  $C$ 's form a coloring of the  $n$ -element subsets of all of  $N$ . Applying Theorem 10.1, we obtain an infinite set all of whose  $n$ -element subsets are assigned the same color by the  $C$ 's. But note that the first  $m$  elements of this infinite set satisfies the condition in Theorem 10.2 for the coloring  $C_t$ , where  $t$  is the last of these first  $m$  elements. Hence  $C_t$  was not a counterexample coloring after all. This is the desired contradiction.

Note that this proof gives no information about how large  $r$  must be relative to  $k, n$ , and  $m$ . It is clear that the proof is highly inexplicit.

However, in this case there is an explicit proof. In fact, Ramsey's original proof in [Ramsey] was explicit and gave iterated exponential bounds for how large  $r$  must be relative to  $k, n$ , and  $m$ .

In [PH] an additional clause is added to the conclusion of Theorem 10.2. We say that a set  $A \subseteq N$  is *relatively large* if the number of elements in  $A$  is numerically at least as large as the minimum element of  $A$ .

The following is proved and studied in [PH]:

**THEOREM 10.3.** *If  $r \gg k, n, m$  and all  $k$ -element subsets of  $\{1, \dots, r\}$  are colored from an  $n$ -element set, then there is a  $\geq m$ -element subset of  $\{1, \dots, r\}$  all of whose  $k$ -element subsets are assigned the same color, and which is relatively large.*

Note that this is just as easy a corollary of Theorem 10.1 as is Theorem 10.2, since, trivially, every infinite subset of  $N$  contains arbitrarily large finite relatively large subsets.

It is shown in [PH] that Theorem 10.3 is not provable in Peano arithmetic (PA), but can be proved just beyond it in, e.g., arithmetic comprehension

(ACA). Furthermore, no provably recursive function of PA is sufficient to bound the required size of  $r$  as a function of  $k, n, m$ .

The concept of relatively large uses integers simultaneously in the role of "elements" and of "number of elements." This dual role is sufficiently unusual in mathematics as to prompt a search for alternatives to Theorem 10.3 that do not use concepts such as relatively large. A particularly attractive alternative is through what we call a *function value theorem*. We first give the infinitary form.

**THEOREM 10.4.** *Let  $F$  be a function from all  $\leq k$ -element subsets of  $N$  into  $N$  and  $m \in N$ . Then there is an infinite set  $A \subseteq \{m, m+1, \dots\}$  such that  $F$  takes on at most  $k+1$  values  $\leq \min(A)$ .*

And here is the straightforward finite form.

**THEOREM 10.5.** *Let  $r \gg k, n, m$  and  $F$  be a function from all  $\leq k$ -element subsets of  $\{1, \dots, r\}$  into  $\{1, \dots, r\}$ . Then there is a  $\geq n$ -element  $A \subseteq \{m, \dots, r\}$  such that  $F$  takes on at most  $k+1$  values  $\leq \min(A)$ .*

Theorem 10.5 has the same metamathematical properties as Theorem 10.3 (in fact, the two can be shown to be equivalent within a weak fragment of PA).

We close the discussion by presenting some *function congruence theorems*.

**THEOREM 10.6.** *For any  $F: N^k \rightarrow N^k$  there are  $x_1 < x_2 < \dots < x_{k+1}$  with  $F(x_1, x_2, \dots, x_k) \equiv F(x_2, x_3, \dots, x_{k+1}) \pmod{2}$ . Furthermore, a bound can be placed on  $x_{k+1}$  which depends on  $k$  but not on  $F$ .*

**THEOREM 10.7.** *For any  $F: N^k \rightarrow N^k$  there are  $1 < x_1 < x_2 < \dots < x_{k+1}$  with  $F(x_1, x_2, \dots, x_k) \equiv F(x_2, x_3, \dots, x_{k+1}) \pmod{x_1}$ . Furthermore, a bound can be placed on  $x_{k+1}$  which depends on  $k$  but not on  $F$ .*

Theorem 10.6 has a bound involving approximately  $k$  iterated exponentials, and no fewer.

Theorem 10.7 cannot be bounded with a provably recursive function of PA.

**Appendix.** Progress towards the construction or discovery of basic mathematical problems about finite objects, with a clear and intuitive meaning, conveying interesting mathematical information, that is readily graspable, and which is independent of ZFC, has been incremental. Here we indicate the current state of the art.

The most convincing independence results in this vein are currently stated in terms of countably infinite functions or sets. Nevertheless, one proves, well within ZFC, that the independent sentences are equivalent to sentences involving only the ring of integers. Unfortunately, we do not know how to directly put the sentences into such finite terms without causing unacceptable complications.



Let  $\otimes : A \times A \rightarrow Q$ , where  $Q$  is the rationals. We say that  $R \subseteq Q \times Q$  is  $\otimes$ -Boolean if  $R$  can be defined by a quantifier free formula in  $(Q, <, \otimes)$ ; i.e., if  $R$  can be defined in terms of conjunction, disjunction, negation, and inequalities between expressions built up from  $\otimes$ , variables, and constants from  $Q$ .

**PROPOSITION 1.** *There is a  $\otimes : Q \times Q \rightarrow Q$  such that for all  $\otimes$ -Boolean  $R \subseteq Q \times Q$  and  $a \in Q$ , there is a  $b \otimes b < a \otimes a$  such that for all  $x < b \otimes b$  and  $y < a \otimes a$ , if  $R(x, y)$  then  $R(x, (b \otimes b) \otimes x)$ .*

The following is proved in [Fr3].

**THEOREM 2.** *Proposition 1 can be proved in ZFC with the use of Mahlo cardinals of every finite order, but cannot be proved in ZFC. In fact, Proposition 1 is provably equivalent to the consistency of  $ZFC + \{ \text{there is a Mahlo cardinal of order } \bar{n} \}_n$ , (within  $RC A_0$ ).*

Let  $F : N^k \rightarrow N$  and  $A$  be any set. We write  $F_{<}[A]$  for  $\{x : F(y_1, \dots, y_k) = x \text{ for some } y_1, \dots, y_k < x \text{ chosen from } A\}$ .

By way of background, note that the following is easily provable and compactly expresses the fundamental principle of definition by induction on the natural numbers.

**THEOREM 3.** *For all  $F : N^k \rightarrow N$  there is an  $A \subseteq N$  with  $N = A \Delta F_{<}[A]$ .  $A$  is necessarily infinite.*

However, the following proposition is provably false (within  $RC A_0$ ).

**PROPOSITION 4.** *For all  $n \gg k$  and  $F : N^k \rightarrow N$ , there is an infinite set  $1, n \in A \subseteq N$  with  $N = A \Delta F_{<}[A]$ .*

Now consider the following weakening of Proposition 4.

**PROPOSITION 5.** *For all  $n \gg k$  and  $F : N^k \rightarrow N$ , there are infinite sets  $1, n \in A_1 \subseteq A_2 \subseteq \dots \subseteq A_k \subseteq N$  with  $A_i + A_i \subseteq A_{i+1} \Delta F_{<}[A_{i+1}]$ ,  $i < k$ .*

It can be shown that Proposition 5 can be proved in ZFC with Mahlo cardinals of every finite order, but not in ZFC. In fact, Proposition 5 can be shown to be equivalent to the 1-consistency of  $ZFC + \{ \text{there is a Mahlo cardinal of order } n \}_n$ , (within  $ACA$ ). Proposition 5 can be proved for each fixed  $k$  (with  $ACA$ ). The rate of growth associated with  $n \gg k$  is bounded by a recursive function but not any provably recursive function of ZFC (even with Mahlo cardinals of any given finite order).

It can also be shown (within  $ACA$ ) that Propositions 1 and 5 are true if and only if they are true in the arithmetic sets.

## REFERENCES

- [Ajtai] M. Ajtai, *On the boundedness of definable linear operators*, Period. Math. Hungar. **5** (1974), 343–352.
- [Co] Paul J. Cohen, *The independence of the continuum hypothesis*, Proc. Nat. Acad. Sci. U.S.A. **50** (1963), 1143–1148; **51** (1964), 105–110.
- [Dales] H. G. Dales, *A discontinuous homomorphism from  $C(X)$* , Amer. J. Math. **101** (1986), 647–734.
- [DMR] M. Davis, Y. Matijasevic, and J. Robinson, *Hilbert's tenth problem. Diophantine equations: Positive aspects of a negative solution*, Proc. Sympos. on the Hilbert Problems (DeKalb, IL, May 1974), vol. 28, Amer. Math. Soc., Providence, RI, 1976, pp. 323–378.
- [DW] H. G. Dales and W. H. Woodin, *An introduction to independence for analysts*, London Math. Soc. Lecture Notes, no. 115, Cambridge Univ. Press, London and New York, 1986.
- [Fe1] S. Feferman, *Systems of predicative analysis*, J. Symbolic Logic **29** (1964), 1–30.
- [Fe2] —, *A more perspicuous formal system for predicativity*, Konstruktionen versus Positionen I, Walter de Gruyter, Berlin, 1979, pp. 68–93.
- [Fr1] H. Friedman, *On the necessary use of abstract set theory*, Adv. in Math. **41** (1981), 209–280.
- [Fr2] —, *Unary Borel functions and second order arithmetic*, Adv. in Math. **50** (1983), 155–159.
- [Fr3] —, *Necessary uses of abstract set theory in finite mathematics*, Adv. in Math. **60** (1986), 92–122.
- [FRS] H. Friedman, N. Robertson, and P. Seymour, *The metamathematics of the graph minor theorem*, Logic and Combinatorics (S. G. Simpson, ed.) Contemp. Math., vol. 65, Amer. Math. Soc., Providence, RI, 1987, pp. 229–261.
- [Fu] L. Fuchs, *Infinite Abelian groups*, vol. II, Academic Press, 1973, pp. 178–181.
- [Go1] Kurt Gödel, *The consistency of the axiom of choice and of the generalized continuum hypothesis*, Ann. Math. Stud., no. 3, Princeton Univ. Press, Princeton, NJ, 1940.
- [Go2] —, *What is Cantor's continuum problem?*, Philosophy of Mathematics (P. Benaceraf and H. Putnam, eds.), Cambridge Univ. Press, London and New York, 1983.
- [Go3] —, *Collected works*, vols. I, II, Oxford Univ. Press, New York, 1986, 1990.
- [GRS] R. Graham, B. Rothschild, and J. Spencer, *Ramsey theory*, Wiley, New York, 1980.
- [HML] *Handbook of mathematical logic*, Studies in Logic and the Foundations of Mathematics (Jon Barwise, ed.), vol. 90, North-Holland, Amsterdam 1977.
- [HMS] L. Harrington, D. Marker, and S. Shelah, *Borel orderings*, Trans. Amer. Math. Soc. **310** (1988), 293–302.
- [HMSS] *Harvey Friedman's research on the foundations of mathematics*, Studies in Logic and the Foundations of Mathematics (L. A. Harrington, M. D. Morley, A. Scedrov, and S. G. Simpson, eds.), vol. 117, Elsevier, Amsterdam, 1985.
- [Hi] G. Higman, *Ordering by divisibility in abstract algebras*, Proc. London Math. Soc. **2** (1952), 326–336.
- [Je1] T. Jech, *Nonprovability of Souslin's hypothesis*, Comment. Math. Univ. Carolin. **8** (1967), 291–305.
- [Je2] —, *Set theory*, Academic Press, 1978.
- [Je3] —, *Multiple forcing*, Cambridge Tracts in Math., vol. 88, Cambridge Univ. Press, London and New York, 1986.
- [Jensen] R. Jensen, *Souslin's hypothesis is incompatible with  $V = L$* , Notices Amer. Math. Soc. **15** (1968), 935.
- [Kriz] I. Kriz, *Well-quasi ordering finite trees with gap-condition. Solution of Harvey Friedman's conjecture*, Ann. of Math. (2) **130** (1989), 215–226.
- [Kr] J. B. Kruskal, *Well-quasi-ordering, the tree theorem, and Vázsonyi's conjecture*, Trans. Amer. Math. Soc. **95** (1960), 210–225.
- [Levy] A. Levy, *Basic set theory*, Springer-Verlag, 1979.
- [Luzin] N. Luzin, *Leçons sur les ensembles analytiques*, Gauthier-Villars, Paris, 1930.

- [Mar] D. A. Martin, *Descriptive set theory: projective sets*, in [HML], pp. 783–815.
- [Mat] Yu. Matijacevic, *Enumerable sets are Diophantine*, Soviet Math. Dokl. **11** 354–357.
- [MS] D. A. Martin and J. R. Steel, *A proof of projective determinacy*, J. Amer. Math. Soc. **2** (1989), 71–125.
- [Na] C. St. J. A. Nash-Williams, *On well-quasi-ordering finite trees*, Proc. Cambridge Philos. Soc. **59** (1963), 833–835.
- [PH] J. Paris and L. Harrington, *A mathematical incompleteness in Peano arithmetic*, in [HML], pp. 1133–1142.
- [Ramsey] F. P. Ramsey, *On a problem of formal logic*, Proc. London Math. Soc. **30** (1930), 264–286.
- [RS] N. Robertson and P. Seymour, ongoing series of papers in J. Combin. Theory Ser. B, 1983–.
- [Sh] S. Shelah, *A compactness theorem for singular cardinals, free algebras, Whitehead problem and transversals*, Israel J. Math. **21** (1975), 319–349.
- [Ship] J. Shipman, *Cardinal conditions for strong Fubini theorems*, Trans. Amer. Math. Soc. (to appear).
- [Si1] S. G. Simpson, *Nonprovability of certain combinatorial properties of finite trees*, in [HMSS], pp. 87–117.
- [Si2] —, *Reverse mathematics*, Proc. Sympos. Pure Math., vol. 42, Amer. Math. Soc., Providence, RI, 1985, pp. 461–471.
- [Smith] R. L. Smith, *The consistency strengths of some finite forms of the Higman and Kruskal theorems*, in [HMSS], pp. 119–135.
- [Smo] C. Smorynski, *The incompleteness theorems*, in [HML], pp. 821–865.
- [Sol] R. Solovay, *A model of set theory in which every set of reals is Lebesgue measurable*, Ann. of Math. (2) **94** (1970), 1–56.
- [Sp] E. Specker, *Additive gruppen von Folgen ganzen Zahlen*, Portugal. Math. **9** (1950), 131–140.
- [ST] R. Solovay and S. Tennenbaum, *Iterated Cohen extensions and Souslin's problem*, Ann. of Math. (2) **94** (1971), 201–245.
- [St] L. J. Stanley, *Borel diagonalization and abstract set theory: recent results of Harvey Friedman*, in [HMSS], pp. 11–86.
- [Tarski] A. Tarski, *A decision method for elementary algebra and geometry*, 2nd rev. ed., Berkeley and Los Angeles, 1951.

DEPARTMENT OF MATHEMATICS, OHIO STATE UNIVERSITY, COLUMBUS, OHIO 43210

## Elliptic Curves and Modular Forms

BENEDICT H. GROSS

An elliptic curve  $E$  over the field  $k$  has a nonsingular plane model of the form

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6,$$

where the coefficients  $a_i$  lie in  $k$ . The set  $E(k)$  of solutions to this equation (in the projective plane) has the structure of an abelian group: the unique point on the line at infinity is taken as the origin and any three collinear points sum to zero. When  $k = \mathbf{C}$  is the field of complex numbers, the theory of elliptic functions identifies  $E(\mathbf{C})$  with a complex torus, so—as a topological group—with the product of two circles. When the field  $k$  is finite,  $E(k)$  is clearly a finite group. When  $k$  is a number field (an extension of finite degree of the field  $\mathbf{Q}$  of rational numbers) the famous theorem of Mordell and Weil states that the group  $E(k)$  is finitely generated.

We will focus our attention on the case when  $k = \mathbf{Q}$ . Since  $E(\mathbf{Q})$  is finitely generated, we have an isomorphism

$$E(\mathbf{Q}) \simeq \mathbf{Z}^r \oplus T,$$

where  $r \geq 0$  is an integer and  $T$  is a finite group. The torsion subgroup  $T$  is easily determined in any given case, and the proof of the Mordell-Weil theorem yields an effective upper bound for the rank  $r$  of  $E$ . To determine if this upper bound is sharp requires a search for rational points.

The following example has been investigated by Bremner and Cassels. Let  $q$  be a prime number with  $q \equiv 5 \pmod{8}$ , and let  $E$  be defined by the equation

$$y^2 = x^3 + qx.$$

Then the subgroup  $T$  of  $E(\mathbf{Q})$  is cyclic of order 2, generated by the point  $P = (0, 0)$ , and the rank  $r$  satisfies  $r \leq 1$ . One suspects that  $r = 1$  in all cases, although this is only known for  $q < 20,000$ . Occasionally, the search for a solution is quite time consuming: for example, when  $q = 2437$  the

smallest point  $P = (x, y)$  of infinite order in  $E(\mathbf{Q})$  has coordinates

$$x = \frac{1058218655773369472688280687468828399922014718555143690966617841}{275081987041241794421770856177032513092966187596374600583396900},$$

$$y = \frac{443090331670870476765298567239328435425666485280521498925653374541937139166973694383354835903889}{4562398640636267034178360393354742958207189280664086915767660421227708280931775118586314953000}.$$

One approach to the determination of the rank is to study the number of solutions to the equation modulo  $p$ . Choose a plane model for  $E$  with integral coefficients and minimal discriminant  $\Delta$ . Let  $A_p$  denote the number of solutions of the reduced equation (including the point at infinity) over  $\mathbf{Z}/p\mathbf{Z}$ , and write  $A_p = p + 1 - a_p$ . The  $L$ -function of  $E$ , which packages this information into an analytic function of the complex variable  $s$ , is defined by the Euler product

$$L(E, s) = \prod_{p|\Delta} (1 - a_p p^{-s})^{-1} \cdot \prod_{p \nmid \Delta} (1 - a_p p^{-s} + p^{1-2s})^{-1}$$

which converges in the half-plane  $\Re(s) > 3/2$ . Expanded out, this product is a Dirichlet series  $\sum_{n \geq 1} a_n \cdot n^{-s}$  with integral coefficients  $a_n$ .

If we formally set  $s = 1$  in the Euler product, we find the formal product  $\prod (p/A_p^0)$ , where  $A_p^0$  is the number of nonsingular points on  $E$  modulo  $p$ . Motivated by the expectation that a large value of  $r$  should lead, on the average, to a large number of solutions modulo  $p$ , Birch and Swinnerton-Dyer conjectured that the order of vanishing of  $L(E, s)$  at the point  $s = 1$  is equal to the rank  $r$ . Aided by Cassels and Tate, they also gave an arithmetic interpretation for the leading term in its Taylor expansion there.

To begin to attack this conjecture, one needs the analytic continuation of  $L(E, s)$  to a neighborhood of  $s = 1$ . Following Taniyama, Shimura, and Weil, one now hopes to prove that the function  $L(E, s)$  is entire by showing that it is the Mellin transform of a modular form. More precisely, let  $N$  be the conductor of the curve  $E$ . This is an integer, with the same prime factors as the minimal discriminant  $\Delta$ , which measures the ramification in the division fields of  $E$ .

**CONJECTURE.** *The function  $f(\tau) = \sum_{n \geq 1} a_n e^{2\pi i n \tau}$ , for  $\tau$  in the upper half-plane, is a cusp form of weight 2 for the congruence subgroup  $\Gamma_0(N)$  of  $\mathrm{SL}_2(\mathbf{Z})$ .*

The group  $\Gamma_0(N)$  consists of integer matrices  $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$  with  $ad - bc = 1$  and  $c \equiv 0 \pmod{N}$ , and  $f(\tau)$  is modular of weight 2 if for every such matrix we have the identity  $f((a\tau + b)/(c\tau + d)) = (c\tau + d)^2 f(\tau)$ . If  $f(\tau)$  is a cusp form, its Mellin transform

$$\Lambda(f, s) = \int_0^\infty f(iy) y^s \frac{dy}{y} = (2\pi)^{-s} \Gamma(s) L(E, s)$$

is entire. Moreover, Carayol has shown that  $f$  is then a "newform" of level  $N$ , and hence an eigenfunction for the Fricke involution:  $f(-1/N\tau) = \lambda \cdot N\tau^2 f(\tau)$ , with  $\lambda = \pm 1$ . This implies that  $\Lambda(f, s)$  satisfies the functional

equation

$$\Lambda(f, s) = \varepsilon \cdot N^{1-s} \Lambda(f, 2-s)$$

with  $\varepsilon = -\lambda$ .

There is now a great deal of theoretical and computational evidence in favor of the conjecture that  $f(\tau)$  is modular, and for a given curve  $E$  it can be checked using a finite amount of computation. For example, the conjecture is true for the curves  $y^2 = x^3 + qx$  mentioned above; its truth for all elliptic curves over  $\mathbf{Q}$  implies Fermat's Last Theorem, by recent work of Ribet. In all that follows, we will assume the conjecture is true for the curve  $E$ , and will derive some geometric and arithmetic consequences.

Let  $X_0(N)$  be the modular curve over  $\mathbf{Q}$  which classifies elliptic curves with a cyclic subgroup of order  $N$ . The work of Eichler and Shimura shows that the newform  $f(\tau)$  determines an elliptic quotient  $E_0$  of the Jacobian of  $X_0(N)$  over  $\mathbf{Q}$ , and Faltings' results on the isogeny conjecture show that  $E_0$  is isogeneous to  $E$ . Hence there is a nonconstant regular map  $\varphi: X_0(N) \rightarrow E$  over  $\mathbf{Q}$  which takes the cusp  $\infty$  of  $X_0(N)$  to the origin of  $E$ . The differential  $2\pi i f(\tau) d\tau$  on the upper half-plane is invariant under  $\Gamma_0(N)$  and defines a regular differential on  $X_0(N)$  over  $\mathbf{Q}$ . Once  $\varphi$  has been chosen, there is a unique invariant differential  $\omega$  on  $E$  which satisfies  $\varphi^*(\omega) = 2\pi i f(\tau) d\tau$  on  $X_0(N)$ .

The following method of constructing points on  $E$  over number fields is due to Birch. Let  $K$  be an imaginary quadratic field of discriminant  $-D$ , where all prime factors of  $N$  are split. Let  $H$  be the Hilbert class-field of  $K$  (the maximal abelian unramified extension, which has finite degree equal to the class-number of  $K$ ). Using the theory of complex multiplication, one can construct Heegner points  $x$  on  $X_0(N)$  over  $H$ . We then define  $P_K$  as the trace of the point  $\varphi(x)$  from  $E(H)$  to  $E(K)$ ; this trace is calculated by adding  $\varphi(x)$  to its conjugates, using the group law on  $E$ . Zagier and I found a formula for its canonical height  $\hat{h}$ , which measures the amount of paper required to record  $P_K$ , in terms of the derivative of the  $L$ -function of  $E$  over  $K$ :

$$L'(E/K, 1) = \frac{\iint_{E(C)} \omega \wedge \bar{\omega}}{\sqrt{D}} \cdot \hat{h}(P_K).$$

This formula implies that the point  $P_K$  has infinite order in  $E(K)$  if and only if  $L'(E/K, 1) \neq 0$ .

The precise conjecture of Birch and Swinnerton-Dyer predicts that when  $P_K$  has infinite order, the group  $E(K)$  has rank 1, and that the finite index  $[E(K) : \mathbf{Z}P_K]$  annihilates the Tate-Šafarevič group of  $E$  over  $K$ . Kolyvagin has recently made a great advance, which essentially proves this. His work brings us close to a proof of the full conjecture of Birch and Swinnerton-Dyer, for modular elliptic curves  $E$  over  $\mathbf{Q}$  where the order of  $L(E, s)$  at  $s = 1$  is either 0 or 1. But the conjecture for those curves where the  $L$ -function vanishes to order  $\geq 2$  remains completely mysterious, as does the central

problem of why the function  $f(\tau)$  attached to an elliptic curve  $E$  over  $\mathbf{Q}$  is a modular form.

### BIBLIOGRAPHIC REMARKS

An excellent introduction to elliptic curves is the survey article by Tate [1]. This work also contains a long list of references. Some excellent introductory works on elliptic curves have also appeared recently in the Springer graduate text series, see [2–4]. For a discussion of modular forms on  $\Gamma_0(N)$ , and the theory of complex multiplication, I would recommend [5]. The formula for the heights of Heegner points is proved in [6]; Kolyvagin's work appears in [7].

1. J. T. Tate, *The arithmetic of elliptic curves*, Invent. Math. **23** (1974), 179–206.
2. D. Husemöller, *Elliptic curves*, Graduate Texts in Math., vol. 111, Springer, 1987.
3. N. Koblitz, *Introduction to elliptic curves and modular forms*, Graduate Texts in Math., vol. 97, Springer, 1984.
4. J. H. Silverman, *The arithmetic of elliptic curves*, Graduate Texts in Math., vol. 106, Springer, 1986.
5. G. Shimura, *Introduction to the arithmetic theory of automorphic functions*, Publ. Math. Soc. Japan II, Iwanomi Sholen and Princeton Univ. Press, 1971.
6. B. H. Gross and D. Zagier, *Heegner points and derivatives of  $L$ -series*, Invent. Math. **84** (1986), 225–320.
7. V. A. Kolyvagin, *Euler systems*, Grothendieck Festschrift, Birkhäuser, 1990.

DEPARTMENT OF MATHEMATICS, HARVARD UNIVERSITY, CAMBRIDGE, MASSACHUSETTS 02138

## Developments in Algebraic Geometry

JOE HARRIS

I should say at the outset that I have no claim to any particular insight into the future of algebraic geometry. What I thought I would do, accordingly, is to talk a little bit about the history of the subject, leading up to its present incarnations, and leave it to you to extrapolate.

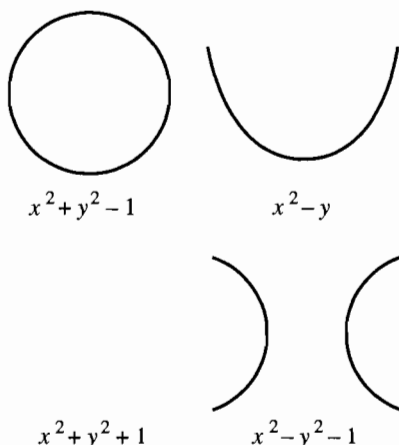
Algebraic geometry is a subject whose development has been marked by fundamental changes in the basic objects studied, and in the approach to their study. For example, one possible definition of the subject—admittedly an extreme one—would be to say that algebraic geometry is “the study of the geometry of those loci defined by polynomial equations.” If we adopt this point of view, we could say that the subject is over two millennia old: the conic sections and quadric surfaces studied by the ancient Greeks happen to be such objects.

A more balanced definition of the subject might be to say that it is the study of the relations between the algebra of polynomials and the geometry of the loci that they define. In this sense, the subject is much younger; it traces its origins back to the introduction by Descartes of the notion of coordinates in the plane, making it possible to describe a conic as the zero locus of a quadratic polynomial  $f(x, y)$ , and relate the algebraic manipulation of that polynomial to geometric operations on the curve itself.

Of course, to Descartes and to mathematicians for some time afterward, “polynomial” meant polynomials  $f_\alpha(x_1, \dots, x_n)$  with real coefficients, and “locus” meant the set of real solutions, that is, the subset  $X$  of  $\mathbb{R}^n$  of vectors  $x = (x_1, \dots, x_n)$  such that  $f_\alpha(x) = 0$ . The basic set-up of algebraic geometry from the time of Descartes until the early nineteenth century was this: one had a collection of polynomials  $f_\alpha(x_1, \dots, x_n) \in \mathbb{R}[x_1, \dots, x_n]$  with real coefficients, and one studied their common zeros in  $n$ -space  $\mathbb{R}^n$ . This was a time when the techniques of the subject were pretty rudimentary, but the problems studied were completely intelligible, even to nonexperts.

For example, consider the simplest type of algebraic variety: a plane curve,

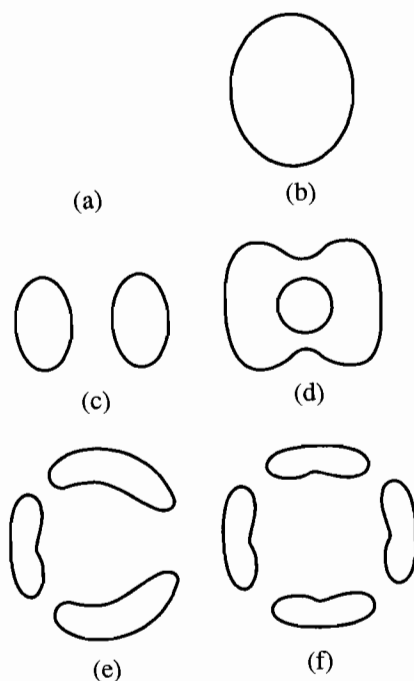




or in other words the zero locus  $X$  of a single polynomial  $f(x, y)$  of degree  $d$  in two variables. If we assume the curve  $C$  is smooth, in the sense that  $f$  does not vanish simultaneously with its two partial derivatives, then  $X$  will be a real 1-manifold, that is, a disjoint union of copies of  $\mathbb{R}$  and  $S^1$ , the latter of which were called “ovals” in the classical language. We may then ask how many arcs and ovals a plane curve may have; and what sort of configuration they may form—that is, which pairs of ovals may be nested. For example, a plane quartic—that is, the zero locus of a fourth-degree polynomial in  $x$  and  $y$ —without arcs may have any number of ovals from none to four; if there are two, they may be nested or not, as in diagrams (c) and (d). The main tool here is simply the fact that no line may meet a quartic curve more than four times, and more generally that another plane curve of degree  $e$  may meet it in at most  $4e$  points; thus, if a quartic contains two nested ovals it can contain no other points, since a line joining such a point to a point interior to the inner of the two nested ovals would meet the curve at least five times.

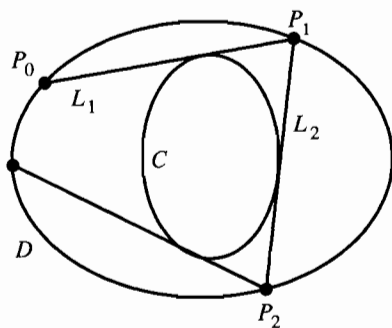
Of course, we may make further distinctions, e.g., between convex and nonconvex ovals; for example, the outer oval of two nested ovals forming a quartic may be either convex or nonconvex, as in figure (d) (the inner one must always be convex; otherwise there would exist a line meeting the curve six times).

The answer to the first of the questions posed above is *Harnack's theorem*, which says that a plane curve  $X$  of degree  $d$  may have any number of ovals from none to  $(d-1)(d-2)/2 + 1$ . It is proved in elementary fashion using the fact that a curve  $C$  passing through a point  $P$  lying on an oval of  $X$  must meet that oval at least twice. For example, in the case above suppose that a quartic curve had five ovals. We could then choose a point  $p_i$  on each of five ovals of  $C$ , and then find a conic curve  $Q$  passing through each of these points;  $Q$  would then have to meet  $C$  in at least ten points, violating the fact that a conic and a quartic can meet in at most eight points. In fact, the bound given by Harnack's theorem is sharp, as may readily be seen by example.

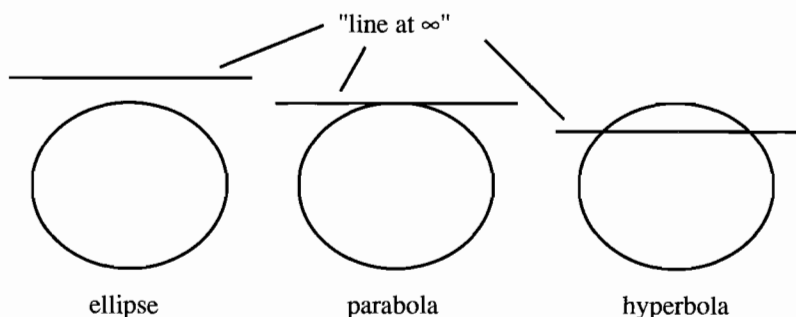


The second question above—what configurations the ovals of a plane curve may form—is, by contrast, still unanswered to this day, even in the case of curves of degree 6, though progress has been made by the Russian school.

To give another example of a problem examined and solved during this period, consider Poncelet's theorem. The original question asks when, given two ellipses  $C$  and  $D$  in the plane, there is a polygon inscribed in one and circumscribed about the other; the answer is a surprising one. To construct such a polygon, starting with a given vertex  $P_0$  on the outer ellipse  $D$ , is easy: we just take the first side  $L_1$  to be one of the two tangent lines to  $C$  through  $P_0$ ; take  $P_1$  to be the other point of intersection of this line with  $D$ ,  $L_2$  the other tangent line to  $C$  through  $P_1$ , and so on. The question is then when this process repeats after a finite number of steps; Poncelet's theorem is that it does or does not independently of the choice of initial point  $P_0$ , so that the pair  $(C, D)$  will either admit a continuous family of inscribed-and-circumscribed polygons or none at all.



**The “classical” period.** The next transformation of the subject of algebraic geometry occurred around the beginning of the nineteenth century. It consisted of two changes in the basic objects considered: the introduction of projective varieties, and of complex coordinates. The effect of the first was that seemingly different varieties in ordinary Euclidean space, or affine space as it is called, might in fact behave the same when completed in projective space: for example, the three types of smooth conics in  $\mathbb{R}^2$  all look like single ovals in  $\mathbb{RP}^2$ ; the difference lies simply in the situation of the “line at infinity” with respect to the projective conic.



Thus, we could say, to understand ordinary plane conics, we should first understand projective conics, whose behavior is more uniform; then consider the different ways in which they may meet a line in  $\mathbb{RP}^2$ . Somewhat more generally, in relation to the question posed above about arcs and ovals, we may see that a smooth curve of degree  $d$  in  $\mathbb{RP}^2$  will consist entirely of ovals; the arcs of the curve in  $\mathbb{R}^2$  will arise when the line at infinity intersects some of the ovals of the curve.

Similarly, looking at the locus of complex zeros of a polynomial, rather than just the real, has the effect of making uniform their behavior: for example, the polynomials  $1 - x^2 - y^2$ ,  $1 + x^2 - y^2$ , and  $1 + x^2 + y^2$  all have isomorphic zero loci in  $\mathbb{C}^2$ —after all, they differ only by a complex linear change of variables—even though their zeros in  $\mathbb{R}^2$  look different like a circle, a hyperbola, and the empty set, respectively. Again, the implicit idea is to understand conics over  $\mathbb{C}$  first, and then to ask what conics over  $\mathbb{R}$  may give rise to the same conic over  $\mathbb{C}$ .

The effect of this change is striking when we consider again the question about the topology of a plane curve  $X$ . If we denote by  $X(\mathbb{C}) \subset \mathbb{CP}^2$  the closure in  $\mathbb{CP}^2$  of the locus of complex solutions of  $f(x, y) = 0$ —equivalently, the locus of the corresponding homogeneous polynomial—we see that all smooth curves of a given degree  $d$  are homeomorphic: they are compact orientable surfaces of genus  $(d-1)(d-2)/2$ , and indeed are isotopically embedded in  $\mathbb{CP}^2$ .

We can use this information to say something about the real zeros of a real polynomial. The locus  $X(\mathbb{R}) \subset \mathbb{RP}^2$  of real points of  $X$  in  $\mathbb{RP}^2$  is just the set

of fixed points of the action of complex conjugation acting on  $X(\mathbb{C})$ . Thus, if  $X(\mathbb{R})$  has  $\delta$  ovals, the quotient  $X(\mathbb{C})/\tau$  will be a 2-manifold with boundary consisting of  $\delta$  copies of  $S^1$ ; if we add  $\delta$  discs  $D^2$  we may complete this to a compact 2-manifold  $Y$ . We may then compute the topological Euler characteristic of  $Y$  as

$$\chi(Y) = \chi(X(\mathbb{C})/\tau) + \delta = \chi(X(\mathbb{C}))/2 + \delta = -d(d-3)/2 + \delta.$$

But of course  $\chi(Y) \leq 2$ , and we deduce that

$$\delta \leq \frac{d(d-3)}{2} + 2.$$

Poncelet's theorem similarly appears in a new light when viewed from this vantage point. In fact, it admits a very simple proof, first observed to me by Phillip Griffiths. We look at the incidence correspondence, consisting of pairs:

$$\Gamma = \{(P, L) : P \in D, L \text{ is tangent to } C, \text{ and } P \in L\}.$$

This is again an algebraic curve, and when we look at its complex points we find that it is a torus, that is, it is isomorphic to the complex plane  $\mathbb{C}$  modulo a lattice  $\Lambda$ . In these terms, we can readily describe the action of passing from one pair  $(P_i, L_{i+1})$  to the next  $(P_{i+1}, L_{i+2})$ : it is just a translation in the plane. If this translation has finite order modulo the lattice, every polygon closes up; if not, none do; and so we get Poncelet's theorem.

In this way, the main focus of the subject shifted, in the first half of the 19th century, from varieties in real Euclidean space—real affine varieties—to complex projective ones. It is worth remarking as well that one of the main motivations for this shift was another sort of uniformity of behavior. It was felt by Poncelet, who was instrumental in bringing about both of these changes, that as a general rule intersections of varieties ought to be preserved. Thus, if two lines in general meet in a point, they should continue to do so even if they become parallel; thus the passage to projective space. By the same token, if a line and a conic meet in two points, they should continue to do so, even if we pull them apart; thus the introduction of complex numbers.



Arguments like the ones above about the number of ovals of a plane curve or Poncelet's theorem represent, from one point of view, the completion of a bargain struck when first passing from the fairly natural environment of real plane curves to complex projective ones: you make life easier for yourself by dealing with better-behaved (if less readily visualizable) objects, with the implicit promise of eventually going back and applying what you learn in this way to the original problem. (This bargain has not always been so faithfully kept; new objects tend to suggest new problems, and old ones are easily forgotten. It is embarrassing, for example, when a mathematician working with a hyperbolic PDE in three variables asks a question about real plane curves, how little we know to this day about them.) Without question, these changes opened the door to a new era in algebraic geometry, that culminated in the work of Noether, Segre, Castelnuovo, Enriques, Severi, and others of the Italian school.

**"Abstract" algebraic geometry.** The basic change from real affine variety to the complex projective one revolutionized the way people thought about algebraic geometry, and there was no going back. One of the reasons these changes stuck was that, while the mental image geometers had of algebraic varieties was altered radically, the formal structure of the subject was much less dramatically altered. Thus, while the words "algebraic curve" conjured up the image of what we would now call a compact Riemann surface, rather than what most people would identify as a curve, many of the old theorems and techniques could still be reproduced word for word in the new context.

Let me explain this in a little more detail, since it is an essential point. Given a collection of polynomials  $f_\alpha \in \mathbb{C}[x_1, \dots, x_n]$ —or equivalently the ideal  $I \subset \mathbb{C}[x_1, \dots, x_n]$  they generate—we associate to them their common zero locus  $X = V(I)$ . In the other direction, if  $X \subset \mathbb{C}^n$  is an algebraic variety we let  $I(X) \subset \mathbb{C}[x_1, \dots, x_n]$  be the ideal of polynomials vanishing on  $X$ . We thus have a two-way correspondence

$$\{\text{subvarieties of } \mathbb{C}^n\} \xrightleftharpoons[V]{I} \{\text{ideals } I \subset \mathbb{C}[x_1, \dots, x_n]\}.$$

Note that this is not by any means bijective: in one direction, the composition of the two is the identity—the definition of a variety  $X \subset \mathbb{A}^n$  amounts to the statement that  $V(I(X)) = X$ —but going the other way the composition is neither injective or surjective. We can fix this up by simply restricting our attention to the image of the map  $V$ , and happily there is a nice characterization of this image (and indeed of the composition  $I \circ V$ ); for any ideal  $I \subset \mathbb{C}[x_1, \dots, x_n]$ , the ideal of functions vanishing on the common zero locus of  $I$  is the radical of  $I$ , i.e.,

$$I(V(I)) = \text{rad}(I).$$

Thus, there is a bijective correspondence between subvarieties  $X \subset \mathbb{C}^n$  and radical ideals  $I \subset \mathbb{C}[x_1, \dots, x_n]$ .

(Note also that if we replace  $\mathbb{C}$  by  $\mathbb{R}$ , the correspondence breaks down a little further: even a radical ideal in  $\mathbb{R}[x_1, \dots, x_n]$  may be nontrivial and still have no common zero locus. Instead, we use this correspondence in effect to *define* the notion of variety over  $\mathbb{R}$ .)

Now, let  $X \subset \mathbb{C}^n$  be a variety and  $I(X)$  its ideal. The quotient ring  $A = A(X) = \mathbb{C}[x_1, \dots, x_n]/I$  is then called the *ring of regular functions* on  $X$ , or the *coordinate ring* of  $X$ . (Note that the condition  $I = \text{rad}(I)$  is equivalent to saying that the ring  $\mathbb{C}[x_1, \dots, x_n]/I$  has no nilpotent elements.) Since  $X$  is the common zero locus of the polynomials  $f \in I$ ,  $X$  is determined by the ring  $A$ ; and indeed virtually every property of  $X$  may be expressed directly in terms of  $A$  rather than of the locus  $X$ . For example, a point of  $X$  is a maximal ideal in  $A(X)$  (in the case of a real variety it is an ideal with residue field  $\mathbb{R}$ ); a map between two such varieties  $X$  and  $Y$  is exactly a ring homomorphism  $A(Y) \rightarrow A(X)$  over  $\mathbb{C}$ ; the dimension of  $X$  is the transcendence degree of the quotient ring of  $A(X)$  over  $\mathbb{C}$ , and so on. The point is, *pretty much the entire subject can be expressed in terms of the algebra of the rings  $A(X)$* . Given that, it is no longer so surprising that the passage from  $\mathbb{R}$  to  $\mathbb{C}$  involves so little actual reworking of the theory: we would expect homomorphisms between  $\mathbb{R}$ -algebras  $A$  and  $B$  to be closely related to homomorphisms between  $A \otimes \mathbb{C}$  and  $B \otimes \mathbb{C}$ , even though the corresponding varieties may be completely different in appearance.

I do not mean, of course, that this passage from the geometric to the algebraic description of algebraic geometry was simply a matter of obvious algebraic analogues of geometric constructions and properties. In fact, it involved a large number of new ideas and techniques. To give you one example, in dealing with compact Riemann surfaces, an object of fundamental importance is its Jacobian variety  $J(X)$ . This is defined classically as the quotient of complex  $g$ -space  $\mathbb{C}^g$  by a lattice  $\Lambda \subset \mathbb{C}^g$  obtained by integrating a basis of holomorphic 1-forms on  $X$  over a collection of cycles forming a basis of the first homology  $H_1(X, \mathbb{Z})$ . The problem of giving an algebraic construction of this essential object is a serious challenge, and was not solved until Andre Weil. In general, the algebraization of the subject was initiated in earnest in the work of Zariski, starting in the 1920s, and was carried out over a number of decades, reaching in some sense its culmination in the work of Serre.

Of course, having reworked the subject of algebraic geometry in this new context, it may be applied over far more fields than just  $\mathbb{R}$  and  $\mathbb{C}$ . Indeed, this is true to an extent that may seem remarkable at first. After all, a variety over a finite field  $k = \mathbb{F}_p$  will consist simply of a finite collection of points; you will not see much difference in the picture of a curve in  $k^3$  and the picture of a surface in  $k^3$ . You could argue that this is at least in part because the field  $k$  is not algebraically closed, but the fact is that a curve in 3-space over the algebraic closure  $\bar{k}$  of  $\mathbb{F}_p$  still does not look that much

different from a surface; both are just countably infinite collections of points.

Nonetheless, geometric statements about real and complex varieties will, for the most part, still be true in this general setting. For example, there is even a Lefschetz fixed point theorem: we can define a cohomology theory (étale cohomology, developed by M. Artin and Grothendieck) for varieties  $X$  over  $\mathbb{F}_p$  that mimics the ordinary topological cohomology of a variety over  $\mathbb{C}$  (albeit with coefficients in the  $l$ -adic numbers  $\mathbb{Q}_l$ ); and then it will be the case that the number of fixed points of an automorphism  $\tau$  of  $X$  will be expressed in terms of the traces of the action of  $\tau$  on the cohomology groups of  $X$ .

Indeed, this is fundamentally related to one of the main constructions of number theory, that of the zeta-function. For  $X$  a variety over the field  $\mathbb{F}_p$  of  $p$  elements, we let  $N_r$  be the number of points of  $X$  over the field  $\mathbb{F}_{p^r}$  with  $q = p^r$  elements. We may then encode this information in the *zeta-function* of  $X$ , defined to be the power series in  $t$ :

$$Z(X, t) = \exp \left( \sum N_r \cdot \frac{t^r}{r} \right).$$

If we take the special case where  $\tau$  is the Frobenius endomorphism, sending each coordinate to its  $p$ th power, then the number  $N_r$  is just the number of fixed points of the  $r$ th power of  $\tau$ . If  $\tau$  has eigenvalues  $\{\lambda_{i,j}\}$  on  $H^i(X)$ , then,

$$N_r = \sum (-1)^i \text{Tr}(\tau^r | H^i(X)) = \sum (-1)^i (\lambda_{i,j})^r,$$

so

$$\sum_r N_r \cdot \frac{t^r}{r} = \sum_{i,j,r} (-1)^i \frac{(\lambda_{i,j} \cdot t)^r}{r} = \sum_{i,j} (-1)^{i+1} \log(1 - \lambda_{i,j} \cdot t)$$

and

$$Z(X, t) = \prod_{i,j} (1 - \lambda_{i,j} \cdot t)^{(-1)^{i+1}} = \frac{P_1(t) \cdot P_3(t) \cdots}{P_0(t) \cdot P_2(t) \cdots},$$

where  $P_i(t) = \det(1 - \tau_i \cdot t)$  is the characteristic polynomial of the action  $\tau_i$  of  $\tau$  on  $H^i(X)$ . We may see in this way that the zeta-function  $Z$  is a rational function, a theorem first proved by Dwork; Deligne carried this further to prove the analogue of the Riemann hypothesis for varieties over finite fields, that the  $P_i$  were polynomials with integer coefficients and roots of absolute value  $p^{-i/2}$ .

Actually, I may be overstating the extent to which one should feel surprised that the Lefschetz fixed point theorem holds in this context. After all, the one key ingredient of the Lefschetz theorem is the notion of intersection of cycles on a manifold: the essential step in the proof is the calculation of the intersection of the diagonal in a product  $X \times X$  with the graph of a map  $f: X \rightarrow X$ , though it is often couched in the language of cup products. At the same time, intersection of cycles is a theory that existed in algebraic geometry some time before it existed in topology; indeed, it was the presence

of this notion in algebraic geometry that supposedly motivated Lefschetz to make the definition in the topological setting.

Finally, in all this talk of the algebraization of the subject, I am ignoring another fundamental shift in the subject: the change from consideration of affine or projective varieties—zero loci of polynomials—to abstract algebraic varieties. This was actually a change common to many branches of mathematics in the early twentieth century: for example, while a group in the nineteenth century meant a subset of either the symmetric or general linear group closed under composition and inverse, the twentieth century introduced the notion of abstract group. Group theory was thus split up into the analysis of abstract groups—what we now think of as group theory—and the study of ways in which a given abstract group could be mapped to the general linear group, or in other words representation theory. Similarly, the notion of abstract algebraic variety—an object locally isomorphic, in a suitable sense, to affine varieties—became the basic object of algebraic geometry.

This did not make for a change so much in the objects studied as in the way they were studied; analogously to the development of group theory, the study of varieties was “factored” into the study of abstract varieties, and then the ways in which a given abstract variety could be embedded in projective space.

**Schemes.** We come now to the latest of the revolutions in the subject of algebraic geometry, the introduction of the theory of schemes by Grothendieck in the 1950s and 1960s. The notion of scheme has had tremendous impact, both in a purely geometric and in an arithmetic setting. To a certain extent, it is possible to describe this impact separately in the two settings, and I will try to do this here.

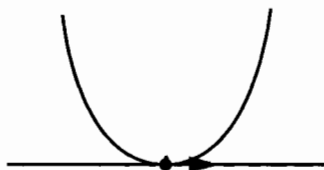
To describe a scheme in the geometric context, recall the basic correspondence introduced earlier between varieties  $X \subset \mathbb{C}^n$  and ideals  $I \subset \mathbb{C}[x_1, \dots, x_n]$ . If one is going to fix up the above correspondence so as to make it bijective, there are naively *two* ways of going about it: we can either restrict the class of objects on the right, or enlarge the class of objects on the left. In classical algebraic geometry, as we have just said, we do the former; in scheme theory, we do the latter. Thus, we more or less *define* an affine scheme  $X \subset \mathbb{C}^n$  to be an object associated to an arbitrary ideal  $I \subset \mathbb{C}[x_1, \dots, x_n]$ . To put it differently, given a finitely generated ring over  $\mathbb{C}$ —that is, a ring of the form  $A = \mathbb{C}[x_1, \dots, x_n]/I$ —we create an object, called  $\text{Spec } A$ , whose ring of functions is the ring  $A$ .

What sense can this possibly make? Just as before, this makes sense to the extent that most of the notions that we actually deal with in algebraic geometry may be defined in terms of rings and ideals. For example, if  $X \subset \mathbb{C}^n$  is the subscheme with ideal  $I = I(X)$ , we *define* a function on  $X$  to be an element of the ring  $A(X) = \mathbb{C}[x_1, \dots, x_n]/I$ ; the intersection of two such varieties  $X, Y \subset \mathbb{C}^n$  is given by the join of their ideals; the data of a map between two such varieties  $X$  and  $Y$  is equivalent to the data of a map



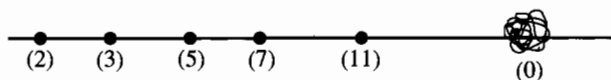
$\varphi: A(Y) \rightarrow A(X)$ ; a point of  $X$  is a prime ideal  $p$  in  $A(X)$ ; the fiber of the map  $X \rightarrow Y$  given by  $\varphi: A(Y) \rightarrow A(X)$  over a point  $p \in X$  is the subscheme of  $Y$  corresponding to the ring  $A(Y)/\varphi^{-1}(p)$ , and so on. The point is, all these things make as much sense whether or not  $I$  is a radical ideal.

In these circumstances, for example, if we wanted to intersect the line  $(y = 0)$  with the conic curve  $(y = x^2)$ , we would take the intersection to be the object  $X = \text{Spec } \mathbb{C}[x, y]/(y, x^2) \subset \mathbb{C}^2$  in the affine plane defined by the ideal  $(y, y - x^2) = (y, x^2)$ . This object has only one point, but it is not the same as the point defined by the ideal  $(x, y)$ : we simply declare a function on  $X$  to be an element of the quotient ring  $\mathbb{C}[x, y]/(y, x^2) = \mathbb{C} \oplus \mathbb{C} \cdot x$ —that is, an expression of the form  $a + bx$ . In other words, we say that a function  $f(x, y)$  on the plane vanishes on  $X$  if and only if it vanishes at the point  $(0, 0)$  and has normal derivative  $\partial f / \partial x$  zero at  $(0, 0)$  as well.



It is interesting to note that one justification for this generalization of the notion of variety comes from the same source as Poncelet's. Again, consider a line and a conic in the plane, and suppose now that the line becomes tangent to the conic. As before, we would like to say that there are still two points of intersection of the two. Classically, it was just said that the line and the conic intersected at the one point "with multiplicity 2," but this is unsatisfactory from a number of points of view. Scheme theory gives us a way of refining it: we say that the intersection of the line  $(x)$  with the conic  $(x - y^2)$  is the scheme given by the ideal  $(x, x - y^2) = (x, y^2)$ . This not only conveys the multiplicity of intersection in the fact that the ideal is not radical, it tells us also the direction from which the two points that coalesced into this one point came.

The second, arithmetic, impact of the notion of scheme arises from a further generalization. To put it simply, we may observe that in the construction of the scheme  $\text{Spec } A$  there does not need to be a ground field at all: the ring  $A$  in general need not contain any field. Thus, for example, we have a fundamentally important scheme  $\text{Spec } \mathbb{Z}$ , whose points (except for  $(0)$ ) correspond to the prime numbers.



In effect, then, we are treating the integers as variables—as functions on our

space  $\text{Spec } \mathbb{Z}$ . This turns out to be one of the most crucial points in the application of schemes to number theory. For example, a diophantine problem—in other words, a variety defined by polynomials with integer coefficients such as  $y^2 = x^3 + 1$ —will give rise to a scheme  $X = \text{Spec}(\mathbb{Z}[x, y]/(y^2 - x^3 - 1))$ . The inclusion  $\mathbb{Z} \subset \mathbb{Z}[x, y]/(y^2 - x^3 - 1)$  then gives a map  $X \rightarrow \text{Spec}(\mathbb{Z})$ , whose fibers (as loosely defined above) are exactly the reductions of the original equation modulo the primes. Nor does the ring have to be finitely generated; geometric objects associated to rings such as power series rings are extremely useful as auxiliary objects in algebraic geometry.

Needless to say, for every “why not” I toss off blithely here, a tremendous amount of foundational work is implicit. For example, consider again the Jacobian of a curve: now that we have described a curve as an object fibered over  $\text{Spec } \mathbb{Z}$ , its Jacobian should be one as well. Actually constructing such an object—showing it exists and has the functorial properties we want—is a project of major proportions (it took Steve Kleiman essentially a semester to describe his solution of this problem in a course I attended). The need for this sort of foundational material has given the subject, unavoidably, a reputation for technical difficulty and inaccessibility. On the other hand, it would be hard to overestimate the power of the ideas implicit in these notions. After all, it is worth bearing in mind, to most mathematicians of the early 19th century the notion of a complex projective variety must have seemed more than a little forbidding as well.

Let me finish by considering what may lie ahead. If you are comfortable with the thesis presented here, that progress in algebraic geometry is reflected as much in its definitions as in its theorems, the natural question to ask is what objects algebraic geometers will be studying in the next century. Currently there are two notions abroad that make a claim to be the natural successor of the notion of scheme (and Manin has even suggested that they should be amalgamated).

The first is the notion of *compactified arithmetic scheme*, developed by Arakelov, Faltings, and others. In this, we take a scheme of finite type over  $\mathbb{Z}$  and add additional structure to it: we throw in the data of a Kähler metric on the “fiber at infinity.” This additional structure in some sense addresses the problem that there is no “compactification” of  $\text{Spec } \mathbb{Z}$  to a projective scheme and allows us to tie together many of the phenomena associated to individual primes. For example, the classical fact that the total degree of a rational function on a projective curve—that is, the same number of zeros minus the number of poles—is zero translates into the product formula, that the product of the valuations of an element of a number field at all primes (including the infinite ones) is one. The notion of arithmetic scheme is, as you might expect, of special interest to number theorists, and indeed played a role in Faltings’ proof of the Mordell conjecture.

The second new notion is that of a *superscheme*. This is a generalization of the notion of scheme, in which we loosen still further the strictures on

the rings we consider: we no longer require that it be commutative. This is not to say that we look at arbitrary noncommutative rings; rather, we look at rings with a  $\mathbb{Z}/2$ -grading and require that they be skew-commutative, in the sense that for  $x$  and  $y$  homogeneous,  $x \cdot y = (-1)^{\deg(X)\deg(Y)} y \cdot x$ . Thus, a commutative ring represents the special case where the grading is trivial; and in general (if we are not in characteristic 2) the odd graded piece will consist of nilpotents, though nilpotents that behave very differently than those considered in the context of "classical" schemes. Much of the motivation for the study of superschemes comes from physics, though the questions that arise in trying to carry standard algebraic geometry over into this new context seem interesting in their own right.

Does either of these two notions embody the future of algebraic geometry; does it lie in some other direction altogether, or will the future bring about a return to the classical questions of the subject? The only general pattern to the development of the subject thus far seems to be a gradual but consistent trade-off of naive geometric intuition for a formal unity (in each case, met with cries of, "It may be a pretty theory, but it's not geometry!"). Whether this continues, how far and in what direction, is anybody's guess.

DEPARTMENT OF MATHEMATICS, BROWN UNIVERSITY, PROVIDENCE, RHODE ISLAND 02912

*Current address:* Department of Mathematics, Harvard University, Cambridge, Massachusetts 02138

## A Century of Lie Theory

ROGER HOWE

The origins of a subject are frequently difficult to trace. The extent to which precursor fields and early investigations, later perceived to have anticipated the emergence of a field or to fit naturally into it, should be annexed to the field can be a matter for vigorous debate. In the case of Lie theory, Sophus Lie was already studying “continuous, finite groups of transformations” in the 1870s, and one could even make a case for including Euclidean geometry as part of Lie theory [Crtn7]. However, in 1888 the first volume of *Theorie der Transformationsgruppen* by “S. Lie unter Mitwirkung von Dr. F. Engel” [LiEn1] was published by Teubner in Leipzig; and also Killing’s classification [Kill] of complex semisimple Lie algebras appeared in *Mathematische Annalen*. These events are the basis for the title of this article.

My assignment, roughly, is to report on the development of Lie theory over the past 100 years, and to extrapolate it into the future. To do this in a uniform, systematic way is, for me and I suspect for anyone, impossible. So this account will be frankly idiosyncratic; I make here a blanket apology to the many investigators whose interesting results will be slighted or ignored completely; or maybe worse, treated clumsily. All I can offer by way of consolation is the remark that it has happened to me too. Similarly, although the bibliography is extensive, it is not at all comprehensive. References are only intended to provide the reader with representative sources of further information. Again I offer apologies to the many authors who will find that I have neglected to mention relevant work of theirs.

1. The first example of a Lie group is Euclidean space  $\mathbf{R}^n$  with vector addition as the group operation, but it is too simple-minded a group to reveal the essential features of Lie theory. Almost as well known, and much more interesting structurally, is  $\mathrm{GL}_n(\mathbf{R})$ , the group of real invertible  $n \times n$  matrices, with matrix multiplication as the group operation. It serves as the basic template for Lie theory in the following sense: any subgroup of  $\mathrm{GL}_n(\mathbf{R})$  which is closed (with respect to the standard topology on  $n \times n$  matrices) is

a Lie group; and conversely, modulo some relatively subtle caveats ignorable at this point (see §2.5), any Lie group is realizable as a closed subgroup of  $GL_n(\mathbf{R})$  for some  $n$  (cf. [Hoch]). Of course,  $GL_n(\mathbf{R})$  contains discrete, even finite, subgroups, but Lie theory, in its most basic form, ignores these. The subgroups of  $GL_n(\mathbf{R})$  which are the immediate subjects of Lie theory are the ones which are the opposite of discrete: the connected ones. The first miracle of Lie theory is that the extremely weak topological hypothesis—closed and connected, when combined with the algebraic condition—subgroup, yields a subset which is a smooth (even analytic) surface (i.e., submanifold). If one then looks at the tangent space to this surface at the identity matrix, one finds it is endowed with a certain algebraic structure, the Lie bracket (which as an operation on matrices is simply commutator). This is the Lie algebra. The second miracle of Lie theory is that, except for the caveats ignored above, this Lie algebra, a vector space with a bilinear nonassociative product, completely determines the group from which it comes.

Without discussing in detail yet the foundational results of Lie theory, we can observe that its essential feature seems to be the enrichment of the algebraic notion of group by the topological notion of continuity: a Lie group is an object which carries in a compatible way the structure of group and of differentiable manifold (in fact analytic manifold; by Hilbert's 5th Problem, finally solved in the early 1950s [Glea, MoZi1, Yama1, Yama2, MoZi2, Kap11], it is enough to require a Lie group to be locally homeomorphic to Euclidean space—no smoothness need be explicitly assumed). Thus continuity, indeed smoothness, seems to be a *sine qua non* of the theory. Therefore, it is interesting to observe that Lie theory has intimate and fruitful interactions with the theory of discrete, in particular finite, groups.

1.1. An important aspect of the connection can be illustrated by a careful study of the bread-and-butter topic of elementary linear algebra, Gaussian elimination. Let  $A$  be an  $n \times n$  matrix with entries  $\{a_{ij} : 1 \leq i, j \leq n\}$ , and let  $z = (z_1, z_2, \dots, z_n)^T$  be a column vector of length  $n$ . Consider the system of linear equations

$$(1.1.1) \quad Ax = z,$$

from which we want to solve another column vector of length  $n$  for  $x$ . We will assume  $A$  is invertible, so that the solution  $x$  exists and is unique. Gaussian elimination is a standard method for solving system (1.1.1). We will discuss it in its naive form, untouched by worries about round-off error.

Writing the system (1.1.1) out long-hand we obtain:

$$(1.1.2) \quad \begin{array}{ccccccc} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & = & z_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n & = & z_2 \\ & \vdots & & \vdots & & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n & = & z_n \end{array}$$

Suppose  $a_{11} \neq 0$ . Then for  $k = 2, 3, \dots, n$ , we can subtract  $a_{k1}/a_{11}$  times the first equation from the  $k$ th equation to arrive at an equivalent system:

$$(1.1.3) \quad \begin{aligned} a'_{11}x_1 + a'_{12}x_2 + a'_{13}x_3 + \dots + a'_{1n}x_n &= z'_1 \\ a'_{22}x_2 + a'_{23}x_3 + \dots + a'_{2n}x_n &= z'_2 \\ &\vdots \\ a'_{n2}x_2 + a'_{n3}x_3 + \dots + a'_{nn}x_n &= z'_n \end{aligned}$$

where  $a'_{1j} = a_{1j}$ ,  $z'_1 = z_1$ , and  $a'_{kj} = a_{kj} - a_{k1}a_{1j}a_{11}^{-1}$ ,  $z'_k = z_k - a_{k1}z_1a_{11}^{-1}$ ,  $k \geq 2$ .

Since the last  $n - 1$  equations of system (1.1.3) involve only the  $n - 1$  unknowns  $x_2, x_3, \dots, x_n$ , we evidently have a recursive procedure for solving the system (1.1.1)–(1.1.2). Providing that  $a'_{22} \neq 0$ , we may subtract  $a'_{k2}/a'_{22}$  times the second equation in system (1.1.3) from the third through  $n$ th equations to obtain a third equivalent system with a subsystem of  $n - 2$  equations in  $n - 2$  unknowns. And soon, after  $n - 1$  steps, we will arrive at a triangular system:

$$(1.1.4) \quad \begin{aligned} b_{11}x_1 + b_{12}x_2 + \dots + b_{1n}x_n &= y_1 \\ b_{22}x_2 + \dots + b_{2n}x_n &= y_2 \\ b_{33}x_3 + \dots + b_{3n}x_n &= y_3 \\ &\vdots \\ b_{nn}x_n &= y_n \end{aligned}$$

This system can of course be solved by “back-substitution.” Of the several slight variants of this procedure, we select the following. First, divide each equation by its leading coefficient to obtain:

$$(1.1.5) \quad \begin{aligned} x_1 + b'_{12}x_2 + b'_{13}x_3 + \dots + b'_{1n}x_n &= y'_1 \\ x_2 + b'_{23}x_3 + \dots + b'_{2n}x_n &= y'_2 \\ &\vdots \\ x_n &= y'_n \end{aligned}$$

where  $b'_{ij} = b_{ii}^{-1}b_{ij}$ ,  $y'_i = b_{ii}^{-1}y_i$ . Now observe we have already solved for  $x_n$ :

$$(1.1.6a) \quad x_n = y'_n.$$

We can therefore compute  $x_{n-1}$  by the simple recipe

$$(1.1.6b) \quad x_{n-1} = y'_{n-1} - b'_{n-1n}x_n,$$

and so on. If we know  $x_n, x_{n-1}, \dots, x_{i+1}$ , then we compute  $x_i$  by the formula

$$(1.1.6c) \quad x_i = y'_i - \sum_{j=i+1}^n b'_{ij}x_j.$$

Thus, under some mild assumptions about the nonvanishing of certain numbers, we have a systematic procedure for performing matrix inversion by means of ordinary (i.e., scalar) arithmetic.

Let us formulate this procedure in terms of matrix manipulations. Let  $L_1$  be the matrix

$$(1.1.7) \quad L_1 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -a_{11}^{-1}a_{12} & 1 & 0 & \cdots & 0 \\ -a_{11}^{-1}a_{13} & 0 & 1 & \cdots & 0 \\ & & & \ddots & \vdots \\ -a_{11}^{-1}a_{1n} & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Then in terms of matrices, the passage from the system (1.1.2) to the system (1.1.3) amounts to the multiplication of both sides of equation (1.1.1) by  $L_1$ : equation (1.1.1) is the matrix version of (1.1.2) and the matrix version of (1.1.3) is

$$(1.1.8) \quad L_1 Ax = L_1 y.$$

Similarly, the second stage of the procedure is equivalent to multiplying the system (1.1.8) by the matrix

$$(1.1.9) \quad L_2 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & (-a'_{22})^{-1}a'_{21} & 1 & 0 & \cdots & 0 \\ 0 & (-a'_{22})^{-1}a'_{31} & 0 & 1 & & 0 \\ \vdots & \vdots & & & \ddots & \vdots \\ 0 & 0 & & & & 0 \\ 0 & (-a'_{22})^{-1}a'_{n1} & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Of the matrices  $L_i$ , we may observe

- (i) Each  $L_i$  has all entries zero above the diagonal; that is,  $L_i$  is lower triangular.
- (ii) Additionally each  $L_i$  has all its diagonal entries equal to 1. Since for a lower triangular matrix the diagonal entries equal the eigenvalues, this is the same as saying all the eigenvalues of  $L_i$  are equal to 1. A matrix with all eigenvalues equal to 1 is called *unipotent*.

Thus the matrices  $L_i$  are unipotent lower triangular matrices. Let us denote the set of all unipotent lower triangular matrices by  $\mathcal{U}$ . We can easily

check that

- (1.1.11) (i) The product of two matrices in  $\overline{\mathcal{U}}$  is also in  $\overline{\mathcal{U}}$ .  
 (ii) The inverse of a matrix in  $\overline{\mathcal{U}}$  is in  $\overline{\mathcal{U}}$  (the analogous fact for unipotent upper triangular matrices is implicit in equations (1.1.6a)–(1.1.6c)). Thus  $\overline{\mathcal{U}}$  is a group. In particular, the successive multiplications of our system (1.1.1) by the  $L_i$  may be achieved by multiplication by a single appropriate element of  $\overline{\mathcal{U}}$ .

The result of these modifications by lower triangular unipotent matrices is the system (1.1.4) which, in contrast to the  $L_i$ , is the system corresponding to an *upper* triangular matrix. Thus passage from (1.1.2) to (1.1.4) is expressed in matrix terms by the equation

$$(1.1.12) \quad \tilde{L}A = B,$$

where  $\tilde{L}$  is a unipotent lower triangular matrix and  $B$  is an upper triangular matrix. Further, the passage from (1.1.4) to (1.1.5) corresponds to a matrix factorization

$$(1.1.13) \quad B = DU,$$

where  $D$  is a diagonal matrix and  $U$  is a unipotent upper triangular matrix. Combining (1.1.12) and (1.1.13) gives us

$$(1.1.14) \quad \tilde{L}A = DU$$

which we can convert to a factorization

$$(1.1.15) \quad A = LDU,$$

where  $L (= \tilde{L}^{-1})$  is lower triangular unipotent,  $D$  is diagonal, and  $U$  is upper triangular unipotent.

Thus, Gaussian elimination produces, and is equivalent to, the factorization (1.1.15) of a “generic” matrix  $A$  into a product of upper and lower triangular unipotent matrices and a diagonal matrix. This equivalence is well known and can be found in elementary textbooks [Hill, Stra].

1.2. However, as we have noted, this procedure will not always work.<sup>1</sup> We can ask what to do when it does not. One can observe that it is always possible to permute the equations so that, after rearrangement, the desired diagonal coefficient is nonzero, and the elimination can proceed. This provides an algorithm that will always work, so elementary texts usually stop their discussion with this, or a similar remark.

However, it is interesting to see what one can do in as systematic a fashion as possible. Let us look again at the system (1.1.1) or (1.1.2), admitting the possibility that  $a_{11} = 0$ . Then we may search down the first column until we find a nonzero coefficient. (There must be one if  $A$  is nonsingular.) Suppose the first row with a nonzero first entry is row  $i_1$ . Then we may add



a multiple of the  $i_1$ th row to the rows below to make zeros of all entries of the first column, except for  $a_{i_1,1}$ . To describe this precisely, let  $E_{ij}$  denote the "matrix unit" which has 1 in the  $i$ th row,  $j$ th column, and zeros everywhere else. Then the process of eliminating all but one entry in the first column amounts to multiplying  $A$  on the left by the matrix

$$(1.2.1) \quad L_1 = I - \sum_{j \geq i_1} a_{i_1,1}^{-1} a_{j,1} E_{ji_1}.$$

Here  $I$  is the identity matrix. Because  $i_1$  was the index of the first row to have a nonzero entry, this matrix will be unipotent lower triangular. Also, note that if  $a_{11} \neq 0$ , then  $i_1 = 1$ , and the matrices  $L_1$  of (1.1.7) and (1.2.1) coincide.

So now we have a matrix

$$(1.2.2) \quad A' = L_1 A$$

which has only one nonzero entry in the first column, in the  $i_1$ th row. Look at the second column. Choose the index  $i_2$  so that

$$(1.2.3) \quad \begin{aligned} & \text{(i) } i_2 \neq i_1, \\ & \text{(ii) } a'_{i_2,2} \neq 0, \\ & \text{(iii) } i_2 \text{ is as small as possible subject to (i) and (ii).} \end{aligned}$$

With  $i_2$  so chosen, we can eliminate all entries in the second column below row  $i_2$  by multiplying by the matrix

$$(1.2.4) \quad L_2 = I - \sum_{j \geq i_2} a'_{i_2,2}{}^{-1} a'_{j,2} E_{ji_2}.$$

The matrix  $L_2$  is unipotent lower triangular. The resulting product

$$(1.2.5) \quad A'' = L_2 A'$$

has the properties

$$(1.2.6) \quad \begin{aligned} & \text{(i) Only row } i_1 \text{ has a nonzero entry in the first column;} \\ & \quad \text{only rows } i_1 \text{ and } i_2 \text{ have nonzero entries in the first} \\ & \quad \text{and second columns.} \\ & \text{(ii) In the second column, all rows below row } i_2 \text{ have a} \\ & \quad \text{zero.} \end{aligned}$$

Note that, with regard to condition (1.2.6)(ii), we should distinguish two cases: if  $i_1 > i_2$ , then (1.2.6)(ii) says that  $a''_{i_2,2} = 0$ , so the second column of  $A''$  will have only one nonzero entry, viz.  $a_{i_2,2}$ ; but if  $i_1 < i_2$ , then row  $i_1$  of  $A'$  passes unchanged to  $A''$ , and it may happen that  $a''_{i_1,2} \neq 0$ .

It should now be evident that we can continue this process of multiplying  $A$  by unipotent lower triangular matrices until we produce

$$(1.2.7) \quad \tilde{B} = \tilde{L}A$$

with the properties:

- (1.2.8) (i) For each  $j$ ,  $1 \leq j \leq n$ , the matrix  $\tilde{B}$  has exactly  $j$  rows with nonzero entries in the first  $j$  columns.  
 (ii) If  $i_j$  is the row which has its first nonzero entry in the  $j$ th column, then  $b_{kj} = 0$  if  $k > i_j$ .

These properties are the analogs for general  $i$  of properties (1.2.6) for the case  $i = 2$ .

Given a matrix  $B$  satisfying conditions (1.2.8) we can produce from it an upper triangular matrix  $\tilde{B}$  simply by permuting its rows: we move row  $i_1$  to row 1, row  $i_2$  to row 2, and so forth. This also amounts to a matrix multiplication:

$$(1.2.9) \quad B = \tilde{P}\tilde{B},$$

where  $\tilde{P}$  is a permutation matrix—a matrix with all entries zero except for one 1 in each row and column, whose effect on a column  $n$ -vector is simply to permute its entries. (Precisely,  $\tilde{P}$  will take the  $i_j$ th entry to the  $j$ th entry.)

Finally, we can factor the upper triangular matrix  $\tilde{B}$  as in (1.1.13). Combining (1.1.13), (1.2.7), and (1.2.9) gives the following result.

**THEOREM 1.2.10.** *Given an arbitrary invertible  $n \times n$  matrix  $A$  there is a factorization*

$$(1.2.11) \quad A = LPDU,$$

where

- (i)  $L$  is unipotent lower triangular,
- (ii)  $P$  is a permutation matrix,
- (iii)  $D$  is diagonal, and
- (iv)  $U$  is unipotent upper triangular.

The factors  $P$  and  $D$  are uniquely determined. Further,  $U$  can be made to satisfy the following condition:

- (1.2.12) Let  $p$  be the permutation of  $\{1, 2, \dots, n\}$  corresponding to the permutation matrix  $P$ . If  $k < l$ , but  $p(k) > p(l)$ , then  $u_{kl}$  (the  $(k, l)$ th entry of  $U$ ) is zero.

If  $U$  satisfies condition (1.2.12), then  $L$  and  $U$  are also uniquely determined.

**REMARKS.** (a) The factor  $L$  in (1.2.11) is related to  $\tilde{L}$  in (1.2.7) by  $L = \tilde{L}^{-1}$ . The factor  $P$  in (1.2.11) is related to  $\tilde{P}$  in (1.2.9) by  $P = \tilde{P}^{-1}$ .

(b) Condition (1.2.12) is just a translation of property (1.2.8)(ii) because, in the notation of (1.2.8), the permutation  $p$  will send  $j$  to  $i_j$ . If  $k < l$  and  $i_k > i_l$ , condition (1.2.8)(ii) says  $\tilde{b}_{i_k l} = 0$ ; but  $\tilde{b}_{i_k l} = b_{kl} = b_{kk}u_{kl}$ . (Here the  $\tilde{b}_{ij}$  are the entries of  $\tilde{B}$ , and likewise for  $B$  and  $U$ .) In particular, the reduction algorithm we have described will produce the factorization (1.2.11) of  $A$  with  $U$  satisfying condition (1.2.12).

1.3. The algorithm we have described is always feasible (at least theoretically, ignoring ill-conditioning) and it leaves nothing to chance or choice. Thus it, and the resulting decomposition (1.2.11), refines the decomposition (1.1.15) of the “generic” matrix  $A$ . It yields a partition of the set of all invertible matrices, i.e., the group  $GL_n$ , into a finite number of sets indexed by permutations. In particular, it yields a precise description of the set of “nongeneric” matrices (those for which the factorization (1.1.15) does not exist). Further, the set of matrices of form (1.2.11) for which  $P$  is a fixed matrix has a very simple structure. Thus Theorem 1.2.10 suggests  $GL_n$  is sort of a “fattened up” version of  $S_n$ , the permutation group on  $n$  letters. In fact the relation, hinted at in Theorem 1.2.10, between  $GL_n$  and  $S_n$  is quite intimate, and generalizes to all semisimple Lie groups. This is the first of the connections between Lie groups and finite groups promised at the outset of this section. This linkage was first brought out in the work of Weyl [Weyl1], so  $S_n$  and its generalizations are called *Weyl groups*.

1.3.1. To strengthen the reader’s belief in the importance of the  $S_n - GL_n$  connection, we point out that the decomposition (1.2.11) has a straightforward and satisfying group-theoretic interpretation. We introduced the group  $\overline{\mathcal{U}}$  of unipotent lower triangular matrices. Let  $\mathcal{U}$  be the group of unipotent upper triangular matrices.

Then the set of  $A$  for which a fixed  $P$  and  $D$  appear in (1.2.11) is simply a  $(\overline{\mathcal{U}}, \mathcal{U})$  double coset. Further, the condition (1.2.12) is simply an irredundancy condition to guarantee that no element of the double coset is written twice. To see this, let  $\{u_{ij} : 1 \leq i < j \leq n\}$  be the above diagonal coordinates of a typical element  $U$  of  $\mathcal{U}$ . Then we can check that

$$(1.3.1.1) \quad \begin{aligned} \mathcal{U} \cap (PD)^{-1} \overline{\mathcal{U}} (PD) &= \mathcal{U} \cap P^{-1} \overline{\mathcal{U}} P \\ &= \{U \in \mathcal{U} : u_{ij} = 0 \text{ if } p(i) > p(j)\}. \end{aligned}$$

This condition is precisely complementary to condition (1.2.12) and only the identity element of  $\mathcal{U}$  can satisfy both.

We can carry this further. Let  $\mathcal{D}$  be the group of invertible diagonal matrices. Then  $\mathcal{D}$  normalizes both  $\mathcal{U}$  and  $\overline{\mathcal{U}}$ , and

$$(1.3.1.2a) \quad \mathcal{B} = \mathcal{D} \cdot \mathcal{U} = \{DU : D \in \mathcal{D}, U \in \mathcal{U}\} = \{UD : D \in \mathcal{D}, u \in \mathcal{U}\}$$

is the group of arbitrary (invertible) upper triangular matrices, and similarly

$$(1.3.1.2b) \quad \overline{\mathcal{B}} = \mathcal{D} \overline{\mathcal{U}}$$

is the group of lower triangular invertible matrices.

Let  $W$  denote the (Weyl) group of permutation matrices. Observe that  $W$  normalizes  $\mathcal{D}$ . Therefore

$$(1.3.1.3) \quad P\mathcal{D} = \{PD : D \in \mathcal{D}\} = \{DP : D \in \mathcal{D}\} = \mathcal{D}P$$

for any  $P$  in  $W$ . Therefore, we see that if we only fix  $P$  in (1.2.11) and let  $L$ ,  $D$ , and  $N$  vary, then we obtain a  $(\overline{\mathcal{U}}, \mathcal{B})$ , or a  $(\overline{\mathcal{B}}, \mathcal{B})$ , or a  $(\overline{\mathcal{B}}, \mathcal{U})$

double coset. Again, we put  $D$  on one side only of  $P$  in (1.2.11) to eliminate redundancy.

Thus (1.2.11) implies the double coset decompositions

$$(1.3.1.4) \quad \begin{aligned} \mathrm{GL}_n &= \overline{\mathcal{B}}W\mathcal{B} = \overline{\mathcal{U}}W\mathcal{B} = \overline{\mathcal{B}}W\mathcal{U} \\ &= \bigcup_{P \in W} \overline{\mathcal{B}}P\mathcal{B} = \bigcup_{P \in W} \overline{\mathcal{U}}P\mathcal{B} = \bigcup_{P \in W} \overline{\mathcal{B}}P\mathcal{U}. \end{aligned}$$

We may also observe that  $\mathcal{U}$  and  $\overline{\mathcal{U}}$ , and  $\mathcal{B}$  and  $\overline{\mathcal{B}}$  are conjugate in  $\mathrm{GL}_n$ . Explicitly, if

$$(1.3.1.5) \quad w_0 = \begin{bmatrix} 0 & & \dots & 1 \\ \vdots & & & \\ & & 1 & \\ & 1 & & \vdots \\ 1 & & \dots & 0 \end{bmatrix}$$

is the permutation matrix corresponding to the permutation which exactly reverses order  $\{1, 2, \dots, n\}$ , then we see easily that  $w_0 = w_0^{-1}$  and

$$(1.3.1.6) \quad \overline{\mathcal{U}} = w_0\mathcal{U}w_0, \quad \overline{\mathcal{B}} = w_0\mathcal{B}w_0.$$

Since  $w_0 \in W$ , we can combine (1.3.1.4) and (1.3.1.6) to obtain

$$(1.3.1.7) \quad \mathrm{GL}_n = \mathcal{B}W\mathcal{B} = \mathcal{U}W\mathcal{B} = \bigcup_{P \in W} \mathcal{U}P\mathcal{B}.$$

This double coset decomposition of  $\mathrm{GL}_n$  into  $(\mathcal{U}, \mathcal{B})$  double cosets parametrized by the (finite) group  $W$  is commonly called the *Bruhat decomposition*. Its analog in a general semisimple or reductive group is a central fact of modern Lie theory. It was described by F. Bruhat [Bruh] for several classes of groups. He was motivated by questions in representation theory. It was also observed in several cases by Gelfand and Naimark [GeNa]. Its existence in a general semisimple group was established by Harish-Chandra [HaCh1], and it is a central feature of the theory of  $(B - N)$ -pairs developed by Tits [Bour, Crtr]. We will give below some examples of its applications.

1.3.2. Another piece of evidence for the importance of  $S_n = W$  in the study of  $\mathrm{GL}_n$  comes from consideration of the dimensions of the double cosets  $\overline{\mathcal{U}}P\mathcal{B}$ . Use of the term “generic” for the elements of the identity coset suggests the following:

- $$(1.3.2.1) \quad \begin{aligned} &(i) \text{ The identity coset } \overline{\mathcal{U}}\mathcal{B} \text{ is an open subvariety of } \\ &\quad \mathrm{GL}_n, \text{ of dimension equal to } n^2, \text{ the same as the dimension of } \mathrm{GL}_n. \\ &(ii) \text{ The other cosets } \overline{\mathcal{U}}P\mathcal{B}, P \neq I, \text{ are subvarieties of} \\ &\quad \text{strictly smaller dimensions.} \end{aligned}$$

These statements are true. Further, the codimension of a coset  $\overline{\mathcal{U}}P\mathcal{B}$  is a familiar combinatorial function on  $S_n$ .

To check statement (1.3.2.1)(i), simply count the number of parameters involved. Elements of the group  $\mathcal{U}$  have  $n(n-1)/2$  lower triangular entries which vary arbitrarily; it has dimension  $n(n-1)/2$ . Similarly  $\mathcal{U}$  has dimension  $n(n-1)/2$ , and since the diagonal entries of  $\mathcal{D}$  are arbitrary subject to being nonzero, it has dimension  $n$ . So the dimension of  $\overline{\mathcal{U}}\mathcal{B} = \overline{\mathcal{U}}\mathcal{D}\mathcal{U}$  is  $2(n(n-1)/2) + n = n^2$ .

On the other hand, in describing the coset  $\overline{\mathcal{U}}P\mathcal{B}$ , we restrict certain of the upper triangular entries  $u_{ij}$ ,  $1 \leq i \leq j \leq n$ , of  $U$  (as in (1.2.11)) to be zero, according to condition (1.2.12). Condition (1.2.12) says we should set  $u_{ij}$ ,  $i < j$ , equal to zero whenever  $p(i) > p(j)$ , i.e., when  $p$  reverses the order of the pair  $(i, j)$ . Thus the total number of parameters needed to describe the coset  $\overline{\mathcal{U}}P\mathcal{B}$  is  $n^2$  minus the number of pairs  $(i, j)$  whose order is reversed by  $p$ ; in other words, the number of pairs reversed by  $p$  is the codimension of  $\overline{\mathcal{U}}P\mathcal{B}$ . But the number of pairs  $(i, j)$  whose order is reversed by  $p$  is a familiar quantity, usually called the *length* of  $p$  [Bour, Hill], and denoted  $l(p)$ , or also  $l(P)$ . Here  $P$  is, as it has been, the permutation matrix representing  $p$ . In summary:

(1.3.2.2) The codimension of the coset  $\overline{\mathcal{U}}P\mathcal{B}$  in  $\mathrm{GL}_n$  is  $l(P)$ , the length of the permutation associated to  $P$ .

Note that  $l(P)$  is also the dimension of the subgroup  $\mathcal{U} \cap P^{-1}\overline{\mathcal{U}}P$ , as described in (1.3.1.1).

1.4. We give here an example of how Theorem 1.2.10 fits into modern mathematics. Consider the set  $G_k^n$  of  $k$ -dimensional linear subspaces of  $n$ -space. The set  $G_k^n$  is called a *Grassmann variety* or *Grassmannian*, after Hermann Grassmann (1809–1877), a German Gymnasiumlehrer whose deep geometrical insight was radically under-appreciated during his lifetime. Note that  $G_1^n$  is the space of lines in  $n$ -space, so is better known as  $(n-1)$ -dimensional projective space; which of course is the backdrop for classical algebraic geometry. The Grassmannians  $G_k^n$  also play a prominent role in classical algebraic geometry [HoPe]. But we will discuss here a more recent development.

If  $Z$  is a  $k$ -dimensional subspace of  $n$ -space, and  $A$  is in  $\mathrm{GL}_n$ , then

$$(1.4.1) \quad A(Z) = \{A(u) : u \in Z\}$$

is another  $k$ -dimensional subspace. Hence,  $\mathrm{GL}_n$  acts by permutations on the Grassmannian  $G_k^n$  of all  $k$ -dimensional spaces. It is an elementary fact in linear algebra that any  $k$ -dimensional subspace can be transformed into any other by recipe (1.4.1), for an appropriate choice of  $A$  in  $\mathrm{GL}_n$ . Thus the action of  $\mathrm{GL}_n$  on  $G_k^n$  is transitive, or in other words,  $G_k^n$  is a *homogeneous space* or *coset space* for  $\mathrm{GL}_n$ . Thus, if we choose a base point  $V_k$  in  $G_k^n$ , we have an identification

$$(1.4.2) \quad G_k^n \simeq \mathrm{GL}_n / \mathcal{P}_k,$$

where  $\mathcal{P}_k$  is the stabilizer of  $V_k$ —the subgroup of  $P \in \mathrm{GL}_n$  such that  $P(V_k) = V_k$ .

Let us choose for  $V_k$  the obvious space of vectors—spanned by the first  $k$  standard basis vectors—of the form

$$[x_1, x_2, \dots, x_k, 0, 0, \dots, 0]^t.$$

The stabilizer  $\mathcal{P}_k$  of this  $V_k$  is easily checked to be the group of matrices of the form

$$(1.4.3) \quad \begin{bmatrix} A_1 & X \\ 0 & A_2 \end{bmatrix}, \quad A_1 \in \mathrm{GL}_k, A_2 \in \mathrm{GL}_{n-k}, X \in M_{k, n-k}.$$

This group  $\mathcal{P}_k$  contains the group  $\mathcal{B}$  of upper triangular matrices. It follows from the Bruhat decomposition (1.3.1.7) that  $\mathrm{GL}_n$  consists of a finite number of  $(\mathcal{U}, \mathcal{P}_k)$  double cosets. Under the projection mapping  $\pi: \mathrm{GL}_n \rightarrow \mathrm{GL}_n/\mathcal{P}_k \simeq G_k^n$ , a  $(\mathcal{U}, \mathcal{P}_k)$  double coset maps to a  $\mathcal{U}$ -orbit. Hence we can conclude from (1.3.1.7) that, under the action of  $\mathcal{U}$ , the Grassmannian  $G_k^n$  breaks up into a finite number of orbits.

A finer analysis, amounting to a continuation of the arguments which led to (1.2.11), (1.3.1.7), and (1.3.2.2), yields the following conclusions.

**THEOREM 1.4.4.** (a) Set  $W_k = W \cap \mathcal{P}_k$  (note that via the isomorphism  $W \simeq S_n$  we have  $W_k \simeq S_k \times S_{n-k}$ ). Then the natural inclusion

$$(1.4.5) \quad W/W_k \rightarrow \mathcal{U} \backslash \mathrm{GL}_n / \mathcal{P}_k$$

is a bijection. Hence, under the action of  $\mathcal{U}$ , the Grassmannian  $G_k^n$  consists of the  $\binom{n}{k}$  orbits

$$(1.4.6) \quad \mathcal{U}\pi(w), \quad w \in W/W_k.$$

Here  $\pi(w)$  is the image of  $w \in W$  in  $G_k^n$  under the projection  $\pi: \mathrm{GL}_n \rightarrow G_k^n \simeq \mathrm{GL}_n/\mathcal{P}_k$ .

(b) Each orbit  $\mathcal{U}\pi(w)$  is a cell, i.e., may be parametrized in a natural way by a vector space. The dimension of the orbit  $\mathcal{U}\pi(w)$  is  $l(wW_k)$ , the length, as an element of  $S_n$ , of the shortest element in the coset  $wW_k$ .<sup>2</sup>

Theorem 1.4.4 has the following consequence. So far the reader may have been thinking of the field of scalars as  $\mathbf{R}$ , the real numbers. We now want them to be  $\mathbf{C}$ , the complex numbers. Then the  $\mathcal{U}$ -orbits  $\mathcal{U}\pi(w)$  are parametrized by complex vector spaces, so their dimensions over  $\mathbf{R}$  are even. Thus in the case of a complex Grassmannian,  $G_k^n(\mathbf{C})$ , Theorem 1.4.4(b) provides a decomposition into even-dimensional cells. General results in algebraic topology [Mass] then guarantee that these cells define a basis for the (rational) homology of  $G_k^n(\mathbf{C})$ . In other words, Theorem 1.4.4(b) provides direct and detailed information on the topology of the complex Grassmannians; it says we may describe the topology of  $G_k^n(\mathbf{C})$  in terms of the combinatorics of  $S_n$ .

This is an interesting result in itself, but it acquires still more significance in view of the basic role that Grassmannians and their cohomology play in the theory of vector bundles. We recall [Ati1, Huse] that a  $k$ -dimensional vector bundle

$$(1.4.7) \quad \begin{array}{c} V \\ \downarrow \\ X \end{array}$$

over a compact Hausdorff space  $X$  gives rise to a map (the “classifying map”)

$$(1.4.8) \quad \gamma_V: X \rightarrow G_k^n$$

for large  $n$ . (Observe there is an obvious injection of  $G_k^n$  into  $G_k^{n+1}$ , so if we have a map (1.4.8) for  $n = n_0$ , we have such a map for all larger  $n$  by composition with these inclusions.) Further, if  $n$  is sufficiently large, then the isomorphism class of  $V$  is determined by the homotopy class of  $\gamma_V$ .

The pullback map of cohomology

$$(1.4.9) \quad \gamma_V^*: H^*(G_k^n) \rightarrow H^*(X)$$

is thus an invariant of the isomorphism class of  $V$ . The inverse images under  $\gamma_V^*$  of certain elements of  $H^*(G_k^n)$  are the “characteristic classes” (Chern classes, Pontrjagin classes, Todd class, etc.) that figure prominently in the Riemann-Roch formula [Hirz], the index formula [AtSi, Gilk] and such matters. These brief indications must suffice for now to suggest how the Bruhat decomposition, which arises in very elementary, classical mathematics, leads directly into sophisticated modern topics.

The discussion given here for  $G_k^n$  extends to all “flag varieties”, homogeneous spaces of the form  $GL_n/\mathcal{P}$ , where  $\mathcal{P}$  is any subgroup containing  $\mathcal{B}$ . Such a variety may be thought of as the set of all nested sequences  $\{0\} = V_0 \subseteq V_1 \subseteq V_2 \subseteq \cdots \subseteq V_k$  of subspaces of specified dimensions. There is an analogous theory for all semisimple groups.

1.5. To begin an account of Lie theory with the Bruhat decomposition, though it is consonant with modern views, is unhistorical in the extreme. In particular, the Bruhat decomposition embodies two aspects of Lie theory which were totally lacking at the outset, but which have come to be seen as essential aspects of the theory as it exists today.

First, it is global. The main point of Theorem 1.2.10 is to refine the “generic” analysis leading to the LDU decomposition to an analysis that describes *all* of  $GL_n$ . The global aspect of the decomposition (1.2.11) (or (1.3.1.7)) was brought out clearly in the previous remark, when the cosets of the Bruhat decomposition were seen to give rise directly to the homology of the Grassmann varieties. By contrast, the original emphasis of Lie theory was local; only the group in a neighborhood of the identity was considered and calculations were mainly carried out in terms of the “infinitesimal group,” now called the Lie algebra. Attention to global features of Lie groups began

to be emphasized in the work of Weyl [Weyl1, Weyl2], who also coined the term "Lie algebra."

Second, it is completely algebraic. The alert reader will have noticed that nowhere in the derivation of decomposition (1.2.11) was anything assumed about the nature of the scalars, except that they formed a field, so the usual operations of addition, subtraction, multiplication, and division could be performed. Further, all the groups we dealt with were algebraic groups, i.e., were defined by algebraic equations, and likewise, the double cosets of the Bruhat decomposition, and the corresponding cells in the Grassmann varieties, are all algebraic varieties, and all our constructions were valid over any field.

Appreciation of the importance of the essentially algebraic nature of the theory of semisimple Lie groups did not develop fully until around 1950. The algebraic viewpoint was developed during the 1950s into the theory of algebraic groups by Chevalley, Borel, Tits, and others [Borel1, 3, 4, Chev1-5, Bour, Tits].

This development had at least two major consequences, both of which were important for the theme of this section, the connection between Lie groups and discrete groups.

1.5.1. First, Chevalley [Chev1] realized that an algebraic version of Lie theory provided a construction of many simple finite groups. If, in our discussion of Gaussian elimination, we take our scalars to belong to a finite field  $F_q$  of  $q$  elements, then we are talking about the finite group  $GL_n(F_q)$ .<sup>3</sup> This is not a simple group, but it almost is. If we restrict the determinant to be 1, then divide out by the group of scalar matrices, we obtain  $PSL_n(F_q)$ , the projective special linear group, which is simple. It had been realized since the 19th century that classical groups (orthogonal, symplectic, unitary, as well as  $GL_n$ ) have forms over finite fields and that, after elimination of some small abelian groups, these groups give rise to finite simple groups. Also, Dickson [Dick] had constructed finite groups corresponding to the exceptional Lie group  $G_2$ . However, Chevalley was first to realize the systematic connection between Lie theory, in its incarnation as the theory of algebraic groups, and the construction of finite simple groups. Refinements of his work yielded all infinite series of finite simple groups, leaving out only what are now known to be 26 "sporadic" simple groups. (Some of these, including the largest, the Fischer-Griess "Monster," have very recently been found also to be related to Lie theory in a more subtle way [FrLM].)

1.5.2. Second, the algebraic point of view led to the conception of a very broad class of discrete groups, known as the *arithmetic groups*. Arithmetic groups are important in algebraic geometry and, especially, are an essential part of a vastly generalized formulation of the theory of automorphic forms that developed during the 1950s and 1960s (see §4.2). The precise definition



of arithmetic group is technical and not especially enlightening,<sup>4</sup> but the rough idea is that it is a group like  $GL_n$  or  $SL_n$ , but whose matrices have integer entries. Thus  $SL_n(\mathbf{Z})$  is a good example. The point is that arithmetic groups are constructed in a methodical way using algebraic groups.

Consider, by contrast, the simple, abstract condition of being a lattice. A subgroup  $\Gamma$  of a Lie group  $G$  is called a *lattice* if

- (1.5.2.1) (i)  $\Gamma$  is discrete.  
 (ii) The coset space  $G/\Gamma$  carries a finite measure invariant under the permutation action of  $G$ .

We observe that if one desires to compare abstract groups with Lie groups, conditions (1.5.2.1) naturally suggest themselves. Imagine we have an abstract group  $\Gamma$ , whose structure we would like to compare with a Lie group  $G$ . To make the comparison, we would want to find a homomorphism  $h$  of  $\Gamma$  into  $G$ . We may as well assume  $h$  is an embedding, for  $G$  will teach us nothing about  $\ker h$ . But if  $h$  is an injection, we may as well identify  $\Gamma$  with  $h(\Gamma)$  and simply consider  $\Gamma$  as a subgroup of  $G$ . As a condition of coherence or compatibility between  $G$  and  $\Gamma$ , to ensure  $G$  is really exercising some control over  $\Gamma$ , we should require  $\Gamma$  to be closed in  $G$ . But if  $\Gamma$  is countable, in particular if  $\Gamma$  is finitely generated, this is equivalent to requiring  $\Gamma$  to be discrete. Finally, finiteness of volume of  $G/\Gamma$  is some guarantee that  $\Gamma$  is big enough to "see" all of  $G$ .<sup>5</sup> To take a very simple example, any abelian group can be embedded in  $\mathbf{R}^n$ , provided only that it is torsion-free and of cardinality not greater than the continuum. But, to be embedded discretely, it must be a free group of  $k$  generators, with  $k \leq n$ ; and, to be a lattice, it must be free of rank  $n$ , i.e., isomorphic to  $\mathbf{Z}^n$ .

Amazingly, it turns out that the abstract concept "lattice in a Lie group" and the concrete construction "arithmetic group" are, though not identical, very closely related. Thus Borel and Harish-Chandra [BoHC] proved that if certain obvious conditions are met, then an arithmetic group is a lattice. For example,  $GL_n(\mathbf{Z})$  is not a lattice in  $GL_n(\mathbf{R})$ , essentially because  $GL_1(\mathbf{Z}) = \{\pm 1\}$  is not a lattice in  $GL_1(\mathbf{R}) = \mathbf{R}^\times$ ; but  $SL_n(\mathbf{Z})$  is a lattice in  $SL_n(\mathbf{R})$ . This fact, though nontrivial, is already implicit in the "reduction theory" of Hermite [Bor13], and the proof of Borel and Harish-Chandra may be regarded as a refinement and generalization of this theory.

It is natural to wonder to what extent the converse is true.<sup>6</sup> It clearly is not true, for a very famous reason—the moduli of Riemann surfaces. Every compact Riemann surface  $X$  (or surface with a finite number of punctures) can be represented in an essentially unique way (i.e., up to conjugacy of  $\Gamma$ ) as a double coset space

$$(1.5.2.2) \quad X \sim SO_2 \backslash SL_2(\mathbf{R}) / \Gamma,$$

where

$$SL_2(\mathbf{R}) = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} ; a, b, c, d \in \mathbf{R}, ad - bc = 1 \right\}$$

is the group of  $2 \times 2$ , real, determinant one matrices,

$$\mathrm{SO}_2 = \left\{ \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} : \theta \in \mathbf{R} \right\}$$

is the subgroup of rotations, and  $\Gamma$  is an appropriate lattice in  $\mathrm{SL}_2(\mathbf{R})$ . But it is well known that the compact Riemann surfaces of a given genus  $g$  form a continuous family which fills out a much studied complex manifold of dimension  $3g - 3$  (the Teichmüller space) [Bers]. As the surface  $X$  moves around continuously, so must the corresponding group  $\Gamma$  (as in (1.5.2.2)). In particular, there are uncountably many of them. Since there are only countably many arithmetic groups—for essentially the same reason that there are only countably many algebraic numbers—most lattices in  $\mathrm{SL}_2(\mathbf{R})$  are not arithmetic.

However, again quite amazingly, this phenomenon of “deformation of lattices” is essentially limited to  $\mathrm{SL}_2$ . Mostow [Most1] showed that if  $G$  is a semisimple Lie group, containing no factors (locally) isomorphic to  $\mathrm{SL}_2(\mathbf{R})$ , then a lattice  $\Gamma$  in  $G$  is *rigid*, that is, if  $\Gamma$  is another lattice in  $G$  and  $\Gamma'$  is sufficiently close (in a fairly straightforward sense) to  $\Gamma$ , then  $\Gamma'$  is actually conjugate to  $\Gamma$  in  $G$ .

Thus if we exclude  $\mathrm{SL}_2(\mathbf{R})$ , we can at least wonder if all lattices are perhaps arithmetic. But it is not true. However, Margulis [Marg1] (see also [Zimm1]) showed that it is very often true. He showed there is a simple condition (that all simple factors have rank at least two) on a semisimple Lie group that guarantees all its lattices are arithmetic. In some sense, “most” Lie groups satisfy Margulis’ criterion. For example, any lattice in  $\mathrm{SL}_n(\mathbf{R})$ ,  $n \geq 3$ , is arithmetic. Thus, the theory of algebraic groups has led to a rather deep understanding of the geometric properties of discrete subgroups of Lie groups. I should mention, however, that the question of which semisimple groups contain nonarithmetic lattices is not yet precisely settled. See [GrPS, Most2] for examples. This is one problem for the future.

**Endnotes.** 1. We can in fact formulate precisely the condition that it will work for a given  $n \times n$  matrix  $A$ . Let  $A_j$  be the leading  $j \times j$  submatrix of  $A$ :

$$A_j = \begin{bmatrix} a_{11}a_{12} & \cdots & a_{1j} \\ a_{21}a_{22} & \cdots & a_{2j} \\ \vdots & & \vdots \\ a_{j1} & \cdots & a_{jj} \end{bmatrix}.$$

Observe that multiplication of  $A$  on the left by a lower triangular unipotent matrix does not change  $\det A_j$ . Thus if  $\{d_i\}$  are the diagonal entries of the matrix  $D$  in (1.1.15), we have  $\det A_j = \det(DU)_j = \prod_{i=1}^j d_i$ . But in order to carry out the reduction process to achieve the factorization, we need  $d_i \neq 0$  for each  $i$ . From our formula relating  $\det A_j$  to the  $d_i$ , we see this

condition may be expressed directly in terms of  $A$  by the requirement that  $\det A_j \neq 0$ ,  $j = 1, 2, \dots, n$ .

2. There is a very beautiful way to count the number of cells of a given dimension. Consider the “Poincaré polynomial”

$$P_{G_k^n}(q) = \sum_{j \geq 0} b_j q^j,$$

where  $b_j$  is the number of cells (i.e.,  $\mathcal{U}$ -orbits) of dimension  $j$ . Consider also the Poincaré polynomial

$$P_W(q) = P_{S_n}(q) = \sum_{w \in W} q^{l(w)}$$

whose coefficients count the number of elements of  $W$  of given length. It turns out that  $P_{W_k}$  divides  $P_W$ , and

$$P_{G_k^n} = P_W / P_{W_k} = P_{S_n} / (P_{S_k} \cdot P_{S_{n-k}}).$$

Explicitly, one has

$$P_{S_n} = \prod_{i=1}^n \frac{q^i - 1}{q - 1}.$$

Hence

$$P_{G_k^n} = \prod_{i=1}^k \frac{q^{n-k+i} - 1}{q^i - 1}.$$

The  $P_{G_k^n}$  are well known in combinatorics as the “Gaussian polynomials” [Proc1, Zeil3].

3. In this situation, the double coset  $\mathcal{U}P\mathcal{B}$  of (1.3.17) can be seen to have order  $q^{n(n-1)/2} (q-1)^n q^{l(P)}$ , where  $l(P)$  is the length of the permutation  $P$ . Summing over  $P$  gives the formula

$$\#(\mathrm{GL}_n(F_q)) = q^m (q-1)^n \sum_{p \in S^n} q^{l(p)} = q^m (q-1)^n P_{S_n}(q),$$

where  $m = n(n-1)/2$  and  $P_{S_n}$  is the Poincaré polynomial of  $S_n$ . Comparison with the easily derived formulas

$$\#(\mathrm{GL}_n(F_q)) = \prod_{i=0}^{n-1} (q^n - q^i)$$

gives a formula for  $P_{S_n}(q)$ . A similar method applies to the Poincaré polynomial for Grassmannians (cf. note 2).

This is a very modest example of the transferral of information from characteristic  $p$  to characteristic zero—a method which, thanks to recent developments in algebraic geometry, has become extremely powerful. Some spectacular examples of it are the Deligne-Lusztig construction of representations of finite Chevalley groups [DeLu], and the Beilinson-Bernstein and Brylinski-Kashiwara proofs of the Kazhdan-Lusztig conjectures [BeBe, BrKa].

4. For the 'satiably curious, here it is: A (linear) algebraic subgroup of  $\mathrm{GL}_n(\mathbb{C})$  is a subgroup which is also an algebraic subvariety, i.e., the set of zeros of a collection of polynomials in  $X_{ij}$ , the coordinate functions on  $M_n$ , and in  $\det^{-1}$ , the reciprocal of the determinant function. Thus  $\mathrm{SL}_n(\mathbb{C})$ , the special linear group, is defined by the equation  $\det g = 1$ ; and  $\mathrm{O}_n(\mathbb{C})$ , the complex orthogonal group, is defined by the equation  $g^t g = 1$ , which can be regarded as the collection of  $n^2$  scalar equations

$$\sum_{j=1}^n g_{ji} g_{jk} = \delta_{ik} \quad \text{for } 1 \leq i, k \leq n.$$

Let  $G \subseteq \mathrm{GL}_n(\mathbb{C})$  be an algebraic subgroup, and let  $I_G$  be the ideal of polynomials vanishing on  $G$ . If we can find a generating set  $\{P_j\}_{j=1}^m$  of polynomials for  $I_G$  such that the coefficients of the  $P_j$  are real, we say  $G$  is *defined over  $\mathbb{R}$* . If we can find such  $P_j$  with coefficients in  $\mathbb{Q}$ , we say  $G$  is *defined over  $\mathbb{Q}$* . If  $G$  is defined over  $\mathbb{R}$ , then

$$G_{\mathbb{R}} = G \cap \mathrm{GL}_n(\mathbb{R})$$

is called the *real points* of  $G$ . Similarly for  $\mathbb{Q}$ .

If  $G$  is a linear algebraic group defined over  $\mathbb{Q}$ , then  $G \cap \mathrm{GL}_n(\mathbb{Z}) = G_{\mathbb{R}} \cap \mathrm{GL}_n(\mathbb{Z})$  is called the subgroup of *integral points* of  $G$ . Two subgroups  $\Gamma_1, \Gamma_2$  of a group  $G$  are called *commensurable* if  $\Gamma_1 \cap \Gamma_2$  has finite index in both  $\Gamma_1$  and  $\Gamma_2$ .

Let  $G_0$  be a Lie group, and  $\Gamma \subseteq G_0$  a discrete subgroup. We say  $\Gamma$  is *arithmetic* if there exist

- (i) a linear algebraic group  $G \subseteq \mathrm{GL}_n(\mathbb{C})$  defined over  $\mathbb{Q}$ , and
- (ii) a homomorphism  $\Psi: G_0 \rightarrow G_{\mathbb{R}}$ , such that
- (iii)  $\ker \Psi$  is compact,
- (iv)  $\mathrm{im} \Psi$  is normal in  $G_{\mathbb{R}}$ ,
- (v)  $G_{\mathbb{R}}/\mathrm{im} \Psi$  is compact, and, most importantly,
- (vi)  $\Psi^{-1}(G \cap \mathrm{GL}_n(\mathbb{Z}))/\ker \Psi$  is commensurable with  $(\Gamma \cdot \ker \Psi)/\ker \Psi$ .

5. It might be thought one should demand that  $G/\Gamma$  be compact; but this would exclude important examples including  $\mathrm{SL}_n(\mathbb{Z})$ . The relaxation of compactness to finiteness of covolume has been very fruitful.

6. A precise question of this nature was formulated by Selberg [Selb1].

## 2. An outline of Lie theory.

2.1. The glue that binds Lie theory together is the notion of a one-parameter group and its infinitesimal generator. For expository purposes, we will consider the one-parameter group first, although this is a revisionist way of proceeding. This discussion can be found in many texts (e.g., [Ster, Gilm, Pont, Hoch, Vara], etc.). But it is so basic, it seems necessary to include it.

Let  $M$  be a manifold. (You can just think of an open set in  $\mathbb{R}^n$  if you wish.) The set  $\mathrm{Diff}(M)$  of diffeomorphisms (smooth, smoothly invertible

4. For the 'satiably curious, here it is: A (linear) algebraic subgroup of  $\mathrm{GL}_n(\mathbb{C})$  is a subgroup which is also an algebraic subvariety, i.e., the set of zeros of a collection of polynomials in  $X_{ij}$ , the coordinate functions on  $M_n$ , and in  $\det^{-1}$ , the reciprocal of the determinant function. Thus  $\mathrm{SL}_n(\mathbb{C})$ , the special linear group, is defined by the equation  $\det g = 1$ ; and  $\mathrm{O}_n(\mathbb{C})$ , the complex orthogonal group, is defined by the equation  $g^t g = 1$ , which can be regarded as the collection of  $n^2$  scalar equations

$$\sum_{j=1}^n g_{ji} g_{jk} = \delta_{ik} \quad \text{for } 1 \leq i, k \leq n.$$

Let  $G \subseteq \mathrm{GL}_n(\mathbb{C})$  be an algebraic subgroup, and let  $I_G$  be the ideal of polynomials vanishing on  $G$ . If we can find a generating set  $\{P_j\}_{j=1}^m$  of polynomials for  $I_G$  such that the coefficients of the  $P_j$  are real, we say  $G$  is *defined over  $\mathbb{R}$* . If we can find such  $P_j$  with coefficients in  $\mathbb{Q}$ , we say  $G$  is *defined over  $\mathbb{Q}$* . If  $G$  is defined over  $\mathbb{R}$ , then

$$G_{\mathbb{R}} = G \cap \mathrm{GL}_n(\mathbb{R})$$

is called the *real points* of  $G$ . Similarly for  $\mathbb{Q}$ .

If  $G$  is a linear algebraic group defined over  $\mathbb{Q}$ , then  $G \cap \mathrm{GL}_n(\mathbb{Z}) = G_{\mathbb{R}} \cap \mathrm{GL}_n(\mathbb{Z})$  is called the subgroup of *integral points* of  $G$ . Two subgroups  $\Gamma_1, \Gamma_2$  of a group  $G$  are called *commensurable* if  $\Gamma_1 \cap \Gamma_2$  has finite index in both  $\Gamma_1$  and  $\Gamma_2$ .

Let  $G_0$  be a Lie group, and  $\Gamma \subseteq G_0$  a discrete subgroup. We say  $\Gamma$  is *arithmetic* if there exist

- (i) a linear algebraic group  $G \subseteq \mathrm{GL}_n(\mathbb{C})$  defined over  $\mathbb{Q}$ , and
- (ii) a homomorphism  $\Psi: G_0 \rightarrow G_{\mathbb{R}}$ , such that
- (iii)  $\ker \Psi$  is compact,
- (iv)  $\mathrm{im} \Psi$  is normal in  $G_{\mathbb{R}}$ ,
- (v)  $G_{\mathbb{R}}/\mathrm{im} \Psi$  is compact, and, most importantly,
- (vi)  $\Psi^{-1}(G \cap \mathrm{GL}_n(\mathbb{Z}))/\ker \Psi$  is commensurable with  $(\Gamma \cdot \ker \Psi)/\ker \Psi$ .

5. It might be thought one should demand that  $G/\Gamma$  be compact; but this would exclude important examples including  $\mathrm{SL}_n(\mathbb{Z})$ . The relaxation of compactness to finiteness of covolume has been very fruitful.

6. A precise question of this nature was formulated by Selberg [Selb1].

## 2. An outline of Lie theory.

2.1. The glue that binds Lie theory together is the notion of a one-parameter group and its infinitesimal generator. For expository purposes, we will consider the one-parameter group first, although this is a revisionist way of proceeding. This discussion can be found in many texts (e.g., [Ster, Gilm, Pont, Hoch, Vara], etc.). But it is so basic, it seems necessary to include it.

Let  $M$  be a manifold. (You can just think of an open set in  $\mathbb{R}^n$  if you wish.) The set  $\mathrm{Diff}(M)$  of diffeomorphisms (smooth, smoothly invertible

mappings) of  $M$  is a group with composition of mappings as the product. The most perspicuous way to think of a one-parameter group of diffeomorphisms of  $M$  is simply as a homomorphism

$$(2.1.1) \quad \begin{aligned} \phi: \mathbf{R} &\rightarrow \text{Diff}(M) \\ t &\rightarrow \phi_t. \end{aligned}$$

However, technical considerations require certain smoothness conditions. These may most conveniently be formulated by requiring that the map

$$(2.1.2) \quad \Phi: \mathbf{R} \times M \rightarrow M, \quad \Phi(t, m) = \phi_t(m), \quad t \in \mathbf{R}, m \in M,$$

be smooth.

Consider a one-parameter group of diffeomorphisms  $\phi_t$  of  $M$ . Fix  $m \in M$ , and consider the curve

$$(2.1.3) \quad \gamma_m(t) = \phi_t(m).$$

Conditions (2.1.2) guarantee this is a smooth curve. It passes through  $m$  at  $t = 0$ . At that point (in time and space), the tangent vector to the curve is

$$(2.1.4) \quad v(m) = \gamma'_m(0) = \frac{d}{dt} \phi_t(m)|_{t=0}.$$

The map  $v: m \rightarrow v(m)$  is a vector field on  $M$ —it assigns to each point  $m$  the tangent vector  $v(m)$ ; condition (2.1.2) guarantees it is a smooth vector field.

**REMARK.** The intuitive geometric connection between the one-parameter group  $\phi_t$  and the vector field  $v$  is most easily seen by taking, temporarily at least,  $M$  to be an open set in  $\mathbf{R}^n$ . Then equation (2.1.4) is equivalent to

$$\phi_t(m) = m + tv(m) + t^2 \varepsilon(t, m),$$

where  $\varepsilon(t)$  is a smooth function of  $t$  and  $m$ . Thus, for small times, the motion  $\phi_t$  displaces  $m$  approximately by the vector  $tv(m)$ ; this approximation becomes more accurate as  $t \rightarrow 0$ . Thus, if we allow ourselves the language of infinitesimals, we may say that after an infinitesimal time  $\varepsilon$ , the point  $m$  moves to  $m + \varepsilon v(m)$ . This was standard parlance in the 19th century, and the actual motion  $m \rightarrow \phi_t(m)$  was thought of as being composed of a very large number of these very small motions  $m \rightarrow m + \varepsilon v(m)$ . (This intuition is justified rigorously by Euler's approximation scheme for solving O.D.E. [GuNi, Zill], etc.) For this reason the vector field  $v$  was called the "infinitesimal generator" of the one-parameter group  $\phi_t$ .

Now consider the tangent vectors to the curve  $\gamma_m$  at other times. We compute

$$(2.1.5) \quad \begin{aligned} \gamma'_m(s) &= \frac{d}{dt} \phi_t(m)|_{t=s} = \frac{d}{dt} ((\phi_{t-s} \cdot \phi_s)(m))|_{t=s} \\ &= \frac{d}{dt} \phi_t(\phi_s(m))|_{t=0} = v(\phi_s(m)). \end{aligned}$$

Thus, the tangent vector to the curve  $\gamma_m$  at any point is the vector assigned by the vector field  $v$  of (2.1.4). In other words, the mapping  $t \rightarrow \gamma_m(t)$  is a solution of the differential equation

$$(2.1.6a) \quad \frac{d\gamma}{dt}(t) = v(\gamma(t)).$$

Since  $\gamma_m(0) = m$ , we see that  $\gamma_m$  is the solution of equation (2.1.6a) with initial condition

$$(2.1.6b) \quad \gamma_m(0) = m.$$

The above reasoning applies for all times  $t$  and all  $m$  in  $M$ . Thus we see that having the one-parameter group  $\phi_t$  gives us solutions, for all time  $t$ , and for all initial conditions  $m$ , of the differential equation (2.1.6a).

On the other hand, suppose we start with the differential equation (2.1.6a). The classical (nineteenth century—contemporaneous with Lie) Existence and Uniqueness Theorem for ordinary differential equations (see, for example, [Ster, LoSt, BiRo]) tells us that given any  $m$ , there is some  $\varepsilon(m) > 0$  such that for  $|t| < \varepsilon(m)$ , there is a solution  $\gamma_m(t)$  of equation (2.1.6a) with initial condition (2.1.6b). Moreover, this solution is unique. An easy extension of this basic result shows that in fact, for each  $m$ , there is some minimum number  $t^-(m) < 0$  and some maximum number  $t^+(m) > 0$  such that  $\gamma_m(t)$  can be defined for  $t^-(m) < t < t^+(m)$ , and if  $t^+(m) < \infty$ , then as  $t \rightarrow t^+(m)$ , the curve  $\gamma_m(t)$  “drops off the edge of the world” in the sense that  $\gamma_m(t)$  has no limit points in  $M$  as  $t \rightarrow t^+(m)$ ; similarly if  $t^-(m) > -\infty$ .

Consider a particular solution curve  $\gamma_m(t)$  of (2.1.6a) with initial condition (2.1.6b). At time  $t = s$ , the curve passes through  $\gamma_m(s)$ . Consider the curve  $\gamma_{m,s}$  obtained from  $\gamma_m$  by shifting the time variable:

$$(2.1.7) \quad \gamma_{m,s}(t) = \gamma_m(s+t).$$

Then we can compute

$$\frac{d}{dt}\gamma_{m,s}(t) = \frac{d}{dt}\gamma_m(s+t) = v(\gamma_m(s+t)) = v(\gamma_{m,s}(t)).$$

Thus  $\gamma_{m,s}$  also satisfies the differential equation (2.1.6a); but  $\gamma_{m,s}$  satisfies the initial condition  $\gamma_{m,s}(0) = \gamma_m(s)$ . From the uniqueness part of the Existence and Uniqueness Theorem, we conclude

$$(2.1.8) \quad \gamma_m(s+t) = \gamma_{m,s}(t) = \gamma_{\gamma_m(s)}(t).$$

Let us now assume that  $\gamma_m(t)$  is defined for all  $t$  and all  $m$ .<sup>1</sup> Then for each  $t$ , we can define a map

$$(2.1.9) \quad \phi_t: M \rightarrow M$$

by the recipe

$$\phi_t(m) = \gamma_m(t).$$

With this change of notation, the relation (2.1.8) turns into

$$\phi_{t+s}(m) = \phi_t(\phi_s(m)),$$

that is,  $\phi_{s+t} = \phi_s \circ \phi_t$ . Since  $\phi_0$  is clearly the identity, we conclude  $\phi_t$  and  $\phi_{-t}$  are mutually inverse mappings. Hence each map  $\phi_t$  is actually bijective and the map  $t \rightarrow \phi_t$  is a homomorphism from  $\mathbf{R}$  to the group of permutations of the points of  $M$ . Further, the Existence and Uniqueness Theorem has some standard complements concerning smoothness in the initial conditions which guarantee that the maps  $\phi_t$  are smooth, hence diffeomorphisms, and even that the map  $\phi$  defined as in (2.1.2) is smooth. Hence the  $\phi_t$  form a one-parameter group of diffeomorphisms of  $M$ .

To sum up, we can enunciate the following correspondence principle, which amounts to a geometric/group theoretic interpretation of the Existence and Uniqueness Theorem for O.D.E.

(2.1.10) To every one-parameter group  $\phi_t$  of diffeomorphisms of a manifold  $M$  is associated a vector field  $v$ , the “infinitesimal generator” of  $\phi_t$ , by equation (2.1.4). Knowledge of  $\phi_t$  is equivalent to the ability to solve, for all initial values  $m$  and all times  $t$ , the differential equations (2.1.6) associated to  $v$ .

Thus there is a one-to-one correspondence between one-parameter groups acting on  $M$  and certain vector fields on  $M$ , namely those for which the equations (2.1.6) can be integrated for all time. (If  $M$  is compact, this will be all vector fields.) In Lie’s time, one was not so fastidious about the global requirement “for all  $m$  for all time,” so one considered simply that there was a one-to-one correspondence between “one-parameter groups” (in the 19th century sense) and vector fields. Today, one achieves this one-to-one correspondence by replacing the one-parameter group by the one-parameter “local group” or “pseudo-group” [GuSt2]. This is not a group but a collection of mappings trying to fit together to be a group. It is the obvious formalization of what you get from the Existence and Uniqueness Theorem. It is a rather cumbersome technical notion which attempts, with only partial success, to restore to us the Eden we lost when we achieved awareness of global problems.

2.2. A fundamental class of examples of one-parameter groups is obtained by taking  $M$  simply to be a vector space, and requiring the  $\phi_t$  to be linear transformations. Since the infinitesimal generator  $v$  of  $\phi_t$  is obtained as a limit,

$$(2.2.1) \quad v(m) = \lim_{t \rightarrow 0} \frac{\phi_t(m) - m}{t} = \lim_{t \rightarrow 0} \left( \frac{\phi_t - 1}{t} \right) (m), \quad m \in M,$$

we see that  $m \rightarrow v(m)$  is likewise a linear transformation. Let us call it  $A$ . Thus

$$(2.2.2) \quad A = \lim_{t \rightarrow 0} \frac{\phi_t - 1}{t},$$



where this limit is taken in the algebra  $\text{End}(M)$  of matrices on  $M$ . With this notation, the differential equations (2.1.6) specialize to

$$(2.2.3) \quad \frac{d\gamma}{dt} = A(\gamma).$$

Equation (2.2.3) will be recognized as a system of constant-coefficient homogeneous linear differential equations, such as occupy a large chunk of introductory courses on ordinary differential equations [GuNi, BoDP, Zill], and form the basis of linear system theory [TiBo, ZaDe].

We know how to solve equations (2.2.3) explicitly in terms of the matrix  $A$ . We form  $\exp A$ , the exponential of  $A$ , by means of the familiar power series for  $\exp$ :

$$(2.2.4) \quad \exp A = 1 + A + \frac{A^2}{2} + \frac{A^3}{6} + \cdots + \frac{A^k}{k!} + \cdots.$$

Then termwise differentiation of the function  $t \rightarrow \exp tA$  yields the equation

$$(2.2.5) \quad \frac{d}{dt}(\exp tA) = A \exp tA.$$

It follows that

$$(2.2.6) \quad \gamma_m(t) = \exp tA(m)$$

is a solution of (2.2.3) with initial value  $m$  at  $t = 0$ . Thus, in this special case, we can recover the one-parameter group from  $A$  defined by (2.2.2) by

$$(2.2.7) \quad \phi_t = \exp tA.$$

Because of formula (2.2.7), one often abuses terminology and calls  $A$  (rather than the associated vector field which assigns  $A(v)$  to  $v$ ) the infinitesimal generator of  $\phi_t$ . Also, because of strong analogies between this special case and the general one-parameter group, one often refers to the procedure of constructing a one-parameter group from the vector field which is its infinitesimal generator as *exponentiating* the vector field.

2.3. We now have a grasp of the bedrock of Lie theory, the connection between a one-parameter group and its infinitesimal generator. The next step, which is the fundamental insight of Lie theory, is how to combine several one-parameter groups into a multi- (but finite!) parameter group—a Lie group. Roughly speaking, one finds that a Lie group is a very coherent collection of one-parameter subgroups.

To firm up ideas, we imagine, in analogy with formula (2.1.2), that we have a manifold  $M$ , and an auxiliary manifold  $G$ , which is parametrizing a group of diffeomorphisms of  $M$ . Thus we have a mapping

$$(2.3.1) \quad \Phi: G \times M \rightarrow M,$$

which we take to be smooth, such that for each  $g \in G$  the map

$$(2.3.2) \quad \phi_g: M \rightarrow M, \quad \phi_g(m) = \Phi(g, m),$$

is a diffeomorphism of  $M$ , and such that the maps  $\phi_g$  form a group; the composition of two of them is a third one, the identity is one, the inverse of one is one, etc. An important example of such a  $G$  occurs when  $M$  is a vector space, and  $G = \text{GL}(M)$  is the group of all invertible linear transformations of  $M$ .

Inside the group  $G$  there will be various one-parameter subgroups, to each of which corresponds a unique infinitesimal generator. A priori these infinitesimal generators are just a set of vector fields. Let us call this set  $\text{Lie}(G)$ . The magic comes in realizing that in fact this seemingly rather unwieldy object, the collection of infinitesimal generators of one-parameter subgroups of  $G$ , has a very precise structure: it is, first, a real vector space; and in addition, it has defined on it a skew-symmetric product—the Lie bracket. (The modern approach to Lie groups, which defines  $\text{Lie}(G)$  as the space of left-invariant vector fields on  $G$ , makes these facts virtually automatic. It is commendable in its efficiency, but it takes a lot of the wonder out of the story.)

The piece of algebraic structure on  $\text{Lie}(G)$  that is easiest to understand is scalar multiplication. If  $t \rightarrow \phi_t$  is a one-parameter group of diffeomorphisms, with infinitesimal generator  $v$ , then  $t \rightarrow \phi_{st}$ ,  $s \in \mathbf{R}$ , is obviously also a one-parameter group of diffeomorphisms, and its infinitesimal generator is easily checked to be  $sv$ . Thus the set  $\text{Lie}(G)$  is closed under multiplication by scalars.

The next observation is that the infinitesimal analog of multiplication of one-parameter groups is simply addition of vector fields. This is easily seen by the following formal, purely local computation. Let  $\phi_t$  and  $\psi_t$  be two one-parameter groups, with infinitesimal generators  $v(m)$  and  $u(m)$ , acting on a region  $M$  in  $\mathbf{R}^n$ . Then for small  $t$ , we have

$$\phi_t(m) = m + tv(m) + t^2\varepsilon_1(m, t), \quad \psi_t(m) = m + tu(m) + t^2\varepsilon_2(m, t).$$

Hence

$$\begin{aligned} (\psi_t \circ \phi_t)(m) &= \psi_t(m + tv(m) + t^2\varepsilon_1(m, t)) \\ &= m + tv(m) + t^2\varepsilon_1(m, t) \\ &\quad + tu(m + tv(m) + t^2\varepsilon_1(m, t)) + t^2\varepsilon_2 \\ &= m + t(v(m) + u(m)) + t^2\varepsilon(m, t). \end{aligned} \tag{2.3.3}$$

Thus  $t \rightarrow \psi_t \circ \phi_t(m)$  is a curve whose tangent vector at  $m$  is  $v(m) + u(m)$ . Of course,  $t \rightarrow \psi_t \circ \phi_t$  is not usually a one-parameter group, but this calculation leads us to hope that, if  $\{\phi_t\}$  and  $\{\psi_t\}$  are subgroups of the group  $G$  of (2.3.1), then there would also be within  $G$  a one-parameter group with  $v + u$  as infinitesimal generator. It is indeed so. In the case when  $G$  is the group  $\text{GL}(V)$  of a real vector space  $V$  (or any closed subgroup thereof), this is guaranteed by the Trotter product formula [Howe7]:

$$\exp(A + B) = \lim_{n \rightarrow \infty} (\exp(A/n) \exp(B/n))^n. \tag{2.3.4}$$

The correspondence between “infinitesimal composition” and addition of infinitesimal generators is very nice, but it leaves us with an enigma. Vector addition is a very faceless operation; for example, the only isomorphism invariant of vector spaces is their dimension. The simple-minded operation of vector addition cannot begin to reflect the extremely rich possibilities for group laws of Lie groups.

Thus we need to put more structure on our infinitesimal generators. A way to do this is suggested by the observation that an obvious way in which vector addition fails to capture general group laws is that it fails to be non-commutative. We could thus ask for a way to reflect the noncommutativity of a group in the infinitesimal generators of its one-parameter subgroups.

A plausible way to do this is to study the commutators of one-parameter groups. This turns out to be an excellent choice. It essentially involves refining calculation (2.3.3) to second order:

$$\begin{aligned}
 \psi_t \circ \phi_t(m) &= \psi_t(m + tv(m) + t^2 \varepsilon_1(m, t)) \\
 &= m + tv(m) + t^2 \varepsilon_1(m, t) + tu(m + tv(m) + t^2 \varepsilon_1(m, t)) \\
 (2.3.5) \quad &+ t^2 \varepsilon_2(m + tv(m) + t^2 \varepsilon_1(m, t)) \\
 &= m + tv(m) + tu(m) + t^2 \partial_{v(m)} u(m) + t^2 \varepsilon_1(m, s) \\
 &+ t^2 \varepsilon_2(m, t) + t^3 \eta,
 \end{aligned}$$

where  $\eta$  is an appropriate smooth function and

$$(2.3.6) \quad \partial_{v(m)}(u)(m) = \lim_{t \rightarrow 0} \frac{u(m + tv(m)) - u(m)}{t}$$

is the directional derivative of  $u$  at  $m$  in the direction of  $v(m)$ . The term  $\partial_{v(m)}(u)(m)$  is not the only second order term, but it is the first term which reflects the interaction between  $\psi_t$  and  $\phi_t$ , and in particular is the only second order term which depends on the order of composition of  $\phi_t$  and  $\psi_t$ . Thus, when we compute the commutator, we find

$$(2.3.7) \quad \psi_t \circ \phi_t \circ \psi_{-t} \circ \phi_{-t}(m) = m + t^2(\partial_{v(m)}(u)(m) - \partial_{u(m)}(v)(m)) + t^3 \tilde{\eta}.$$

Thus, although the curve  $t \rightarrow \psi_t \circ \phi_t \circ \psi_{-t} \circ \phi_{-t}$  is not necessarily smooth (it may have a cusp at  $t = 0$ ), its geometric tangent vector at  $t = 0$  is

$$(2.3.8) \quad \partial_{v(m)}(u)(m) - \partial_{u(m)}(v)(m).$$

A more rigorous result valid for pairs of matrices is the commutator formula [Howe7]:

$$(2.3.9) \quad \lim_{n \rightarrow \infty} (\exp(A/n) \exp(B/n) \exp(-A/n) \exp(-B/n))^n = \exp[A, B],$$

where

$$(2.3.10) \quad [A, B] = AB - BA$$

is the *commutator* of  $A$  and  $B$ .

We can see from (2.3.3) and (2.3.7), in a formal way (which was good enough for Lie), that if we have a group with the structure of a differentiable manifold acting on another manifold, then the set of infinitesimal generators of its one-parameter groups should be a vector space endowed with an antisymmetric product, given by (2.3.8) (which is now generally referred to as the *Lie bracket* of the vector fields  $u$  and  $v$ ). With more work than we have done here, this can be shown rigorously. From formulas (2.3.4) and (2.3.9), we can confidently make the more modest assertion that the set of infinitesimal generators (in the sense of formula (2.2.7) and the remark following) of one-parameter groups of a closed subgroup of  $\mathrm{GL}_n(\mathbf{R})$  forms a linear subspace of the space  $M_n(\mathbf{R})$  of  $n \times n$  matrices, and is closed under the commutator operation (2.3.10). In other words, the set of infinitesimal generators of a Lie group (of the concrete sorts we have been discussing) forms what is now called a Lie algebra. (This terminology was introduced by Hermann Weyl; the original term was “infinitesimal group.”)

2.4. The incredible thing is that this bilinear product, the Lie bracket or commutator, virtually determines the group that gives rise to it. When one considers this, and then the tight control that Lie theory exercises over finite group theory (briefly described in §1.5), the tight control that finite reflection groups exercise over Lie theory (briefly described below in §§2.9, 2.10), and the manifold applications of Lie theory within mathematics and to physics (see §§3.1 and 4), it is hard to avoid a sense of awe.

To see how the commutator controls the group law, consider the following elementary calculations in  $M_n(\mathbf{R})$ . For matrices  $A, B$ , set

$$(2.4.1) \quad \begin{aligned} L_A(B) &= AB, & R_A(B) &= BA, \\ \mathrm{ad}_A(B) &= [A, B] = (L_A - R_A)(B). \end{aligned}$$

Observe that the maps

$$(2.4.2) \quad L: A \rightarrow L_A, \quad R: A \rightarrow R_A$$

are, respectively, a homomorphism and an antihomomorphism of  $M_n(\mathbf{R})$  into  $\mathrm{End}(M_n(\mathbf{R}))$ . In particular, if  $P$  is any polynomial in one variable, then

$$(2.4.3) \quad P(L_A) = L_{P(A)}, \quad P(R_A) = R_{P(A)}.$$

These identities extend to convergent power series. In particular,

$$(2.4.4) \quad L_{\exp(tA)} R_{\exp(tB)}: C \rightarrow \exp(tA) C \exp(tB), \quad C \in M_n(\mathbf{R}),$$

is a one-parameter group of linear transformations of  $M_n(\mathbf{R})$ . One easily computes that its infinitesimal generator is

$$(2.4.5) \quad L_A + R_B.$$

Taking  $B = -A$  gives the famous formula

$$(2.4.6) \quad \exp(tA) C \exp(tA)^{-1} = \exp(\mathrm{ad}_A)(C).$$

If we follow the common practice of denoting the action of  $GL_n$  on  $M_n$  by conjugation as  $\text{Ad}$ :

$$(2.4.7) \quad \text{Ad } g(B) = gBg^{-1},$$

then we can write

$$(2.4.8) \quad \text{Ad exp}(tA) = \exp(t \text{ad}_A).$$

Consider the  $n$ th power mapping  $A \rightarrow A^n$ . This is a smooth mapping from  $M_n(\mathbf{R})$  to itself. Consider its derivative, which we will denote by  $DA^n$ . For each point  $A$ ,  $DA^n$  is a linear map from  $M_n(\mathbf{R})$  to itself, defined in the standard way (cf. [Lang2, LoSt], etc.),

$$(2.4.9) \quad DA^n(B) = \lim_{t \rightarrow 0} \frac{(A + tB)^n - A^n}{t}.$$

By a computation redolent of freshman calculus, we find

$$(2.4.10) \quad DA^n = \sum L_{A^k} R_{A^{n-k-1}} = \sum (L_A)^k (R_A)^{n-k-1}.$$

Multiplying by  $\text{ad } A$  gives

$$(2.4.11) \quad (\text{ad } A)(DA^n) = L_{A^n} - R_{A^n}.$$

Taking linear combinations over various  $n$  gives

$$(2.4.12) \quad \text{ad } A(DP) = L_{P(A)} - R_{P(A)}$$

for any one-variable polynomial  $P$ . This identity extends to convergent power series. In particular,

$$(2.4.13) \quad \begin{aligned} (\text{ad } A)(D \exp A) &= L_{\exp A} - R_{\exp A} = (L_{\exp A} (R_{\exp A})^{-1} - 1) R_{\exp A} \\ &= (\exp(\text{ad}_A) - 1) R_{\exp A}. \end{aligned}$$

Formulas (2.4.11)–(2.4.13) are simply a convenient means to express some formal identities in power series in  $R_A$  and  $L_A$ . Consequently, we may divide (2.4.13) by  $\text{ad } A$  to obtain

$$(2.4.14) \quad D \exp A = \eta(\text{ad } A) R_{\exp A},$$

where

$$\eta(x) = \frac{\exp x - 1}{x} = 1 + \frac{x}{2} + \frac{x^2}{6} + \cdots + \frac{x^m}{(m+1)!} + \cdots.$$

Consider a product  $\exp B \exp A$  for two matrices  $A, B$ . Since  $\exp$  is analytic and invertible near 1, we know that if  $A, B$  are small enough, there is an analytic function  $C(A, B)$  such that

$$(2.4.15) \quad \exp C(A, B) = \exp B \exp A.$$

Differentiate (2.4.15) with respect to  $B$  near  $B = 0$ . This gives

$$D \exp A (\partial_B C(A, 0)) = B \exp A.$$

Using formula (2.4.14) for  $D \exp A$  gives

$$(2.4.16) \quad \partial_B C(A, 0) = \eta(\operatorname{ad} A)^{-1}(B).$$

Formula (2.4.16) has a very important consequence. The maps  $L_{\exp tB}$  define a one-parameter group of linear maps of  $M_n(\mathbf{R})$ , with associated infinitesimal generator  $v(X) = BX$ . Formula (2.4.16) says, if we use coordinates around 1 (the identity matrix) obtained by pushing forward the usual linear coordinates around 0 via  $\exp$  (so-called exponential coordinates or canonical coordinates), then the infinitesimal generator of  $L_{\exp tB}$  has the form

$$(2.4.17) \quad \tilde{v}(X) = \eta(\operatorname{ad} X)^{-1}(B).^2$$

But we observe this vector field is expressible solely in terms of  $X$ ,  $B$ , and the commutator operation. It follows that we can express the function  $C(A, B)$  of (2.4.15) as a power series in multiple commutators in  $A$  and  $B$ . The terms in this power series can be found explicitly by successive differentiation of (2.4.17). The first few terms are

$$\begin{aligned} C(A, B) = & A + B + \frac{1}{2}[B, A] + \frac{1}{12}([A[A, B]] + [B, [B, A]]) \\ & + \frac{1}{24}[B, [A, [B, A]]] + \cdots \end{aligned}$$

The full formula, known as the *Baker-Campbell-Hausdorff formula* can be found in many places [Jaco1, Serr2, HaSc].

From formulas (2.4.14)–(2.4.17) we can make the following somewhat technical but crucial observation:

$$(2.4.18) \quad \begin{array}{l} \text{Suppose } g \subseteq M_n(\mathbf{R}) \text{ is a Lie subalgebra, i.e., a subspace} \\ \text{of } M_n(\mathbf{R}) \text{ closed under the commutator operation. Then, if} \\ A, B \text{ are in } g \text{ and close enough to } 0, \text{ the element } C(A, B) \\ \text{of formula (2.4.15) is also in } g, \text{ and can be computed strictly} \\ \text{in terms of the commutator operation in } g. \end{array}$$

Thus, in particular, if  $U$  is a small neighborhood of 0 in  $g$ , then  $\exp U$  defines a “local group.” For the general Lie group,<sup>3</sup> this can be established by appealing to Darboux’s Theorem [Ster, Chev3, Vara], a general qualitative result on systems of first order P.D.E., of the same vintage as Lie’s work.

2.5. Let us now stand back and see what we have found out. Let  $G$  be a Lie group. By the *Lie algebra of  $G$*  we understand the set of infinitesimal generators of one-parameter subgroups of  $G$ , endowed with a structure of vector space by means of formula (2.3.3) or (2.3.4), and with the bilinear skew-symmetric Lie bracket by means of formula (2.3.8) or (2.3.10).<sup>4</sup> As above, we denote the Lie algebra of  $G$  by  $\operatorname{Lie}(G)$ . Formulas (2.3.3), (2.3.4), (2.3.7), and (2.3.9) show that the assignment  $G \rightarrow \operatorname{Lie}(G)$  is functorial in the following sense. Let  $\mathfrak{g}$  and  $\mathfrak{h}$  be Lie algebras.<sup>4</sup> A *homomorphism* from  $\mathfrak{g}$  to  $\mathfrak{h}$  is a linear map  $\alpha: \mathfrak{g} \rightarrow \mathfrak{h}$  which takes Lie bracket to Lie bracket. Let  $G$  and

$H$  be Lie groups, and let  $\gamma: G \rightarrow H$  be a (smooth) group homomorphism. Define a mapping

$$(2.5.1) \quad d\gamma: \text{Lie}(G) \rightarrow \text{Lie}(H)$$

by the obvious rule: if  $\beta_t$  is a one-parameter subgroup of  $G$ , with infinitesimal generator  $x$ , then  $d\gamma(x)$  is the infinitesimal generator of the one-parameter subgroup  $\gamma(\beta_t)$ . Formulas (2.3.3), (2.3.4), (2.3.7), and (2.3.9) show that  $d\gamma$  is a homomorphism of Lie algebras. Clearly a composition  $\gamma' \circ \gamma$  of group homomorphisms gives rise to a composition of Lie algebra homomorphisms:

$$(2.5.2) \quad d(\gamma' \circ \gamma) = d\gamma' \circ d\gamma.$$

Thus we have two classes of structures—one, Lie groups, with both geometric and algebraic aspects, and the other, Lie algebras, which are purely algebraic objects. Each of these classes has a notion of homomorphism between objects, and so forms a category. We have a correspondence between the two classes of objects, taking a Lie group  $G$  to its Lie algebra  $\text{Lie}(G)$ . This correspondence takes homomorphisms to homomorphisms and preserves composition, so it is a functor [Jaco2]. It follows from results of Lie or from Ado's Theorem (see Endnote 4) that  $G \rightarrow \text{Lie}(G)$  is surjective—every Lie algebra is the Lie algebra of some Lie group.

Thus to complete our understanding of the correspondence  $G \rightarrow \text{Lie}(G)$  we need to know how many different groups correspond to the same Lie algebra (or isomorphic Lie algebras). To determine this, we use a key technical lemma [Chev3, Serr2, Ster].

**LEMMA 2.5.3.** *Suppose  $G$  is a Lie group, with Lie algebra  $\mathfrak{g}$ , and  $\mathfrak{h} \subseteq \mathfrak{g}$  is a Lie subalgebra. Then there is a connected Lie group  $H$  with Lie algebra  $\mathfrak{h}$ , and an injective homomorphism  $j: H \rightarrow G$  such that  $dj: \mathfrak{h} \rightarrow \mathfrak{g}$  is simply the inclusion map.*

The delicate aspect of this result, of course, is that  $j(H)$ , the subgroup generated by  $\exp \mathfrak{h}$ , may not be closed in  $G$ . Lines of irrational slope in a torus are the familiar example. Thus, to have the correct topological structure,  $H$  must be constructed outside of  $G$ , then injected into  $G$ . The proof of Lemma 2.5.3 is an elaboration of observation (2.4.18). The argument requires some care and is somewhat tedious, but is basically straightforward.

Now consider two connected Lie groups  $G_1$  and  $G_2$  whose Lie algebras are isomorphic. We will abuse notation and denote them by the same letter,  $\mathfrak{g}$ . Then the Lie algebra of the product group  $G_1 \times G_2$  is just the sum  $\mathfrak{g} + \mathfrak{g}$  of two copies of  $\mathfrak{g}$ . The diagonal

$$(2.5.4) \quad \mathfrak{g}_\Delta = \{(x, x) : x \in \mathfrak{g}\}$$

is a Lie subalgebra of  $\mathfrak{g} + \mathfrak{g}$ . Let  $G_\Delta$  be the group and  $j: G_\Delta \rightarrow G_1 \times G_2$  be the homomorphism provided by Lemma 2.5.3 for the subalgebra  $\mathfrak{g}_\Delta$ .

Let  $p_i$  be the projection map from  $G_1 \times G_2$  onto  $G_i$ ,  $i = 1, 2$ . The composition  $p_i \circ j$  is a homomorphism from  $G_\Delta$  to  $G_i$ , and the associated map  $d(p_i \circ j)$  is obviously an isomorphism of Lie algebras. It follows that  $p_i \circ j$  is a diffeomorphism in the neighborhood of the identity. Since it is a homomorphism, we find by translating from the identity to a general point that  $p_i \circ j$  is locally a diffeomorphism at every point of  $G$ . Hence  $p_i \circ j$  is a covering map (cf. [Hu, Tits, Hoch], etc.).

It follows that, if  $G_i$  is simply connected (cf. [Hu, Tits, Hoch], etc.), then  $p_i \circ j$  must be an isomorphism. If both  $G_i$  are simply connected, then  $G_1 \simeq G_\Delta \simeq G_2$ . Thus up to isomorphism there is a unique simply connected group with Lie algebra  $\mathfrak{g}$ .

On the other hand, given any connected group  $G$  with Lie algebra  $\mathfrak{g}$ , it is routine to check that the standard construction (cf. [Hu, Hoch], etc.) of the universal cover  $\tilde{G}$  of  $G$  allows one to define a group structure on  $\tilde{G}$  such that the natural projection map  $\pi: \tilde{G} \rightarrow G$  is a group homomorphism. The kernel of  $\pi$  must be a discrete and normal subgroup of  $\tilde{G}$ ; a simple argument implies that a discrete normal subgroup of a connected group is central.

The above discussion has outlined the main considerations in the proof of the following theorem, which summarizes the main foundational facts of Lie theory (Lie's Theorems 1, 2, 3 and their converses [Gilm, Tits, Vara], etc.).

**THEOREM 2.5.5.** (a) *For each Lie algebra  $\mathfrak{g}$ , there is a unique (up to canonical isomorphism) connected and simply connected Lie group  $\tilde{G}$  with  $\text{Lie}(\tilde{G}) = \mathfrak{g}$ .*

(b) *Further, if  $\mathfrak{g}$  and  $\mathfrak{h}$  are Lie algebras with associated connected and simply connected groups  $\tilde{G}$  and  $\tilde{H}$ , and  $\beta: \mathfrak{g} \rightarrow \mathfrak{h}$  is a homomorphism of Lie algebras, then there is a unique homomorphism of groups  $\alpha: \tilde{G} \rightarrow \tilde{H}$  such that*

$$(2.5.6) \quad \begin{array}{ccc} \mathfrak{g} & \xrightarrow{\beta} & \mathfrak{h} \\ \exp \downarrow & & \downarrow \exp \\ \tilde{G} & \xrightarrow{\alpha} & \tilde{H} \end{array}$$

*commutes, i.e.,  $\beta = d\alpha$ . Conversely, given  $\alpha$  we have seen how to construct  $\beta = d\alpha$ .*

(c) *If  $G$  is another Lie connected group with Lie algebra  $\mathfrak{g}$ , then*

$$(2.5.7) \quad G \simeq \tilde{G}/L$$

*where  $L$  is a discrete subgroup of the center of  $\tilde{G}$ ; given any such  $L$ , the quotient group  $\tilde{G}/L$  is a Lie group with Lie algebra  $\mathfrak{g}$ .*

**REMARKS.** (a) Parts (a) and (b) can be more cryptically summarized as: the functor  $\tilde{G} \rightarrow \text{Lie}(\tilde{G})$  from (the category of) connected, simply connected Lie groups to (the category of) Lie algebras is an equivalence of categories.



(b) Although Ado's Theorem guarantees that any Lie algebra can be embedded in  $M_n(\mathbf{R})$ , it is not true that any Lie group can be embedded in  $GL_n(\mathbf{R})$ . For a given simply connected group  $\tilde{G}$ , it may happen that only proper quotients of  $\tilde{G}$  will embed in  $GL_n(\mathbf{R})$ , or it may happen that only  $\tilde{G}$  itself, and no proper quotients of it, will embed in  $GL_n(\mathbf{R})$ . For example,  $SL_m(\mathbf{C})$  is simply connected. Its center is  $\mathbf{Z}/m\mathbf{Z}$ . Any group covered by  $SL_m(\mathbf{C})$  may be embedded in  $GL_n(\mathbf{C})$  for some  $n$ . The group  $\mathcal{U}$  of unipotent upper triangular real matrices is simply connected. Its center is  $\mathbf{R}$ . No group properly covered by  $\mathcal{U}$  can be embedded in  $GL_n(\mathbf{R})$ . The compact orthogonal group  $SO_m$  has a fundamental group equal to  $\mathbf{Z}/2\mathbf{Z}$ ; its universal cover is the spin group  $Spin_m$ , constructed by means of Clifford algebras [Huse, Jaco2]. The symplectic group  $Sp_{2m}(\mathbf{R})$  in  $2m$  variables (the isometry group of a nondegenerate, skew-symmetric form, cf. §3.2, [Helg2, Arti], etc.) has fundamental group  $\mathbf{Z}$ . No proper cover of it can be embedded in  $GL_n(\mathbf{R})$  for any  $n$ .

(c) The construction which associates to a Lie group  $G$  its Lie algebra  $Lie(G)$  is akin to differentiation: formulas (2.3.3) and (2.3.7) show that the vector space structure on  $Lie(G)$  reflects first derivatives and the Lie bracket is somehow built from second derivatives. Theorem 2.5.5 reveals the remarkable extent to which the essentially linear object  $Lie(G)$  determines the nonlinear object  $G$ . The faithfulness with which  $Lie(G)$  reflects the structure of  $G$  allows one in many situations to replace a calculation on  $G$  with a much simpler calculation on  $Lie(G)$ . This is a key to the power of Lie theory.

(d) Since typically we think of commutativity (as opposed to non-commutativity) as contributing to simplicity, it is interesting to note that in many places in the foundations of Lie theory the presence of commutativity makes life difficult. Existence of a nontrivial center is what makes Ado's Theorem difficult (since when the center is trivial, the adjoint representation  $ad$  is faithful). Theorem 2.5.5(c) makes clear the role of the center of  $\tilde{G}$  in the nonbijectivity of the correspondence  $G \rightarrow Lie(G)$ . The failure of Lie subgroups of a given Lie group to be closed is likewise essentially an abelian phenomenon: the standard example of a winding line on a torus captures its essence.

(e) The classical exponential map  $\exp: \mathbf{R} \rightarrow \mathbf{R}^{+\times}$  is of course an isomorphism of groups. Thus in Lie theory the distinction between the additive group  $\mathbf{R}$  and the multiplicative group  $\mathbf{R}^\times$  is blurred. This blurring is essential to the theory. However, in the theory of algebraic groups, the distinction between additive groups, out of which one builds unipotent groups, and multiplicative groups, which are associated with full reducibility, is very important.

2.6. The first application of the theory summarized in Theorem 2.5.5 is to the structure of Lie groups themselves. One proves structural facts about

Lie algebras, then transfers them to Lie groups by means of Theorem 2.5.5. For example, if  $\mathfrak{g}$  is a Lie algebra, and  $\mathfrak{j} \subseteq \mathfrak{g}$  is a Lie subalgebra such that  $[\mathfrak{g}, \mathfrak{j}] \subseteq \mathfrak{j}$ , we say  $\mathfrak{j}$  is an *ideal* in  $\mathfrak{g}$ . If  $G$  is a connected Lie group and  $J \subseteq G$  is a connected normal Lie subgroup, then  $\text{Lie}(J)$  is an ideal in  $\text{Lie}(G)$ ; the obvious converse also holds. If  $\mathfrak{j} \subseteq \mathfrak{g}$  is an ideal, then the quotient space  $\mathfrak{g}/\mathfrak{j}$  inherits a natural Lie algebra structure from  $\mathfrak{g}$ .

Here are the basic structural facts about Lie algebras. A Lie algebra  $\mathfrak{g}$  is *simple* if it has no ideals other than 0 and itself. The *commutator ideal* is

$$(2.6.1) \quad C(\mathfrak{g}) = \mathfrak{g}^{(2)} = [\mathfrak{g}, \mathfrak{g}] = \text{span of } \{[x, y] : x, y \in \mathfrak{g}\}.$$

The *commutator series*  $C^i(\mathfrak{g})$  is defined by

$$(2.6.2) \quad C^{i+1}(\mathfrak{g}) = C(C^i(\mathfrak{g})).$$

The *descending central series*  $\mathfrak{g}^{(i)}$  is defined by

$$(2.6.3) \quad \mathfrak{g}^{(i+1)} = [\mathfrak{g}, \mathfrak{g}^{(i)}].$$

The Lie algebra  $\mathfrak{g}$  is called *solvable* (in  $i-1$  steps) if  $C^i(\mathfrak{g}) = \{0\}$  for some  $i$ ; and  $\mathfrak{g}$  is called *nilpotent* (in  $i-1$  steps) if  $\mathfrak{g}^{(i)} = \{0\}$  for some  $i$ . If  $\mathfrak{g}^{(2)} = C(\mathfrak{g}) = \{0\}$ , then  $\mathfrak{g}$  is called *commutative* or *abelian*. For a general Lie algebra  $\mathfrak{g}$ , the *radical* of  $\mathfrak{g}$ ,  $R(\mathfrak{g})$ , is the maximum solvable ideal in  $\mathfrak{g}$  (this exists); and the *nilradical* of  $\mathfrak{g}$ ,  $N(\mathfrak{g})$ , is the maximum nilpotent ideal (this also exists). Clearly  $N(\mathfrak{g}) \subseteq R(\mathfrak{g})$ .

The reader may assume that all the terminology above is parallel to the similar group-theoretic terminology.

**THEOREM 2.6.4** (cf. [Hump, Jaco1, Serr2, Vara], etc.). *Let  $\mathfrak{g}$  be a Lie algebra.*

(i) *We can write*

$$(2.6.5) \quad \mathfrak{g} = R(\mathfrak{g}) + \mathfrak{s}_1 + \mathfrak{s}_2 + \mathfrak{s}_3 + \cdots + \mathfrak{s}_k,$$

*where the  $\mathfrak{s}_i$  are simple nonabelian Lie subalgebras of  $\mathfrak{g}$  and the sum is direct. Then*

$$(2.6.6) \quad \mathfrak{g}/R(\mathfrak{g}) = \mathfrak{s}_1 + \mathfrak{s}_2 + \cdots + \mathfrak{s}_k$$

*and the  $\mathfrak{s}_i$  are exactly the minimal simple ideals in  $\mathfrak{g}/R(\mathfrak{g})$ .*

(ii)  *$R(\mathfrak{g})/N(\mathfrak{g})$  is abelian; equivalently  $R(\mathfrak{g})^{(2)} \subseteq N(\mathfrak{g})$ .*

2.7. To flesh out this structure theorem, we would like to describe  $R(\mathfrak{g})$  and the  $\mathfrak{s}_i$ . It turns out that the structure of solvable Lie algebras is too flabby to permit a detailed description of all of them. However, we have a standard example of a solvable Lie algebra, namely  $\mathfrak{b}$ , the Lie algebra of upper triangular matrices (which of course is the Lie algebra of  $\mathcal{B}$ , the group of invertible upper triangular matrices). (Note that the commutator ideal of  $\mathfrak{b}$  is  $\mathfrak{u}$ , the Lie algebra of strictly upper triangular matrices (which is the Lie algebra of  $\mathcal{U}$ , the group of unipotent upper triangular matrices).) In general,

we content ourselves with showing that a general solvable Lie algebra “looks like”  $\mathfrak{b}$  in the following sense. (For the following result we use complex scalars rather than real scalars for the same reason one uses complex scalars in discussing Jordan canonical form. Thus, let  $\mathfrak{b}_{\mathbb{C}}$  denote the Lie algebra of upper triangular matrices with complex entries.)

**THEOREM 2.7.1 (Lie’s Theorem).** *Let  $\mathfrak{g} \subseteq M_n(\mathbb{C})$  be a solvable Lie subalgebra. Then  $\mathfrak{g}$  is conjugate, by an element of  $\mathrm{GL}_n(\mathbb{C})$ , to a subalgebra of  $\mathfrak{b}_{\mathbb{C}}$ .*

**REMARKS.** (a) Lie’s Theorem plays an important role in the representation theory of semisimple Lie algebras. See §3.5, especially Lemma 3.5.3.7.

(b) The group-theoretic version of this, known as the Lie-Kolchin Theorem [Kolc, Serr2], is that a connected solvable Lie subgroup of  $\mathrm{GL}_n(\mathbb{C})$  can be conjugated to be upper triangular. The generalization of this to algebraic groups is the Borel Fixed Point Theorem: a connected algebraic group acting rationally on a complete algebraic variety has a fixed point. This result plays a pivotal role in the theory of algebraic groups, especially the classification of simple algebraic groups [Chev2].

2.8. In contrast to the somewhat loose situation for solvable Lie algebras, the situation for simple Lie algebras is extremely rigid, and the classification of simple complex Lie algebras, due mainly to Killing, is just 100 years old. It is an absolutely gorgeous chapter of mathematics and it continues today to inspire research. There are several excellent accounts of this currently available (cf. [Hump, Jaco1, Serr1], etc.). Here we will discuss its outline, in order to bring to the fore the role played by  $\mathfrak{sl}_2$ , and to emphasize the dominant influence of the geometry of finite reflection groups.

The smallest nonabelian simple Lie algebra is  $\mathfrak{sl}_2$ , the  $2 \times 2$  matrices of trace zero (sometimes known also as the three-dimensional simple Lie algebra or TDS). It has a basis

$$(2.8.1a) \quad e^+ = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad e^- = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad h = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

which satisfy commutation relations

$$(2.8.1b) \quad [h, e^+] = 2e^+, \quad [h, e^-] = -2e^-, \quad [e^+, e^-] = h.$$

The algebra  $\mathfrak{sl}_2$  with its associated group  $\mathrm{SL}_2$  (the group of  $2 \times 2$  matrices of determinant 1) is basic to understanding the whole family of simple Lie algebras. In fact, a careful approach to the structure theory of simple Lie algebras would first provide a detailed analysis of the linear representations of  $\mathfrak{sl}_2$ . We will not do this (see, however, §3.5.1, especially Proposition 3.5.1.9), but let us at least remark that if we set

$$(2.8.2a) \quad w = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = \exp(\pi(e^+ - e^-)/2) \in \mathrm{SL}_2$$

then conjugation by  $w$  interchanges  $e^+$  and  $e^-$  and reverses the sign of  $h$ :

$$(2.8.2b) \quad \operatorname{Ad} w(e^+) = -e^-, \quad \operatorname{Ad} w(e^-) = -e^+, \quad \operatorname{Ad} w(h) = -h.$$

It is probably not too extravagant to say that, just as a general Lie group can be regarded as a coherent family of one-parameter groups, so a (semi)simple Lie group is a very coherent constellation of copies of  $\mathrm{SL}_2$ . Since  $\mathrm{SL}_2(\mathbf{R})$  is much less malleable than  $\mathbf{R}$  itself, the corresponding list of simple Lie algebras (as opposed to all Lie algebras) is rather short and highly structured.

The basic strategy of the classification is to investigate the structure imposed on a simple Lie algebra by its action on itself by conjugation. Thus let  $\mathfrak{g}$  be a simple complex nonabelian Lie algebra, with Lie bracket  $[\cdot, \cdot]$ . For  $x \in \mathfrak{g}$ , recall (cf. formula (2.4.1) or Endnote 4))

$$\operatorname{ad} x(y) = [x, y].$$

The first observation is this: the (generalized) eigenspace decomposition of  $\operatorname{ad} x$  gives a grading on  $\mathfrak{g}$ . Thus suppose  $y_1, y_2$  are eigenvectors for  $\operatorname{ad} x$ , with eigenvalues  $\lambda_1$ . Using the Jacobi identity (Endnote 4) we find

$$\begin{aligned} \operatorname{ad} x([y_1, y_2]) &= [x, [y_1, y_2]] = [[x, y_1], y_2] + [y_1, [x, y_2]] \\ &= (\lambda_1 + \lambda_2)[y_1, y_2]. \end{aligned}$$

That is, the bracket  $[y_1, y_2]$  is an eigenvector for  $\operatorname{ad} x$ , with eigenvalue  $\lambda_1 + \lambda_2$ . With slightly more work, one sees that if

$$(2.8.3) \quad \mathfrak{g} = \sum_{\lambda} \mathfrak{g}(x, \lambda)$$

is the decomposition of  $\mathfrak{g}$  into generalized eigenspaces for  $\operatorname{ad} x$ , then

$$(2.8.4) \quad [\mathfrak{g}(x, \lambda), \mathfrak{g}(x, \mu)] \subseteq \mathfrak{g}(x, \lambda + \mu).$$

Thus the Jacobi identity imposes strong consistency conditions on the Lie bracket. In particular,  $\mathfrak{g}(x, 0)$  is a Lie subalgebra of  $\mathfrak{g}$ , and each  $\mathfrak{g}(x, \lambda)$  is a module for  $\mathfrak{g}(x, 0)$ .

To make maximum use of observation (2.8.4), we would like to find an  $x$  for which the decomposition (2.8.3) is as fine as possible. In fact, there are many such  $x$  and they are easy to find. Consider the characteristic polynomial

$$(2.8.5) \quad \det(\operatorname{ad}(x) - \lambda I) = \sum_{i=0}^l \lambda^i \alpha_i(x), \quad x \in \mathfrak{g},$$

where  $l = \dim \mathfrak{g}$  and the  $\alpha_i(x)$  are appropriate polynomials on  $\mathfrak{g}$ . If  $\alpha_i(x) = 0$  for  $0 \leq i < k$ , but  $\alpha_k(x) \neq 0$ , then  $\dim \mathfrak{g}(x, 0) = k$  (here  $\mathfrak{g}(x, 0)$  is as in (2.8.3) for  $\lambda = 0$ ). Let  $r$  be the smallest number such that the polynomial  $\alpha_r$  on  $\mathfrak{g}$  is not identically zero. We call  $r$  the *rank* of  $\mathfrak{g}$ , and we say  $x \in \mathfrak{g}$  is *regular* if  $\alpha_r(x) \neq 0$ , equivalently if  $\dim \mathfrak{g}(x, 0)$  is as small as possible.

Suppose  $x \in \mathfrak{g}$  is regular. Consider  $y \in \mathfrak{g}(x, 0)$ . We know  $\text{ad } y|_{\mathfrak{g}(x, \lambda)}$   $\subseteq \mathfrak{g}(x, \lambda)$  for each eigenvalue  $\lambda$  of  $\text{ad } x$ . Since  $\text{ad } x|_{\mathfrak{g}(x, \lambda)}$  is invertible for all  $\lambda \neq 0$ , we see that if  $y$  is sufficiently close to  $x$ , then also  $\text{ad } y|_{\mathfrak{g}(x, \lambda)}$  is invertible. (Or, to keep the discussion algebraic, we could observe  $\text{ad}(x + ty)|_{\mathfrak{g}(x, \lambda)}$  will be invertible for all but a finite number of scalars  $t$ .) Since  $\dim \mathfrak{g}(y, 0) \geq \dim \mathfrak{g}(x, 0)$ , if  $\text{ad } y|_{\mathfrak{g}(x, \lambda)}$  is invertible for all  $\lambda \neq 0$ , we must have  $\mathfrak{g}(y, 0) \supseteq \mathfrak{g}(x, 0)$ . Since this is true for an open set of  $y \in \mathfrak{g}(x, 0)$ , it is true for all  $y \in \mathfrak{g}(x, 0)$ . Hence  $\text{ad } y|_{\mathfrak{g}(x, 0)}$  is nilpotent for all  $y \in \mathfrak{g}(x, 0)$ . Hence by Engel's Theorem (cf. [Hump, Jaco1], etc.)  $\mathfrak{g}(x, 0)$  is a nilpotent Lie algebra. It can also easily be checked to be its own normalizer in  $\mathfrak{g}$ . Such subalgebras are called *Cartan subalgebras*. An elementary argument (using the fact that a polynomial equation  $p(v) = 0$  in a complex vector space has a solution set of real codimension 2, hence the complement of the solution set must be connected), shows that, for a complex Lie algebra, all Cartan subalgebras are conjugate (by the adjoint action of the associated Lie group). Hence, although this construction appears to depend on choosing some arbitrary element of  $\mathfrak{g}$ , in fact it is essentially canonical.

By an argument like that for finding simultaneous generalized eigenspaces for commuting operators, we find we can refine decomposition (2.8.3) to a decomposition<sup>5</sup>

$$(2.8.6) \quad \mathfrak{g} = \mathfrak{a} + \sum \mathfrak{g}_\alpha,$$

where  $\mathfrak{a}$  is a Cartan subalgebra of  $\mathfrak{g}$ , and  $\mathfrak{g}_\alpha$  is a simultaneous generalized eigenspace for all  $x \in \mathfrak{a}$ .

Precisely, this means that we have labeled  $\mathfrak{g}_\alpha$  by a linear functional  $\alpha \in \mathfrak{a}^*$ , the dual of  $\mathfrak{a}$ , with the property that, if  $I_{\mathfrak{g}_\alpha}$  denotes the identity map on  $\mathfrak{g}_\alpha$ , then  $\alpha(x)I_{\mathfrak{g}_\alpha} - \text{ad } x|_{\mathfrak{g}_\alpha}$  is nilpotent for all  $x \in \mathfrak{a}$ . The  $\mathfrak{g}_\alpha$  are called *root spaces*, and the  $\alpha \in \mathfrak{a}^*$ ,  $\alpha \neq 0$ , such that  $\mathfrak{g}_\alpha \neq \{0\}$ , are called *roots*. We denote the set of roots by  $\Sigma$ .

To see how  $\mathfrak{sl}_2$  can emerge from this situation, suppose we have a pair of elements  $x \in \mathfrak{g}_\alpha$  and  $y \in \mathfrak{g}_{-\alpha}$ . To keep things as simple as possible, suppose that  $x, y$  are both simultaneous eigenvectors (as opposed to generalized eigenvectors) for  $\mathfrak{a}$ . Consider the bracket  $[x, y]$ . By (2.8.4) it belongs to  $\mathfrak{g}_0$ . Consider the three (exhaustive and mutually exclusive) following possibilities:

$$(2.8.7) \quad \begin{aligned} & \text{(i) } \alpha([x, y]) \neq 0, \\ & \text{(ii) } \alpha([x, y]) = 0 \text{ but } [x, y] \neq 0, \\ & \text{(iii) } [x, y] = 0. \end{aligned}$$

If possibility (iii) holds, then  $x$  and  $y$  span a two-dimensional abelian subalgebra of  $\mathfrak{g}$ . Observe that, since  $\text{ad } x$  and  $\text{ad } y$  commute and are individually nilpotent, the product  $\text{ad } x \text{ ad } y$  will also be nilpotent. If possibility (ii) holds,

then  $z = [x, y]$  will commute with  $x$  and  $y$ . Hence  $x, y$ , and  $z$  span a three-dimensional, two-step nilpotent Lie algebra  $h$ , commonly known as a *Heisenberg Lie algebra* (see §3.1.3). Further, we observe that since  $\operatorname{ad} x$  and  $\operatorname{ad} y$  are individually nilpotent, and  $h$  is nilpotent (hence solvable), the action of  $\operatorname{ad} h$  on  $\mathfrak{g}$  consists of nilpotent operators. In particular, it follows that  $\beta([x, y]) = 0$  for all roots  $\beta$ , not just  $\alpha$ . Finally, suppose that (i) holds. By scaling  $x$  or  $y$  or both, we can arrange that  $\alpha([x, y]) = 2$ . Then comparison with formulas (2.8.1) shows that  $x, y$ , and  $[x, y]$  form a standard basis for a copy of  $\mathfrak{sl}_2$ .

Up to here, our discussion has been completely general and applies to any Lie algebra. A key point is to show that if  $\mathfrak{g}$  is simple and nonabelian then of the three alternatives (2.8.7), only alternative (i) is possible. The usual way to do this is via Cartan's criterion (cf. [Hump, Jaco, Serr2, Vara], etc.). This involves the Killing form. This is the symmetric bilinear form on  $\mathfrak{g}$  defined by

$$(2.8.8) \quad B_K(x, y) = \operatorname{trace}(\operatorname{ad} x \operatorname{ad} y), \quad x, y \in \mathfrak{g}.$$

Cartan's criterion says that the Killing form on a simple nonabelian Lie algebra is nondegenerate. One then sees that when  $\mathfrak{g}$  is decomposed as in (2.8.6), the Killing form must be nondegenerate on  $\mathfrak{a}$ , must be trivial on each  $\mathfrak{g}_\alpha$ , and must pair  $\mathfrak{g}_\alpha$  and  $\mathfrak{g}_{-\alpha}$  nondegenerately. From these basic observations (and a thorough grasp of  $\mathfrak{sl}_2$ ) one can eliminate the occurrence of possibilities (2.8.7)(ii) and (iii). At the same time, one concludes that  $-\alpha$  is a root if  $\alpha$  is, that  $\dim \mathfrak{g}_\alpha = 1$  for all roots  $\alpha$ , and (hence) that  $\mathfrak{a}$  is commutative and the action of  $\mathfrak{a}$  on  $\mathfrak{g}$  by  $\operatorname{ad}$  is diagonalizable.

Thus, for each root  $\alpha$ , one finds that the Lie subalgebra of  $\mathfrak{g}$  generated by  $\mathfrak{g}_\alpha$ , and  $\mathfrak{g}_{-\alpha}$  is a copy of  $\mathfrak{sl}_2$ ; the interaction between these various  $\mathfrak{sl}_2$ 's defines the structure of  $\mathfrak{g}$ . We should remark that the proof [Hump, Jaco1, Serr2] of Cartan's criterion, which underlies the analysis described above, also is based on the anticipation that  $\mathfrak{sl}_2$  will appear inside  $\mathfrak{g}$  in the ways that it does. One could avoid the use of Cartan's criterion by developing more fully the consequences of the trichotomy (2.8.7). Thus at all stages in the analysis of the structure of  $\mathfrak{g}$ , we are relying on properties of  $\mathfrak{sl}_2$ .

2.9. To get a strong grasp on  $\mathfrak{g}$ , we need to understand the structure of the set  $\Sigma$  of roots. This set turns out to have a very tight, highly symmetric structure, imposed by the  $\mathfrak{sl}_2$ 's generated by opposing pairs  $\mathfrak{g}_\alpha, \mathfrak{g}_{-\alpha}$ ,  $\alpha \in \Sigma$ , of root spaces. Let  $L$  be the subgroup of  $\mathfrak{a}^*$  generated by  $\Sigma$ ; the standard name for  $L$  is the *root lattice*. The nondegeneracy of the Killing form implies that  $L$  is a discrete subgroup of  $\mathfrak{a}^*$ , of rank equal to  $\dim \mathfrak{a}^*$  ( $= \dim \mathfrak{a} = \operatorname{rank} \mathfrak{g}$ ). (Also,  $\Sigma$  spans  $\mathfrak{a}^*$ .) Since the Killing form on  $\mathfrak{a}$  is nondegenerate, we can use it to identify  $\mathfrak{a}$  and  $\mathfrak{a}^*$ . Then we can transfer the Killing form to  $\mathfrak{a}^*$ , and restrict it to  $L$ . Thus  $L$  is equipped in a natural way with an inner product. Denote the dualized Killing form by  $B_K^*$ .

For  $\alpha \in \Sigma$ , consider the copy of  $\mathfrak{sl}_2$  generated by  $\mathbf{g}_\alpha$  and  $\mathbf{g}_{-\alpha}$  and consider the corresponding copy of  $SL_2$  obtained by exponentiation, which acts on  $\mathbf{g}$  by conjugation. Let  $w_\alpha$  be the element in this copy of  $SL_2$  corresponding to the element  $w$  of formula (2.8.2). Then  $w_\alpha$  acts on  $\mathbf{g}$ , preserving  $\mathbf{a}$ , so by duality  $w_\alpha$  acts on  $\mathbf{a}^*$ , preserving  $\Sigma$ , hence preserving  $L$ . A computation shows that

$$w_\alpha(\lambda) = \lambda - \frac{2B_K^*(\lambda, \alpha)}{B_K^*(\alpha, \alpha)}\alpha, \quad \lambda \in L \subseteq \mathbf{a}^*.$$

(Recall  $B_K^*$  is the dualized Killing form on  $\mathbf{a}^*$ .) In geometrical terms, this says  $w_\alpha$  is reflection in the hyperplane perpendicular to  $\alpha$ . Let  $W_{\mathbf{g}} = W$  be the group generated by the  $w_\alpha$ . It is called the *Weyl group* (of the pair  $(\mathbf{g}, \mathbf{a})$ , or just of  $\mathbf{g}$ , since  $\mathbf{a}$  is unique up to conjugation). Since  $W$  preserves the finite set of roots  $\Sigma$ , it must be a finite group. For the example  $\mathbf{g} = \mathfrak{sl}_n$ , as described in Endnote 5, the group  $W$  is just  $S_n$ , the symmetric group on  $n$  letters.

Thus we have associated to  $\mathbf{g}$  a finite group  $W$ , which is generated by reflections, and which acts on a lattice  $L$ , preserving a distinguished finite set  $\Sigma$ . These very elementary data determine  $\mathbf{g}$ .

2.10. In fact, just the group  $W$ , acting not on the lattice  $L$  but on its real span  $\mathbf{a}_{\mathbf{R}}^*$ , comes very close to determining  $\mathbf{g}$ , and the classification of finite groups of orthogonal transformations generated by reflections is very beautiful and intimately related to the classification of simple Lie algebras. Since most accounts of the classification mix together the Weyl group and the root system, we would like to make explicit here how much depends on the Weyl group alone.

The idea behind the classification of finite reflection groups is as elementary as it is elegant. Also, it is geometric to its core. Let  $W$  be a finite group acting on  $\mathbf{R}^n$ , and generated by reflections in hyperplanes. Let  $R \subseteq W$  denote the set of reflections. For each reflection  $r \in R$ , let  $H_r$  be the hyperplane fixed by  $r$ . We call the  $H_r$  the *reflection hyperplanes* of  $W$ . The set  $\mathbf{R}^n - \bigcup_{r \in R} H_r$  obtained by deleting the  $H_r$  is a finite union of open convex cones. One such cone  $C$  is called an open *Weyl chamber*; its closure  $\overline{C}$  is called a closed Weyl chamber. Choose one such Weyl chamber  $C_0$ , and call it the *fundamental chamber*. The intersection of  $\overline{C}_0$  with the hyperplane  $H_r$  will be some closed cone in  $H_r$ . Call  $H_r$  a *face plane* of  $C_0$  if  $H_r \cap \overline{C}_0$  has relative interior in  $H_r$ . The intersection  $H_r \cap \overline{C}_0$  will be called a *face* of  $\overline{C}_0$  (or of  $C_0$ ). An easy argument shows that the reflections in the face planes of  $C_0$  generate  $W$ . (More precisely, if  $C$  is any other Weyl chamber, and a line from a general point of  $C_0$  to a general point of  $C$  passes through  $l$  hyperplanes, then  $C_0$  can be moved to  $C$  by a product of  $l$  reflections in the faces of  $C_0$ . Hence the group generated by reflections in the face planes of  $C_0$  acts transitively on the Weyl chambers, hence contains reflections in

all hyperplanes, hence equals  $W$ . A more careful argument, proceeding by induction on word length, shows that  $W$  acts simply transitively on the Weyl chambers [Bour, Hilr].)

Thus we want to understand the relations between the reflections in the various face planes of  $C_0$ . Consider two face planes  $H_r$  and  $H_s$  of  $C_0$ . Then  $H_r \cap H_s$  has codimension 2, and the group generated by  $r$  and  $s$  factors to the plane  $\mathbf{R}^n / (H_r \cap H_s)$ . The planar situation can easily be completely analyzed. The reflections  $r$  and  $s$  generate a dihedral group of order  $2m$ ,  $m \geq 2$ , and the lines  $L_r = H_r / (H_r \cap H_s)$  and  $L_s = H_s / (H_r \cap H_s)$  meet at an angle  $\pi/m$ . See Figure 2.10.1

All the lines in the figure are the images modulo  $H_r \cap H_s$  of hyperplanes  $H_{r'}$ ,  $r' \in R$ , since they are transforms by the group generated by  $r$  and  $s$  of  $L_r$  and  $L_s$ . The angle between any two adjacent lines (which is also the dihedral angle between the corresponding hyperplanes) is always  $\pi/m$ . If  $H_r$  and  $H_s$  are both to bound a common Weyl chamber, the lines  $L_r$  and  $L_s$  must be adjacent.

In particular, the dihedral angle between  $H_r$  and  $H_s$  is always acute or a right angle. Thus, if  $u_r$  and  $u_s$  are the normal unit vectors to  $H_r$  and  $H_s$ , pointing outward from  $C_0$ , the angle between  $u_r$  and  $u_s$  is obtuse, a fact which can be expressed by saying that the dot product  $u_r \cdot u_s$  is nonpositive. From this observation, an easy argument shows that the set of all  $u_r$ , for  $H_r$  a face of  $C_0$ , are independent. Thus, if we assume, without essential loss of generality, that there are no vectors fixed by all of  $W$ , the vectors  $u_r$ , for  $H_r$  a face plane of  $C_0$ , form a basis for  $\mathbf{R}^n$ . Thus  $\overline{C}_0$  is a simplicial cone; precisely, it is the cone generated by the vectors  $-u_r^*$ , where  $u_r^*$  is the basis of  $\mathbf{R}^n$  dual to the basis  $u_r$ ,  $H_r$  a face plane of  $C_0$ . Since the dihedral angles between the faces of  $C_0$  are acute, we call  $C_0$  an *acute simplicial cone*. This geometry of a Weyl chamber is important in other places besides the classification of simple Lie algebras. For example, it is a key ingredient in the Langlands-Vogan classification of irreducible admissible representations [Knap2, Voga1, Wall] (cf. §3.6.4).

We now have the key to the classification of finite reflection groups. Since the external unit normals  $u_r$  to the faces of  $\overline{C}_0$  are a basis for  $\mathbf{R}^n$ , the

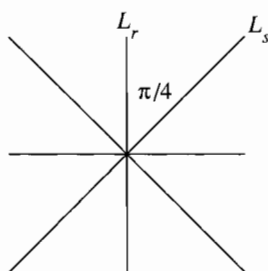


Figure 2.10.1



geometry of  $C_0$  and hence  $W$  itself, is entirely determined by the inner products  $u_r \cdot u_s$  between pairs of external normals. On the other hand, we have seen from our analysis above of the planar case that these inner products are related to the structure of  $W$  via the formula

$$(2.10.2) \quad u_r \cdot u_s = -\cos(\pi/m_{rs}),$$

where  $m_{rs}$  is the order of the product  $rs$  (see Figure 2.10.1).

The problem is to find out what the numbers  $m_{rs}$  can be. An obvious restriction is that the Gram matrix of the  $u_r$ 's, whose entries are the inner products  $u_r \cdot u_s$ , should be positive definite. Here yet another miracle occurs: this simple necessary condition is sufficient to completely determine all possibilities for  $W$ . Moreover, the list of possibilities is quite short, and the computations necessary to limit the list to the actual possibilities are quite easy [Coxe, GrBe].

The result is usually expressed in terms of *Coxeter graphs*. For each face plane  $H_r$  of  $C_0$  one creates a node; then the nodes for  $H_r$  and  $H_s$  are connected by  $m_{rs} - 2$  lines. Alternatively, if  $m_{rs} > 3$ , one labels the line between node  $r$  and node  $s$  by the number  $m_{rs}$ . If two nodes are not connected, then the corresponding reflections commute with each other. If the Coxeter graph of  $W$  is disconnected, then  $W$  is a direct product of the group corresponding to the two pieces. Thus it is only necessary to record the connected Coxeter graphs. Doing so produces the list in Figure 2.10.3.

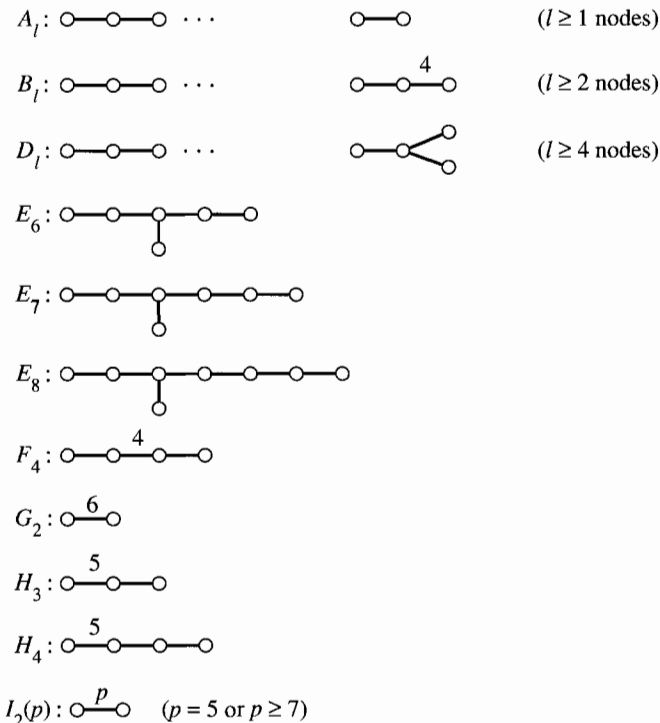


Figure 2.10.3. List of Coxeter graphs.

From our discussion of simple Lie algebras we know that their Weyl groups are contained in this list. However, not all these groups are Weyl groups, because, as we saw, the Weyl groups must leave invariant a lattice  $L$ , and not all the groups in the list above do this.<sup>6</sup> Additionally, for a simple Lie algebra  $\mathfrak{g}$ , we have the data of the root system  $\Sigma$  (see (2.8.6) et infra). From the discussion of the connection between the root system  $\Sigma$  and its Weyl group  $W$ , we see that the elements of  $\Sigma$  are normal to the reflection hyperplanes. But since  $\Sigma$  is contained in the lattice  $L$ , the elements of  $\Sigma$  may not be unit vectors. Instead, they are characterized as the shortest vectors in  $L$  normal to the reflection hyperplanes of  $W$ .<sup>7</sup> It turns out (from considerations of conjugacy) that for the simple root systems only two root lengths are possible, and a change of root length can occur only between nodes which are connected by an even number of lines. To record this extra structure, one refines the Coxeter graph to what is usually called the *Dynkin diagram*, which puts an arrow across junctions with even numbers of lines, pointing in the direction of the longer roots. The resulting list (see Figure 2.10.4) contains four infinite sequences, corresponding to classical groups, and five more “exceptional” groups.

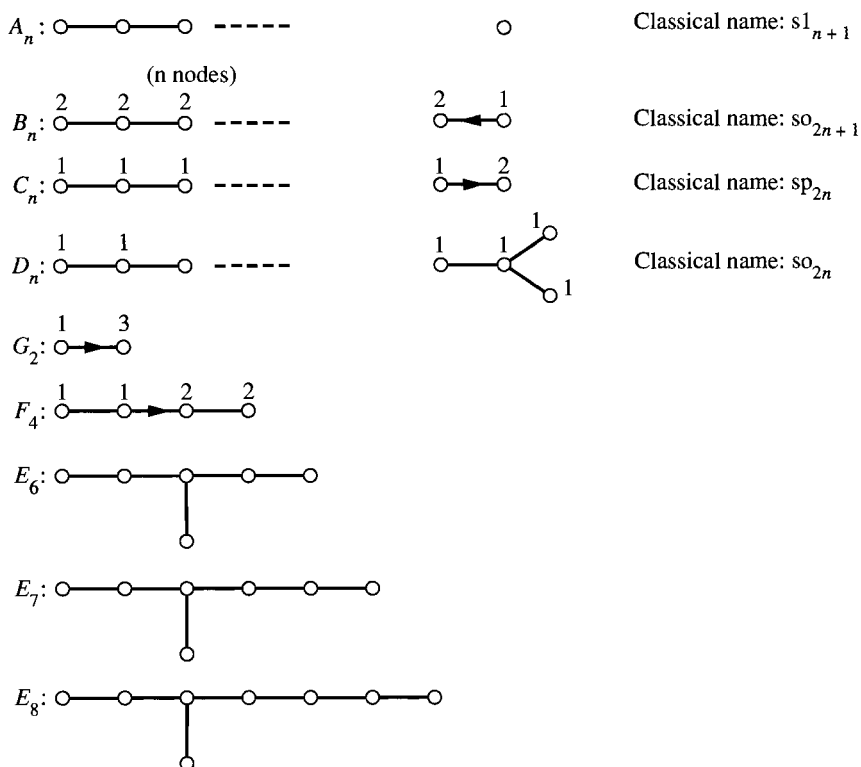
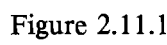


FIGURE 2.10.4. List of Dynkin diagrams.

Here rectangular blocks contain the main classes of objects involved in the theory. Ovals contain important structural information. Arrows proceeding between boxes indicate flow of information. Arrows going in opposite directions between two boxes means the objects in the boxes are each recoverable from the other.



2.12. To emphasize that a given simple Lie algebra is recoverable from its root system (understood as a certain generating subset of a lattice  $L$  equipped with an inner product) we state the theorem of Serre [Hump, Serr1] describing the structure of simple  $\mathfrak{g}$  in terms of generators and relations based on the structure of the root systems.

The data we need are three sets of symbols  $X_\alpha$ ,  $Y_\alpha$ , and  $H_\alpha$ . Each triple is a standard basis for a copy of  $\mathfrak{sl}_2$ . The triples are labeled by a set of *fundamental roots*: one chooses a fundamental Weyl chamber in  $\mathfrak{a}_R^*$ , and the root vectors perpendicular to the face planes of the Weyl chamber are the fundamental roots. For each fundamental root  $\alpha$ , one chooses  $X_\alpha \in \mathfrak{g}_\alpha$  and  $Y_\alpha \in \mathfrak{g}_{-\alpha}$  such that if  $H_\alpha = [X_\alpha, Y_\alpha]$ , then these three elements of  $\mathfrak{g}$  are a standard basis of  $\mathfrak{sl}_2$  (see the discussion following (2.8.7)). To describe how these  $\mathfrak{sl}_2$ 's fit together, we at least need to know the eigenvalues of the  $X_\beta$  under the action of  $H_\alpha$ : set

$$(2.12.1) \quad \text{ad } H_\alpha(X_\beta) = n_{\alpha\beta} X_\beta.$$

From the representation theory of  $\mathfrak{sl}_2$ , one knows the  $n_{\alpha\beta}$  are integers; and of course  $n_{\alpha\alpha} = 2$ . For  $\alpha \neq \beta$ , the integer  $n_{\alpha\beta}$  is nonpositive, and is computed in terms of the geometry of the root system [Hump, Jaco1, Serr1]. The array of integers  $\{n_{\alpha\beta}\}$  is called the *Cartan matrix* of  $\mathfrak{g}$  or of  $\Sigma$ . Serre's result says this is essentially all the data we need to specify  $\mathfrak{g}$ .

**THEOREM 2.12.2 (Serre).** *Let  $\Sigma$  be the root system of a simple Lie algebra  $\mathfrak{g}$ . Let  $F = \{\alpha\}$  be a set of fundamental roots for  $\Sigma$ , and let  $n_{\alpha\beta}$  be defined by equation (2.12.1). Let  $\{X_\alpha, Y_\alpha, H_\alpha : \alpha \in F\}$  be a set of symbols. Let  $\tilde{\mathfrak{g}}$  be the Lie algebra generated by the  $X_\alpha$ ,  $Y_\alpha$ , and  $H_\alpha$ , subject to the following commutation relations:*

$$(2.12.3) \quad \begin{aligned} & \text{(a) } [X_\alpha, Y_\alpha] = H_\alpha, \quad [H_\alpha, H_\beta] = 0, \\ & \quad [X_\alpha, Y_\beta] = 0, \quad \alpha \neq \beta, \\ & \text{(b) } [H_\alpha, X_\beta] = n_{\alpha\beta} X_\beta, \quad [H_\alpha, Y_\beta] = -n_{\alpha\beta} Y_\beta, \\ & \text{(c) } \text{ad } X_\alpha^{1-n_{\alpha\beta}}(X_\beta) = 0, \quad \text{ad } Y_\alpha^{1-n_{\alpha\beta}}(Y_\beta) = 0. \end{aligned}$$

Then  $\tilde{\mathfrak{g}} \simeq \mathfrak{g}$ .

2.13. **REMARKS.** (a) Descriptions, much more involved than Theorem 2.12.2 but in a similar spirit, of the Chevalley groups (over various fields) associated to simple Lie algebras have been given [Crtr, Stei1, 2].

(b) Serre's Theorem has assumed considerable significance in connection with Kac-Moody Lie algebras. In [Kac2] and [Mood1], Kac and Moody independently observed that one could take a "generalized Cartan matrix" of integers  $n_{\alpha\beta}$  (satisfying  $n_{\alpha\alpha} = 2$ ,  $n_{\alpha\beta} \leq 0$ ) and define a Lie algebra  $\tilde{\mathfrak{g}}$  by means of relations (2.12.3). The resulting Lie algebras are infinite dimensional unless the  $n_{\alpha\beta}$  came from the known list of finite-dimensional simple Lie algebras, but they share many of the important properties of finite-dimensional

simple Lie algebras. In particular, each Lie algebra has an associated root system  $\Sigma$  (which will usually be infinite) and a Weyl group  $W$ , which is a reflection group with respect to a (possibly indefinite, even degenerate) inner product. Somewhat later, these infinite-dimensional algebras, especially the “affine” ones, whose associated inner product is positive semidefinite, were realized to be related with a range of fascinating phenomena, including power series identities (the Macdonald identities, Rogers-Ramanujan identities, etc. [Macd2, Kac4, Kost3, LeMi, Lepo2]); completely integrable Hamiltonian systems (Korteweg-de Vries equation, Toda lattice, etc. [AdvM, DJKM1, 2, GoWa1, 2, Syme1, 2, Kost2]); the Fischer-Griess “Monster,” the largest sporadic group [CoNo, FrLM, Kac5]; the representations of graphs [DIRi, Ring, Gabr, Kac6, 7], etc.; and two-dimensional conformal field theories [BePZ, Gawe, Witt]. Work on these various topics is currently proceeding at a furious pace.

(c) Reflection groups, and especially root systems, figure significantly in a variety of contexts outside the classification of simple Lie algebras, some of them quite surprising. We will list a sample of these appearances.

(i) Coxeter [Coxe] was interested in reflection groups because of their connection with regular polytopes. It has long been understood that the symmetry groups of the platonic solids are reflection groups. The famous tessellations of the sphere associated to the regular polyhedra just show the intersection of the sphere with the Weyl chambers for the corresponding reflection group. Similarly, in higher dimensions, one can construct regular polytopes using reflection groups. Especially attractive is the four-dimensional polytope whose three-dimensional faces are 120 regular dodecahedra. Its symmetry group is  $H_4$  in the list (2.10.3). (However,  $H_4$  is not a Weyl group; Weyl groups are associated only to the more mundane regular solids.)

(ii) Reflection groups have an honored place in invariant theory, owing to Chevalley’s theorem [Helg, BeGr, Chev6] complemented by Shephard and Todd [ShTo]:

**THEOREM 2.13.1** (Chevalley, Shephard-Todd). *Let the finite group  $G$  act on the real vector space  $V$ . Let  $P(V)$  be the algebra of polynomials on  $V$ , and let  $P(V)^G$  be the subalgebra of polynomials invariant under the action of  $G$ . Then  $P(V)^G$  is a polynomial algebra (necessarily in  $\dim V$  variables) if and only if  $G$  is generated by reflections.*

The classical example of course is the action of the symmetric group on  $\mathbf{R}^n$  by permutation of the coordinates. For this action, Lagrange’s Theorem [Jaco2, Lang3, Macd1] says the invariant polynomials, usually called symmetric polynomials, are all expressible in terms of the “elementary symmetric polynomials”

$$(2.13.2) \quad \sigma_l(x) = \sum_{1 \leq i_1 < i_2 < \cdots < i_l \leq n} x_{i_1} x_{i_2} \cdots x_{i_l}.$$

In addition to Theorem 2.13.2 there is a very beautiful description of  $P(V)$  as a  $G$ -module in terms of "harmonic polynomials" [Helg1, Chev6]. This structure is involved in significant ways in the representation theory of semisimple groups, and the ideal theory of the universal enveloping algebras of simple Lie algebras. In particular, Theorem 2.13.2 guarantees that the center of the universal enveloping algebra of a simple Lie algebra is a polynomial ring (cf. [Helg2, Wall2, Hump], etc.).

(iii) Much more recent is the application of root systems and reflection groups to problems in linear algebra defined by "representations of graphs". Let  $\Gamma$  be a directed graph: a collection of nodes joined by edges with a sense of direction, i.e., which proceed from one node to another node, but not backwards. (We also permit an edge to connect a node to itself.) A *representation* of the graph  $\Gamma$  is an assignment of a vector space  $V_i$  to each node  $i$ , and a linear transformation  $T_{ij}: V_i \rightarrow V_j$  to each edge from  $i$  to  $j$ . There is an obvious notion of equivalence for two such representations: if  $\{U_i, S_{ij}\}$  is another representation of  $\Gamma$ , it is equivalent to the first one if there are linear isomorphisms  $J_i: U_i \rightarrow V_i$  such that the diagrams

$$\begin{array}{ccc} U_i & \xrightarrow{S_{ij}} & U_j \\ J_i \downarrow & & \downarrow J_j \\ V_i & \xrightarrow{T_{ij}} & V_j \end{array}$$

commute. There is also an obvious notion of direct sum, so the representations of  $\Gamma$  form an abelian category. The problem of representations of graphs is to describe (up to equivalence) the indecomposable objects in the category.

We note that several standard problems of linear algebra are formulable as graph representation problems. For example, the solution of linear equations, solved by Gaussian elimination, the essence of which is the notion of *rank* of a linear transformation, amounts to the representation problem for graph (a), and Jordan canonical form amounts to the solution of the representation problem for the one node graph (b).

Gabriel [Gabr] discovered that a graph  $\Gamma$  has only a finite number of indecomposable representations precisely when the associated undirected graph is a Dynkin diagram of type  $A$ ,  $D$ , or  $E$ . This is remarkable, but the relation goes deeper: the indecomposables are naturally labeled by elements of the root system associated to the Dynkin diagram. Further, Bernstein, Gelfand, and Ponomarev [BeGP] showed there were functors, between representation categories of various  $\Gamma$  with the same Dynkin diagram as undirected graph, which imitated the action of the Weyl group. Kac [Kac6, 7] showed that the representations of more complicated graphs could also be analyzed in terms of root systems and Weyl groups of Kac-Moody Lie algebras.

(iv) The Dynkin diagrams (or Coxeter graphs) of types  $A$ ,  $D$ ,  $E$  also make

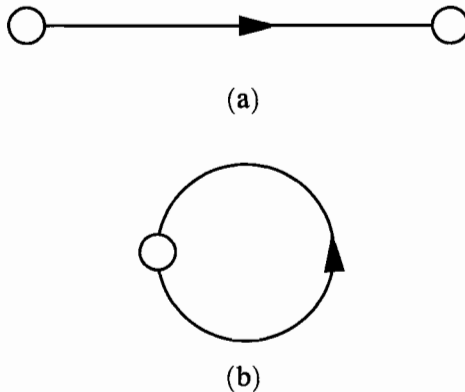


Figure 2.13.3

a fascinating appearance in algebraic geometry, in connection with the classification of isolated singularities of algebraic surfaces. To analyze such a singularity, the algebraic geometer “blows it up” until he achieves a non-singular surface [Lauf, Hirz1, Koda]. In the process, the singular point is replaced by a system of curves. In the case of isolated singularities, the curves are all just projective lines (Riemann spheres—we use “curve” in the sense of a complex one-dimensional (hence real two-dimensional) manifold). To describe the resulting array of curves, one constructs a graph by creating a node for each projective line, and connecting two nodes by the negative of the intersection number of the two associated curves. (A  $p$ -manifold and a  $q$ -manifold inside a  $(p+q)$ -manifold can be expected “generically” to intersect transversally in a finite number of points. To each point of intersection, one can associate a  $\pm 1$  according as the orientation provided by local coordinates on the submanifolds agrees or disagrees with the orientation of the ambient manifold. The sum of these  $\pm 1$ ’s, over all points of intersection, is the intersection number.) It turns out that in the case of simple isolated singularities, the resulting graph is always a Coxeter graph of type  $A_n$ ,  $D_n$ ,  $E_6$ ,  $E_7$ , or  $E_8$  [Arno3, Looi].

This remarkable result has been analyzed in two different ways. One is in terms of finite subgroups of  $SU_2$ . These have been understood since Klein [Klei, GrBe, Miln] and are of course themselves closely related to three-dimensional reflection subgroups. For any  $g \neq 1$  in  $SU_2$ , the only point of  $C^2$  fixed by  $g$  is the origin. Hence, for any finite  $G \subseteq SU_2$  the image of the origin in the quotient space  $C^2/G$  is an isolated singularity. All the isolated surface singularities arise in this way. Further, it is possible to recapture the Coxeter graph of the singularity directly from the representation theory of  $G$ . Let  $\hat{G}$  be the unitary dual of  $G$ —the set of its (equivalence classes of) irreducible (unitary) representations. Let  $\rho_0$  be the given representation of  $G$  on  $C^2$ . Define a graph whose nodes are the elements of  $\hat{G}$ , and such that  $\sigma_1$  and  $\sigma_2$  are connected if and only if  $\sigma_2$  is a component of  $\sigma_1 \otimes \rho_0$ .

(This is a symmetric relation.) The resulting graph is the graph associated geometrically to the singularity  $C^2/G$  [Lamo, Looi].

There is also a direct connection between the simple Lie algebra  $\mathfrak{g}$  with graph of type  $A_n$ ,  $D_n$ , or  $E_n$ , and the singularity with the same graph. We call an element  $x$  of  $\mathfrak{g}$  *nilpotent* if  $\text{ad } x$  is nilpotent. The set of nilpotent elements forms an algebraic subvariety  $\eta$  of  $\mathfrak{g}$ , of codimension equal to the rank of  $\mathfrak{g}$ . Let  $G$  be the Lie group associated to  $\mathfrak{g}$ , and let  $\text{Ad}$  be the action of  $G$  on  $\mathfrak{g}$  by conjugation. Then  $\eta$  may be characterized as the set of zeros of the  $\text{Ad } G$ -invariant polynomials which vanish at the origin. Further,  $\eta$  consists of only finitely many  $\text{Ad } G$ -orbits. (For  $\mathfrak{g} = \mathfrak{gl}_n$ , these are described by Jordan canonical form.) There is one  $G$ -orbit which is open-and-dense, consisting of the so-called *regular nilpotent* elements. The complement in  $\eta$  of the regular nilpotent elements is the singular set of  $\eta$ , and has codimension 2. Denote it by  $\eta_1$ . If one takes a two-dimensional slice in  $\eta$ , at a typical point of  $\eta_1$  and transverse to  $\eta_1$ , this two-dimensional variety will have an isolated singularity, of the type corresponding to the Coxeter graph of  $\mathfrak{g}$  [Brie, Slod].

(v) Finally, to emphasize how innocently, and from what seemingly meager contexts, the root system of simple Lie groups can arise, consider the question of integral quadratic forms. Let  $V$  be a real vector space,  $L \subseteq V$  a lattice, and  $B(\cdot, \cdot)$  a (positive-definite) inner product on  $V$ , such that  $B(l, l')$  is an integer if  $l, l' \in L$ . The question is to describe the isometry classes of such forms, modulo automorphisms of  $L$ .

Suppose  $l_0 \in L$  has  $B$ -norm equal to 1:  $B(l_0, l_0) = 1$ . Then for any  $l \in L$  the difference  $l - B(l, l_0)l_0$  is orthogonal to  $l_0$ . Hence if  $U_0$  is the line through  $l_0$ , and  $U_0^\perp$  is the hyperplane orthogonal to  $U_0$ , then

$$L = (L \cap U_0) + (L \cap U_0^\perp)$$

(orthogonal direct sum).

Hence for purposes of our classification problem, we may as well assume there are no vectors in  $L$  of length 1. Consider next the possibility of vectors  $l$  such that  $B(l, l) = 2$ . Let  $L_1 \subseteq L$  be the lattice spanned by such vectors. Then  $L_1$  decomposes into an orthogonal direct sum of lattices, each one of which is naturally isometric to the root lattice (with appropriately scaled Killing form) of one of the simple Lie algebras, of type  $A$ ,  $D$ , or  $E$ ; and the vectors  $l$  with  $B(l, l) = 2$  form the root system of the appropriate type.

The root lattice of  $E_8$  is particularly significant in this context. Given  $V, L, B$  as above, define

$$(2.13.4) \quad L^* = \{v \in V : B(v, l) \in \mathbb{Z} \text{ for all } l \in L\}.$$

Then  $L^*$  is a lattice, and  $L^{**} = L$ . By our assumption on  $L$ , we have  $L \subseteq L^*$ . The quotient group  $L^*/L$  is clearly an invariant of the isometry class of  $L$ . Of particular interest are the *self-dual* or *unimodular* lattices, for which  $L = L^*$ . Of course  $\mathbb{Z}^n$ , with its usual inner product, is self-dual;



but a more interesting problem is to find an *even* unimodular lattice, i.e., one for which  $B(l, l)$  is even for all  $l \in L$ . It turns out there are none in dimensions less than 8, and that the root lattice of  $E_8$  is the unique example in dimension 8 [FrLM, Sloa, CoSh].

To reinforce the opinion that the facts just recited are not merely curiosities, but are worthy of contemplation, we recall that the Leech lattice, which is the unique even unimodular lattice in 24 dimensions such that  $B(l, l) \geq 4$  for all  $l \neq 0$ , is deeply involved with the sporadic simple groups, especially the Conway groups and the Monster [CoNo, FrLM, Thom].

(d) The classification of simple Lie algebras over fields of positive characteristic is much more delicate than in characteristic zero, because of the failure of Lie's Theorem (Theorem 2.7.1) and related problems. Although, the last word has not been said on this, nearly the last is contained in [StWi], which shows that the types  $A - G$ , plus a family of other algebras, analogous to Lie algebras in characteristic zero which are infinite dimensional, constitute all simple Lie algebras over an algebraically closed field of characteristic  $p \geq 7$ .

**Endnotes.** 1. In other words,  $t^+(m) = +\infty$  and  $t^-(m) = -\infty$  for all  $m$ . Note that equation (2.1.8) implies that  $t^\pm(\gamma_s(m)) = t^\pm(m) - s$ . Hence if there exists  $\varepsilon > 0$  such that  $t^-(m) \leq -\varepsilon$  and  $t^+(m) \geq \varepsilon$  for all points  $m$ , then  $t^\pm(m) = \pm\infty$  for all  $m$ . That is, if we can solve (2.1.6) in a uniform interval for all initial conditions, we can solve it for all time.

2. The function

$$\eta(x)^{-1} = \frac{x}{e^x - 1} = \sum_{n=0}^{\infty} B_n x^{2n} - \frac{x}{2},$$

where the  $B_n$  are the Bernoulli numbers (cf. [Hirz, Lang5], etc.).

3. Some readers may be bothered by the fact that we have not given a formal definition of Lie group. We present one here for them. A Lie group is a smooth manifold  $G$  endowed with a group structure such that the maps

$$\begin{aligned} G \times G &\rightarrow G, & G &\rightarrow G, \\ (x, y) &\rightarrow xy, & x &\rightarrow x^{-1} \end{aligned}$$

of multiplication and inversion (or, equivalently, the single map  $(x, y) \rightarrow xy^{-1}$ ) are smooth. Clearly, by letting  $G$  act on itself by left translations, we can realize  $G$  as a group of diffeomorphisms of a smooth manifold.

4. The question presents itself: will any skew-symmetric product on a vector space define a Lie algebra, in the sense that it arises as the set of infinitesimal generators of a Lie group? The answer is negative. There is an additional identity that needs to be satisfied, the *Jacobi identity*:

$$[A, [B, C]] + [B, [C, A]] + [C, [A, B]] = 0.$$

This is easy to verify for either of our concrete Lie algebras (of vector fields

or matrices). It is an infinitesimal analog of the associative law for group multiplication.

Lie showed that given a vector space with a skew-symmetric product satisfying the Jacobi identity he could construct a group (or what is now called a "local group") with that Lie algebra. Later Ado [**Hump, Jaco, Vara**], showed a vector space with a skew-symmetric product satisfying the Jacobi identity could be realized as a Lie algebra of matrices, i.e., a subspace of  $M_n(\mathbf{R})$ , closed under the commutator operation. Thus we see that a satisfactory definition of abstract Lie algebra, without reference to a group, is: a vector space endowed with a skew-symmetric bilinear form satisfying the Jacobi identity.

The efficiency with which the Jacobi identity captures the essence of Lie algebra structure is shown by the following two observations. First, given a Lie algebra  $\mathfrak{g}$  with bracket operation  $[\cdot, \cdot]$ , define

$$\begin{aligned}\text{ad } x : \mathfrak{g} &\rightarrow \mathfrak{g}, \\ \text{ad } x(y) &= [x, y], \quad x, y \in \mathfrak{g}.\end{aligned}$$

Then  $\text{ad} : \mathfrak{g} \rightarrow \text{End}(\mathfrak{g})$  is a linear map. The Jacobi identity says

$$[\text{ad } x, \text{ad } y] = \text{ad}[x, y]$$

(where the  $[\cdot, \cdot]$  on the left-hand side is the commutator of operators (2.3.10)). That is, the map  $\text{ad}$  preserves Lie bracket, and so is a representation of  $\mathfrak{g}$  on itself. (From formula (2.4.8), we see  $\text{ad}$  is the infinitesimal version of conjugation.) In particular, if  $\mathfrak{g}$  has no center (i.e., no nonzero elements  $x$  such that  $[x, y] = 0$  for all  $y \in \mathfrak{g}$ ), then  $\text{ad}$  gives an isomorphism of  $\mathfrak{g}$  with a Lie algebra of matrices (Ado's Theorem for such  $\mathfrak{g}$ ). In particular, if  $\mathfrak{g}$  is a nonabelian simple Lie algebra, then  $\text{ad}$  provides a faithful matrix representation of  $\mathfrak{g}$ .

Second, the Jacobi identity also says

$$\text{ad } x([y, z]) = [\text{ad } x(y), z] + [y, \text{ad } x(z)].$$

This says that  $\text{ad } x$  is a derivation [**Lang3**] of  $\mathfrak{g}$ . It follows from purely formal properties of  $\exp$  that  $\exp(\text{ad } x)$  must be an automorphism. Thus a Lie algebra structure always comes with a group of automorphisms, provided by conjugation by the associated group. This implies that the more complicated the Lie algebra structure, the more symmetrical it must be. If this observation is too vague to produce the very clean list of simple Lie algebras, at least it is consistent with the existence of such a list.

5. The standard example is  $\mathfrak{gl}_n = M_n(\mathbf{R})$ . One can take for  $\mathfrak{a}$  the diagonal matrices. Then the  $\mathfrak{g}_\alpha$ 's are the lines generated by the matrix units  $E_{ij}$  (cf. formula (1.2.1)). If we use the usual coordinates  $a_i$  on  $\mathfrak{a}$ , then the roots  $\alpha$  are  $a_i - a_j$ .

6. This issue, like so many others, is settled simply by considering the pairwise relations between generators. It is easy to see that a reflection group in the plane generated by  $r$  and  $s$  can only preserve a lattice if  $m_{rs} =$

2, 3, 4, or 6. Comparison of lists (2.10.3) and (2.10.4) reveals it is exactly the Coxeter graphs which have an  $m_{rs}$  other than 2, 3, 4, or 6 which do not survive to become Dynkin diagrams.

7. This is for “reduced” root systems, which is what is encountered in classifying simple complex Lie algebras. For real Lie algebras, nonreduced root systems, e.g.,  $BC_n$ , can also occur [Helg2, Serr1].

**3. Representation theory.** Research into representations (actions on vector spaces via linear transformations) of Lie groups, motivated on one hand by physics [FISz, Mack1, ITGT1–17, Barg3] and on the other by the theory of automorphic forms [GGPS, JaLa, Weill1, BoCa] with deep roots in classical analysis and with strong ties to differential equations, and of course also propelled by its internal dynamics, has been a major part of the mathematical enterprise since roughly World War II. Considering the diversity of motivations, goals, people, and methods involved, the subject displays a remarkable amount of unity. A major source of the unity is the philosophy of the *orbit method* (also known by the more fashionable term *geometric quantization* [Blat, Kiri, Kost1, Sour]). Although we can only sample from the wide range of results that have been established, the overall coherence provided by the viewpoint of the orbit method allows us to convey much more of the subject than would otherwise be possible. An interesting technical point, however, is that the orbit method is almost exclusively a method of *interpretation*, a way of organizing results into a coherent (and often very beautiful) pattern. It provides little in the way of technical tools for proofs or computations. Thus, for example, several of the major results of Harish-Chandra on representations of semisimple groups have found elegant interpretations in terms of the orbit method [Ross1, 2, DuVe, DuHV]. However, these interpretations have provided no short-cuts to Harish-Chandra’s proofs of these results.

A proper discussion of representation theory requires an aggravatingly long technical preparation. We are going to try to ignore that here. For the convenience of the reader, basic definitions and constructions have been summarized in Appendix 1. The discussion below refers to Appendix 1 as necessary. The reader who finds these references too distracting may wish to acquaint himself, at least in a rough way, with Appendix 1 before reading the main body of this section.

**3.1. An example: the quantum harmonic oscillator.** To illustrate the potential uses of representation theory, and its attraction, I can produce no better example than the spectral analysis of the quantum mechanical harmonic oscillator. This is elementary almost to the point of simple-mindedness, yet it contains the seeds of extremely varied developments that form subjects of active current research. In particular, it is basic for the orbit method to be discussed later. Also, it exhibits the extreme elegance of the best Lie algebraic computations.

2, 3, 4, or 6. Comparison of lists (2.10.3) and (2.10.4) reveals it is exactly the Coxeter graphs which have an  $m_{rs}$  other than 2, 3, 4, or 6 which do not survive to become Dynkin diagrams.

7. This is for “reduced” root systems, which is what is encountered in classifying simple complex Lie algebras. For real Lie algebras, nonreduced root systems, e.g.,  $BC_n$ , can also occur [Helg2, Serr1].

**3. Representation theory.** Research into representations (actions on vector spaces via linear transformations) of Lie groups, motivated on one hand by physics [FISz, Mack1, ITGT1–17, Barg3] and on the other by the theory of automorphic forms [GGPS, JaLa, Weill1, BoCa] with deep roots in classical analysis and with strong ties to differential equations, and of course also propelled by its internal dynamics, has been a major part of the mathematical enterprise since roughly World War II. Considering the diversity of motivations, goals, people, and methods involved, the subject displays a remarkable amount of unity. A major source of the unity is the philosophy of the *orbit method* (also known by the more fashionable term *geometric quantization* [Blat, Kiri, Kost1, Sour]). Although we can only sample from the wide range of results that have been established, the overall coherence provided by the viewpoint of the orbit method allows us to convey much more of the subject than would otherwise be possible. An interesting technical point, however, is that the orbit method is almost exclusively a method of *interpretation*, a way of organizing results into a coherent (and often very beautiful) pattern. It provides little in the way of technical tools for proofs or computations. Thus, for example, several of the major results of Harish-Chandra on representations of semisimple groups have found elegant interpretations in terms of the orbit method [Ross1, 2, DuVe, DuHV]. However, these interpretations have provided no short-cuts to Harish-Chandra’s proofs of these results.

A proper discussion of representation theory requires an aggravatingly long technical preparation. We are going to try to ignore that here. For the convenience of the reader, basic definitions and constructions have been summarized in Appendix 1. The discussion below refers to Appendix 1 as necessary. The reader who finds these references too distracting may wish to acquaint himself, at least in a rough way, with Appendix 1 before reading the main body of this section.

**3.1. An example: the quantum harmonic oscillator.** To illustrate the potential uses of representation theory, and its attraction, I can produce no better example than the spectral analysis of the quantum mechanical harmonic oscillator. This is elementary almost to the point of simple-mindedness, yet it contains the seeds of extremely varied developments that form subjects of active current research. In particular, it is basic for the orbit method to be discussed later. Also, it exhibits the extreme elegance of the best Lie algebraic computations.

3.1.1. A quantum mechanical system is defined by a selfadjoint operator called the *Hamiltonian operator* on a Hilbert space  $\mathcal{H}$  [Mack3]. Analysis of the system involves describing the spectral decomposition, especially the eigenvalues and eigenvectors, of the Hamiltonian. For the one-dimensional quantum harmonic oscillator, the Hilbert space is  $L^2(\mathbf{R})$ , and the Hamiltonian is [Shan]

$$(3.1.1.1) \quad T = \frac{d^2}{dx^2} - x^2.$$

To find the spectrum of  $T$ , consider the operators  $p, q$  on  $L^2(\mathbf{R})$  defined by

$$(3.1.1.2) \quad p(f)(x) = \frac{df}{dx}(x), \quad q(f)(x) = ix f(x)$$

for  $f$  sufficiently nice in  $L^2(X)$ . It is easy to check that the four operators  $T, p, q$ , and  $1$ , the identity operator, span a four-dimensional Lie algebra: the commutators

$$(3.1.1.3) \quad [A, B] = AB - BA$$

of two of these operators is a linear combination of some or all of them. Indeed, easy computations show

$$(3.1.1.4) \quad \begin{aligned} \text{(a)} \quad [p, q] &= i, \\ \text{(b)} \quad [T, p] &= -2iq, \quad [T, q] = 2ip, \end{aligned}$$

and of course the commutator of  $1$  with anything is zero.

Let us set

$$(3.1.1.5a) \quad \mathbf{a} = \frac{d}{dx} + x = p - iq, \quad \mathbf{a}^+ = \frac{d}{dx} - x = p + iq.$$

Then we observe

$$(3.1.1.5b) \quad [\mathbf{a}^+, \mathbf{a}] = 2,$$

$$(3.1.1.5c) \quad \mathbf{a}^+ = -\mathbf{a}^*,$$

where  $\mathbf{a}^*$  indicates the operator on  $L^2(\mathbf{R})$  adjoint to  $\mathbf{a}$ , and

$$(3.1.1.5d) \quad T = \frac{1}{2}(\mathbf{a}^+ \mathbf{a} + \mathbf{a} \mathbf{a}^+).$$

Further we can see that the vector

$$(3.1.1.5e) \quad v_0 = e^{-x^2/2}$$

is annihilated by  $\mathbf{a}$ :

$$(3.1.1.5f) \quad \mathbf{a}v_0 = 0.$$

Now let us forget we are dealing with specific operators on  $L^2(\mathbf{R})$ . Let us simply suppose we have some Hilbert space on which are defined two

operators,  $\mathbf{a}$  and  $\mathbf{a}^+$ , satisfying relations (3.1.1.5b,c), such that there is a vector  $v_0$  annihilated by the operator  $\mathbf{a}$ . Define

$$(3.1.1.6) \quad v_j = (\mathbf{a}^+)^j(v_0) = \mathbf{a}^+(v_{j-1}), \quad j = 1, 2, 3, \dots$$

I claim

$$(3.1.1.7) \quad \mathbf{a}(v_j) = -2jv_{j-1}.$$

This may be easily verified by use of the commutator identity

$$[\mathbf{a}, (\mathbf{a}^+)^k] = \sum_{j=0}^{k-1} (\mathbf{a}^+)^j [\mathbf{a}, \mathbf{a}^+] (\mathbf{a}^+)^{k-j-1} = -2k(\mathbf{a}^+)^{k-1}.$$

Using (3.1.1.6) and (3.1.1.7) we can verify that, if  $T$  is defined by formula (3.1.1.5d) then

$$(3.1.1.8) \quad T(v_j) = -(2j+1)v_j.$$

Thus the  $v_j$  are eigenvectors for  $T$ . Since  $T$  is selfadjoint, this means the  $v_j$  are mutually orthogonal. We can even determine the Hilbert space norms of the  $v_j$ 's. If the inner product is denoted by  $(\ , \ )$  we can compute

$$(v_j, v_j) = (\mathbf{a}^+ v_{j-1}, \mathbf{a}^+ v_{j-1}) = -(\mathbf{a}\mathbf{a}^+ v_{j-1}, v_{j-1}) = 2j(v_{j-1}, v_{j-1}).$$

Hence

$$(3.1.1.9) \quad (v_j, v_j) = 2^j j! (v_0, v_0).$$

It follows that if we put

$$(3.1.1.10) \quad u_j = (2^j j! (v_0, v_0))^{-1/2} v_j,$$

then the  $u_j$  form an orthogonal sequence of eigenvectors for  $T$ , and

$$(3.1.1.11) \quad \mathbf{a}u_j = -(2j)^{1/2} u_{j-1}, \quad \mathbf{a}^+ u_j = (2(j+1))^{1/2} u_{j+1}.$$

If we now return to the concrete situation which gave rise to equations (3.1.1.5), we see that the commutation relations (3.1.1.4) (which follow from (3.1.1.5a,b,d) allow us to construct what can be shown to be an orthonormal eigenbasis for  $T$ , and in particular to determine its spectrum.

3.1.2. The structure revealed by the calculations above has significance far beyond its application to the determination of the spectrum of the harmonic oscillator. In particular, the commutation relations (3.1.1.4a) between  $p$  and  $q$ , or (3.1.1.5a) between  $\mathbf{a}$  and  $\mathbf{a}^+$ , which are known as Heisenberg's *Canonical Commutation Relations* (CCR for short) (cf. [Mack3, Shan, Weyl3], etc.), have been found to be fundamental to quantum mechanics. They imply the uncertainty principle, which asserts that no particle state (i.e., vector in  $L^2(\mathbf{R})$ ) can exist for which momentum and position are simultaneously well defined (i.e., which is a simultaneous eigenvector for  $p$ , the "momentum

operator," and  $q$ , the "position operator"). See [DyMc, Foll1, Körn, Shan], etc.

Further, equation (3.1.1.7) shows that a triple  $(\mathbf{a}, \mathbf{a}^+, v_0)$  consisting of two operators  $\mathbf{a}, \mathbf{a}^+$  satisfying (3.1.1.5b,c) together with a vector  $v_0$  satisfying (3.1.1.5f) is essentially unique. This may be taken as a version of another foundational result of quantum mechanics, the Stone-von Neumann Theorem (cf. Theorem 3.3.2.4 and [Cart, Foll, Mack3, Howe4, vNeu], etc.), which asserts the uniqueness, under appropriate technical hypotheses, of the canonical commutation relations. (We note that some sort of condition, such as (3.1.1.5f), is needed to supplement the CCR (3.1.1.5a) in order to guarantee uniqueness. The possibilities for nonuniqueness were exploited by J. Bernstein to obtain interesting results in distribution theory [Bern1, Bern2, Bor12].)

3.1.3. The uniqueness result of §3.1 has an easy extension to larger systems of operators. Let  $\{p_j, q_j\}_{j=1}^n$  be a collection of  $2n$  operators satisfying the following relations (known again as the Canonical Commutation Relations):

$$(3.1.3.1) \quad [p_j, p_k] = 0 = [q_j, q_k] \quad [p_j, q_k] = i \delta_{jk}.$$

Then the  $p$ 's and  $q$ 's, together with 1, the identity operator, span a  $(2n+1)$ -dimensional Lie algebra, now widely known as the *Heisenberg Lie algebra*. The Heisenberg algebra may be realized on  $L^2(\mathbf{R}^n)$  by taking  $q_j$  to be multiplication by  $ix_j$  and  $p_j$  to be partial differentiation with respect to  $x_j$ . The Stone von-Neumann Theorem applies also to these systems and asserts, again under some natural hypotheses, that the realization of the  $p$ 's and  $q$ 's by  $ix_j$ 's and  $\frac{\partial}{\partial x_j}$ 's is essentially unique. One form of this result amounts to a classification of the irreducible unitary representations (see §A.1.7) of a certain nilpotent Lie group, known as the Heisenberg group (see §3.3 and also [Cart, Foll, Howe4, Moor], etc.). This is a basic step in the classification of the unitary dual (see §A.1.7) of nilpotent and solvable Lie groups [AuKo, Kiri, Moor, Puka3].

The Heisenberg Lie algebra is closely connected not only with the harmonic oscillator, but with many other important equations of physics, both classical and quantum [Sthr, Howe6, Engl]. Extended to infinite numbers of variables, it plays a key role in quantum field theory [Sega1, Shal, Thir] and the theory of "loop groups" and vertex algebras [Garl, FrLm, FrKa, Kac1-7, KaPe, Lepo1, Lepo2].

In addition to these applications to physics, mathematical structures attached to the CCR are important in algebraic geometry (invariant theory [Howe1], abelian varieties [Cart, Igus, Mumf]), number theory (theory of  $\theta$ -series [Cart, Gelb2, Howe5, HoPS, KuMi1, 2, 3, LiVe, ToWa1, 2], etc.,  $K$ -theory [Rama]), and differential equations (Hamiltonian systems [Olve] (cf. §3.2), pseudo-differential and Fourier integral operators [FePh, Foll1, GuSt1, Howe3, 4], several complex variables [Foll2, FoSt, Stan], and  $D$ -modules

[Bor11, Bern1, Bern2]). Some of these topics will be touched on in the discussion which follows.

3.2. *The orbit method.* The philosophy which describes a large portion of the representation theory of Lie groups is a descendant of the correspondence principle of early quantum mechanics [Bohr, Iken, Jamm]. Since it is a philosophy and not a theorem, it is difficult to formulate in such a way that is not clearly false in some cases, but still appears to have content. But roughly the idea is that, if  $G$  is a connected Lie group, then for each “classical dynamical system” for  $G$ , there should be a corresponding “quantum dynamical system,” which would be a unitary representation.

3.2.1. What could this mean? The key to the matter is symplectic geometry [AbMa, Grom, GuSt, Wein]. This is geometry based on a skew-symmetric bilinear form, in contrast to Euclidean or Riemannian geometry, which is based on a symmetric bilinear form. It is a slippery, less tangible kind of geometry; there is no notion of “distance” or “angles” in symplectic geometry. However, somewhat latterly because of its elusive nature, symplectic geometry has come to be seen to be of fundamental importance. Lie theory in particular seems to be steeped in symplecticism, owing to the anti-symmetry of the Lie bracket.

Let  $V$  be a finite-dimensional real vector space. A *symplectic form*  $\langle \cdot, \cdot \rangle$  on  $V$  is a nondegenerate skew-symmetric bilinear form. Nondegeneracy means that the map  $\alpha: V \rightarrow V^*$  defined by

$$(3.2.1.1) \quad \alpha(v)(v') = \langle v', v \rangle, \quad v, v' \in V,$$

is an isomorphism. Standard elementary arguments [Lang3, Jaco2] show that for  $V$  to have a symplectic form,  $V$  must have even dimension, say  $2n$ . Further, given  $n$ , there is essentially just one symplectic form. Precisely, we can, again by very elementary arguments, always find a *symplectic basis* for  $V$ , that is, a basis  $\{e_i, f_i\}_{1 \leq i \leq n}$ , such that

$$(3.2.1.2) \quad \langle e_i, e_j \rangle = 0 = \langle f_i, f_j \rangle, \quad \langle f_i, e_j \rangle = \delta_{ij}, \quad 1 \leq i, j \leq n,$$

where  $\delta_{ij}$  is Kronecker's delta. If  $x_i, y_i$  are the coordinates with respect to the symplectic basis (we call them *symplectic coordinates*), then

$$(3.2.1.3) \quad \langle v, v' \rangle = \sum_{i=1}^n x'_i y_i - x_i y'_i.$$

From a symplectic form on  $V$ , we can construct a Lie algebra structure on  $C^\infty(V, \mathbf{R})$ , the real-valued smooth functions on  $V$ ; the Lie bracket in this case is known as the *Poisson bracket*. In formulas, in the coordinates of (3.2.1.3), we have

$$(3.2.1.4) \quad \{P, Q\} = \sum_{i=1}^n \frac{\partial P}{\partial y_i} \frac{\partial Q}{\partial x_i} - \frac{\partial P}{\partial x_i} \frac{\partial Q}{\partial y_i}, \quad P, Q \in C^\infty(V).$$



There are at least three conceptual ways of thinking about this formula. Much of the richness of Hamiltonian mechanics stems from the fact that they all yield the same answer, formula (3.2.1.4).

First, recall that the derivative or differential

$$(3.2.1.5) \quad dP = \sum_{i=1}^n \frac{\partial P}{\partial x_i} dx_i + \frac{\partial P}{\partial y_i} dy_i$$

is a function on  $V$  with values in  $V^*$ . We know the symplectic form defines an isomorphism  $\alpha$  from  $V$  to  $V^*$ . Thus we can consider  $\alpha^{-1}(dP)$  and  $\alpha^{-1}(dQ)$ , which are  $V$ -valued functions on  $V$ . We can compute

$$(3.2.1.6) \quad \alpha^{-1}(dP) = \sum_{i=1}^n \frac{\partial P}{\partial y_i} e_i - \frac{\partial P}{\partial x_i} f_i$$

and from this that the Poisson bracket may be expressed as

$$(3.2.1.7) \quad \{P, Q\} = \langle \alpha^{-1}(dP), \alpha^{-1}(dQ) \rangle.$$

Second, we can regard the  $V$ -valued function  $\alpha^{-1}(dP)$  as defining a vector field on  $V$ . (Indeed, this is the correct thing to do from the point of view of differential geometry.) We can then differentiate a function with respect to  $\alpha^{-1}(dP)$ . The Poisson bracket can also be expressed in these terms:

$$(3.2.1.8) \quad \{P, Q\} = \alpha^{-1}(dP)(Q).$$

Third, if we think of both  $\alpha^{-1}(dP)$  and  $\alpha^{-1}(dQ)$  as vector fields, then we can consider their Lie bracket, as in formula (2.3.8), and we have

$$(3.2.1.9) \quad [\alpha^{-1}(dP), \alpha^{-1}(dQ)] = \alpha^{-1}(d\{P, Q\}).$$

This formula shows that the map

$$(3.2.1.10) \quad P \rightarrow \alpha^{-1}(dP)$$

is a Lie algebra homomorphism from  $C^\infty(V; \mathbf{R})$ , equipped with the Poisson bracket, to the space of vector fields on  $V$ , with their natural Lie bracket.

This third interpretation of Poisson bracket leads one to ask what the image of the map (3.2.1.10) looks like. From the form (3.2.1.6) of  $\alpha^{-1}(dP)$  it is clear that it cannot be an arbitrary vector field; its coefficients must satisfy the obvious "integrability conditions" imposed by the equality of mixed partial derivatives, namely, if we write a vector field

$$(3.2.1.11a) \quad v = \sum a_i e_i + b_i f_i,$$

then if  $v = \alpha^{-1}(dP)$  for some  $P \in C^\infty(V; \mathbf{R})$  we must have

$$(3.2.1.11b) \quad \frac{\partial a_i}{\partial y_j} = \frac{\partial a_j}{\partial y_i}, \quad \frac{\partial a_i}{\partial x_j} = -\frac{\partial b_j}{\partial y_i}, \quad \frac{\partial b_i}{\partial x_j} = \frac{\partial b_j}{\partial x_i}.$$

Conversely, the Poincaré Lemma [Gold, Ster] tells us the conditions (3.2.1.11b) do guarantee that  $v$  will be of the form  $\alpha^{-1}(dP)$ . But of more interest are the following equivalent geometric interpretations of the integrability conditions.

**PROPOSITION 3.2.1.12.** *A vector field  $v$  is in the image of map (3.2.1.10) if and only if*

*(i) the natural action of  $v$  on  $C^\infty(V)$  is a derivation of the Poisson bracket, i.e.,*

$$(3.2.1.13) \quad v(\{P, Q\}) = \{v(P), Q\} + \{P, v(Q)\},$$

*or*

*(ii) the natural action of  $v$  on exterior forms on  $V$  annihilates the form*

$$(3.2.1.14) \quad \omega = \sum_{i=1}^n dx_i \wedge dy_i.$$

In terms of the one-parameter group  $\varphi_t$  (or local group) generated by  $v$ , as described in §2.1, condition (3.2.1.13) says that the  $\varphi_t$  will be automorphisms of the Poisson bracket, and the equivalent condition (3.2.1.14) says the  $\varphi_t$  will preserve the differential form  $\omega$ . Clearly the diffeomorphisms satisfying either of these conditions will form a group, which is sometimes called the group of *symplectomorphisms*. (A more traditional term is *canonical transformation*.) Roughly speaking, the vector fields satisfying the equivalent conditions of Proposition 3.2.1.12 form the Lie algebra of this group; consequently we will denote the space of them by  $\text{Vect}_{\text{Sp}}(V)$ . This allows us to summarize the discussion just above by saying the map (3.2.1.10) takes  $C^\infty(V; \mathbf{R})$  to  $\text{Vect}_{\text{Sp}}(V)$ .

An important technical point about the map (3.2.1.10) is that it is almost but not quite an isomorphism: it has a one-dimensional kernel, consisting of the constant functions. Also, it is easy to check from formula (3.2.1.4) (by letting  $P$  be a fixed function, and letting  $Q$  vary through the coordinate functions  $x_i, y_i$ ) that the constants are precisely the center of the Lie algebra  $C^\infty(V; \mathbf{R})$  with Poisson bracket. Thus we have an exact sequence

$$(3.2.1.15) \quad 0 \rightarrow \mathbf{R} \rightarrow C^\infty(V; \mathbf{R}) \xrightarrow{\alpha^{-1} \circ d} \text{Vect}_{\text{Sp}}(V) \rightarrow 0$$

which exhibits  $C^\infty(V; \mathbf{R})$  as a one-dimensional central extension of  $\text{Vect}_{\text{Sp}}(V)$ .

To illustrate the difference the central extension (3.2.1.15) makes, consider the Lie algebra generated by the coordinate functions  $x_i, y_i$ . It is easy to check that

$$(3.2.1.16) \quad \alpha^{-1}(dx_i) = -f_i, \quad \alpha^{-1}(dy_i) = e_i.$$

Hence the vector fields  $\alpha^{-1}(d\lambda)$ ,  $\lambda \in V^*$ , are just the directional derivatives on  $V$ ; they form an abelian Lie algebra, whose corresponding group is just  $V$ , acting on itself by translations. However, under Poisson bracket, the  $x_i$  and  $y_i$  generate a nonabelian Lie algebra: we have

$$(3.2.1.17) \quad \{x_i, x_j\} = 0 = \{y_i, y_j\}, \quad \{y_j, x_i\} = \delta_{ij}.$$

These are simply a version of the CCR (see (3.1.3.1); the normalization here is slightly different from (3.1.3.1)).

Hence the Lie algebra generated by  $V^*$  under  $\{, \}$  is a  $(2n+1)$ -dimensional, two-step nilpotent Heisenberg algebra

$$(3.2.1.18) \quad \mathfrak{h}(V) = V^* \oplus \mathbf{R}.$$

Although when realized via the Poisson bracket, the Heisenberg Lie algebra is described in terms of  $V^*$ , it is more natural to describe it in terms of  $V$ , which is easy to do since we have identified  $V$  and  $V^*$  via the map  $\alpha$  of formula (3.2.1.1). Thus we prefer to write

$$(3.2.1.19a) \quad \mathfrak{h}(V) = V \oplus \mathbf{R}.$$

Then the Lie bracket looks like

$$(3.2.1.19b) \quad [(v, t), (v', t')] = (0, \langle v, v' \rangle), \quad v, v' \in V, t, t' \in \mathbf{R}.$$

Finally, to conclude this subsection, we note that the space  $S^2(V^*)$  of homogeneous quadratic polynomials forms a Lie algebra under the Poisson bracket. This algebra normalizes the Heisenberg Lie algebra  $\mathfrak{h}(V)$  discussed just above, and via the map (3.2.1.10) it is sent isomorphically to the Lie algebra  $\mathfrak{sp}(V)$  of the symplectic group  $\mathrm{Sp}(V)$  of linear transformations of  $V$  which preserve  $\langle, \rangle$ . (See §3.5.5 for more discussion of this remarkable realization of  $\mathfrak{sp}(V)$ .)

**3.2.2.** We can use the discussion of §3.2.1 to define a *symplectic manifold*  $M$  in a manner entirely analogous to the usual definition ([Gold, Helg2, AbMa, Ster], etc.) of smooth manifold: one covers the underlying point set of the manifold  $M$  with local coordinate patches, such that the local coordinate functions are the coordinates with respect to a standard symplectic basis of a symplectic vector space; instead of letting the coordinate changes on overlapping charts be arbitrary diffeomorphisms, one requires them to be symplectomorphisms. Then if one interprets Proposition 3.2.1.12 using the standard language of differentiable manifolds (see references just above), one sees  $M$  has the following properties:

- (i) There is a distinguished closed exterior 2-form  $\omega$  on  $M$ , i.e., a section of  $\Lambda^2 T^*(M)$ , with the property that the alternating bilinear form induced by  $\omega$  on the tangent space at each point of  $M$  is a symplectic form. (The 2-form  $\omega$  will have the form (3.2.1.14) in each local chart.)
- (ii) The space  $C^\infty(M)$  is endowed with a Lie algebra structure, called the Poisson bracket, and denoted  $\{, \}$ . This will satisfy the appropriately coordinate-free versions of properties (3.2.1.7), (3.2.1.8), and (3.2.1.9). (On each coordinate patch, the bracket  $\{, \}$  will be given by formula (3.2.1.4).)

(3.2.2.1)

Alternately, one could define a symplectic manifold  $M$  as one having a distinguished closed 2-form, as in (3.2.2.1)(i), or as having a Poisson bracket structure on  $C^\infty(M; \mathbf{R})$ , as in (3.2.2.1)(ii). Some basic lemmas (Darboux's Theorem) then guarantee that  $M$  can be covered by local coordinate charts, in the way we imagined to begin with ([AbMa, Olve, Ster]).

In any case, the Poisson bracket gives us a homomorphism of Lie algebras

$$C^\infty(M; \mathbf{R}) \rightarrow \text{Vect}_{\text{Sp}}(M),$$

where again  $\text{Vect}_{\text{Sp}}(M)$  is the Lie algebra of vector fields which generate (local) one-parameter groups of symplectomorphisms. The kernel of the map is the space of locally constant functions on  $M$ . Since the characteristic functions of the connected components of  $M$  form a canonical basis for this space, we may identify it with the 0th cohomology group  $H^0(M)$ . Also, we have seen that via the map  $\alpha$  of formula (3.2.1.1), the space  $\text{Vect}_{\text{Sp}}(M)$  is identified with the closed 1-forms on  $M$ , and the map from  $C^\infty(M; \mathbf{R})$  is simply exterior differentiation. Hence the cokernel of this map is identified to the first deRham cohomology group  $H^1(M)$ . Thus we have an exact sequence

$$(3.2.2.2) \quad 0 \rightarrow H^0(M) \rightarrow C^\infty(M; \mathbf{R}) \rightarrow \text{Vect}_{\text{Sp}}(M) \rightarrow H^1(M) \rightarrow 0.$$

There are three main sources of examples of symplectic manifolds.

(a) *Cotangent bundles*: If  $M$  is any manifold, then  $T^*(M)$ , the cotangent bundle of  $M$ , is in a natural way a symplectic manifold [AbMa, Blat, Ster].

(b) *Kähler manifolds* [LaBe, Hart, Weil3]: Let  $U$  be a complex vector space, and let  $(\cdot, \cdot)$  be a Hermitian inner product on  $U$ . Then the imaginary part of  $(\cdot, \cdot)$  defines a symplectic form on the real vector space obtained from  $U$  by restricting scalars. A Kähler manifold is a complex manifold  $M$  which is endowed with a Hermitian metric on its holomorphic tangent bundle, whose imaginary part is a closed  $(1, 1)$ -form, and which thus defines a symplectic structure on  $M$ . Kähler manifolds are significant because they include all nonsingular projective algebraic varieties: complex projective space  $\mathbf{CP}^n$  possesses a Kähler metric, the Fubini-Study metric [GrHa], the unique metric invariant under the action of the unitary group  $U_{n+1}$  on  $\mathbf{CP}^n$ ; and any nonsingular projective subvariety of  $\mathbf{CP}^n$  inherits this metric by restriction. For purposes of obtaining symplectic manifolds, one can equally well consider "pseudo-Kähler" manifolds, defined in the same way as Kähler manifolds, except the Hermitian "metric" need not be positive definite.

(c) *Coadjoint orbits*: For us, this is the most important class of examples. Let  $G$  be a Lie group, write  $\text{Lie}(G) = \mathfrak{g}$ , and let  $\mathfrak{g}^*$  be the dual space to  $\mathfrak{g}$ . The group  $G$  acts on  $\mathfrak{g}$  via  $\text{Ad}$ , the adjoint action, and therefore acts on  $\mathfrak{g}^*$  via the contragredient to  $\text{Ad}$ , called the *coadjoint action*, and denoted  $\text{Ad}^*$ . Consider  $\lambda \in \mathfrak{g}^*$ . Let

$$(3.2.2.3) \quad R_\lambda = \{g \in G: \text{Ad}^* g(\lambda) = \lambda\}$$

be the stabilizer or isotropy group of  $\lambda$ , the subgroup of  $G$  which leaves  $\lambda$  fixed. Its Lie algebra is

$$(3.2.2.4) \quad \mathfrak{r}_\lambda = \{x \in \mathfrak{g} : \text{ad}^*(x)(\lambda) = 0\}.$$

The map

$$(3.2.2.5) \quad e_\lambda : g \rightarrow \text{Ad}^* g(\lambda)$$

defines a surjective,  $G$ -equivariant map from the coset space  $G/R_\lambda$  to

$$(3.2.2.6) \quad \mathcal{O}_\lambda = \{\text{Ad}^* g(\lambda) : g \in G\},$$

the  $\text{Ad}^* G$  orbit through  $\lambda$ . Differentiating the map  $e_\lambda$  at the origin gives an isomorphism

$$(3.2.2.7) \quad \mathfrak{g}/\mathfrak{r}_\lambda \simeq T(\mathcal{O}_\lambda)_\lambda$$

of the quotient  $\mathfrak{g}/\mathfrak{r}_\lambda$  with the tangent space to  $\mathcal{O}_\lambda$  at  $\lambda$ .

Consider on  $\mathfrak{g}$  the antisymmetric bilinear form

$$(3.2.2.8) \quad \langle x, y \rangle_\lambda = \lambda([x, y]).$$

One can easily check that the radical of the form  $\langle \cdot, \cdot \rangle_\lambda$ —defined as

$$\{x \in \mathfrak{g} : \langle x, y \rangle_\lambda = 0 \text{ for all } y \in \mathfrak{g}\},$$

that is, the vectors which are orthogonal to everything with respect to the form  $\langle \cdot, \cdot \rangle_\lambda$  on  $\mathfrak{g}$ —is precisely  $\mathfrak{r}_\lambda$ . Hence the form  $\langle \cdot, \cdot \rangle_\lambda$  factors to define a non-degenerate form on the quotient  $\mathfrak{g}/\mathfrak{r}_\lambda$ . In view of the isomorphism (3.2.2.7), we can push  $\langle \cdot, \cdot \rangle_\lambda$  forward to define a symplectic form on the tangent space  $T(\mathcal{O}_\lambda)_\lambda$  to  $\mathcal{O}_\lambda$  at  $\lambda$ . Since this can be done at every point of  $\mathfrak{g}^*$ , and since it is a canonical construction, this will produce a  $G$ -invariant differential 2-form which induces a symplectic form on the tangent space to  $\mathcal{O}_\lambda$  at every point. A computation shows [GuSt, AbMa] that this canonically defined 2-form is in fact closed. (It should not be surprising that this is essentially a consequence of the Jacobi identity.) Hence  $\mathcal{O}_\lambda$  is a symplectic manifold; further  $G$  acts transitively on  $\mathcal{O}_\lambda$  via symplectomorphisms.

Some coadjoint orbits are isomorphic to cotangent bundles, and others support Kähler or pseudo-Kähler metrics.

Lie [LiEn, vol. 2, p. 294] was apparently aware of the symplectic structure on coadjoint orbits, or at least the associated Poisson bracket, but it was subsequently forgotten until the 1960s when its importance for representation theory was appreciated [Bere, Blat, Kiri, Kost1].

3.2.3. Let  $G$  be a Lie group and let  $M$  be a connected symplectic manifold. Suppose  $G$  acts on  $M$  by symplectomorphisms. Differentiating the action of  $G$  yields a homomorphism  $\beta$  from  $\text{Lie}(G)$  to  $\text{Vect}_{\text{Sp}}(M)$ . Denote the image of  $\text{Lie}(G)$  in  $\text{Vect}_{\text{Sp}}(M)$  by  $\mathfrak{g}$ . We would like to lift  $\mathfrak{g}$  to a subalgebra of  $C^\infty(M)$ . According to the sequence (3.2.2.2) there are two obstructions to doing this. The first is that  $\mathfrak{g}$  may not be in the image of

the map from  $C^\infty(M)$  to  $\text{Vect}_{\text{Sp}}(M)$ , that is, some elements of  $\mathfrak{g}$  may represent nontrivial cohomology in  $H^1(M)$ . If  $M$  is simply connected, then  $H^1(M) = 0$  [Mass], so we can eliminate this obstruction by passing to a covering of  $M$  if necessary. So suppose  $\mathfrak{g}$  is in the image of  $C^\infty(M)$ . Denote the inverse image of  $\mathfrak{g}$  in  $C^\infty(M)$ , via the sequence (3.2.2.2), by  $\tilde{\mathfrak{g}}$ . Then we have a diagram:

$$\begin{array}{ccccccc} & & & \text{Lie}(G) & & & \\ & & & \beta \downarrow & & & \\ 0 & \rightarrow & \mathbf{R} & \rightarrow & \tilde{\mathfrak{g}} & \rightarrow & \mathfrak{g} \rightarrow 0 \end{array}$$

The Lie algebra  $\tilde{\mathfrak{g}}$  is a central extension of  $\mathfrak{g}$  by  $\mathbf{R}$ , and thus defines a certain cohomology class  $\gamma$  in  $H^2(\mathfrak{g}; \mathbf{R})$  (see [Jaco1, Kost1]). We can lift the homomorphism  $\beta$  to a homomorphism

$$\tilde{\beta}: \text{Lie}(G) \rightarrow \tilde{\mathfrak{g}} \subseteq C^\infty(M; \mathbf{R})$$

if and only if the pullback  $\beta^*(\gamma) \in H^2(\text{Lie}(G); \mathbf{R})$  vanishes. If this happens, then there is a choice of liftings  $\tilde{\beta}$  of  $\beta$ , corresponding to the homomorphisms of  $\text{Lie}(G)$  to  $\mathbf{R}$  (which form the group  $H^1(\text{Lie}(G); \mathbf{R}) \simeq (\mathfrak{g}/\mathfrak{g}^{(2)})^*$ ).

By a *Hamiltonian action*  $\beta$  of  $G$  on  $M$ , we mean an action of  $G$  on  $M$  by symplectomorphisms, together with a compatible homomorphism

$$\tilde{\beta}: \text{Lie}(G) \rightarrow C^\infty(M; \mathbf{R})$$

such that the diagram

$$(3.2.3.1) \quad \begin{array}{ccccccc} & & & & \text{Lie}(G) & & \\ & & & \tilde{\beta} \swarrow & \downarrow \beta & & \\ 0 & \rightarrow & \mathbf{C} & \rightarrow & C^\infty(M; \mathbf{R}) & \rightarrow & \text{Vect}_{\text{Sp}}(M) \rightarrow 0 \end{array}$$

commutes [GuSt 1, Kirw, Kost1].

REMARKS. (a) A standard basic fact about a semisimple Lie algebra  $\mathfrak{s}$  is that  $H^2(\mathfrak{s}; \mathbf{R}) = H^1(\mathfrak{s}; \mathbf{R}) = 0$  [Jaco1]. Thus if  $G$  is semisimple, then any action of  $G$  by symplectomorphisms is automatically Hamiltonian, in a unique way.

(b) For a general Lie group  $G$ , a symplectic action of  $G$  may be regarded as a Hamiltonian action of an appropriate central extension of  $G$ ; thus the action of a symplectic vector space on itself by translations comes from a Hamiltonian action of the associated Heisenberg group, as in formulas (3.2.1.16)–(3.2.1.19).

Suppose we have a Hamiltonian action  $\beta$  of  $G$  on the symplectic manifold  $M$ . By duality, the homomorphism  $\tilde{\beta}: \text{Lie}(G) \rightarrow C^\infty(M; \mathbf{R})$  gives us a mapping

$$\begin{aligned} \mu_\beta: M &\rightarrow \mathfrak{g}^*, \\ \mu_\beta(m)(x) &= \tilde{\beta}(x)(m), \quad m \in M, x \in \mathfrak{g}. \end{aligned}$$

It is easy to see that the mapping  $\mu_\beta$  is equivariant for the action of  $G$ . Because  $\mu_\beta$  describes the angular momentum of a particle in a particular

case (the action of  $0_3$  on  $\mathbf{R}^3 \times \mathbf{R}^{3*} \simeq T^*(\mathbf{R}^3)$  [AbMa, GuSt1]), it is called the *moment map*.

The geometry of the moment map for a general Hamiltonian action is quite interesting, and quite relevant for representation theory [Ati2, GuSt3, Kirw2, DuHV]. But right now we focus on the case when  $G$  acts transitively on  $M$ . In this case, the image of  $\mu_\beta$  is clearly a single coadjoint orbit. Further, an elementary argument shows that  $\mu_\beta$  must be locally a diffeomorphism. Thus any homogeneous Hamiltonian  $G$ -action must be a covering space of some coadjoint orbit [GuSt1, Kost1]. Or in other words, up to coverings, coadjoint orbits provide the universal examples of transitive Hamiltonian  $G$ -actions.

3.2.4. At the start of §3.2 we made a vague reference to the notion of a “classical dynamical system” for  $G$ . Now we can specify that we will take this to mean a Hamiltonian  $G$ -action. The rationale for this choice comes from the Hamiltonian version of classical mechanics, which shows that a classical conservative dynamical system satisfying Newton’s Laws can be expressed as a Hamiltonian action of  $\mathbf{R}$  [AbMa, Arno]; besides this it has been observed to work.

Given this meaning of “classical dynamical system,” the discussion of §3.2.3 can be taken as showing that the irreducible, i.e., transitive, classical dynamical systems for  $G$  correspond to coverings of coadjoint orbits. Thus the principle enunciated rather imprecisely at the start of §3.2 can now be stated more clearly: we hope to be able to associate irreducible unitary representations to (covers of) coadjoint orbits for  $G$ . The extent to which this hope is realized will be surveyed in the next subsections.

3.3. *Nilpotent groups*. The hope expressed in §3.2.4 is realized perfectly for nilpotent groups, as was discovered by Kirillov [Kiri, Puka1, Moor]. (Stating things this way is, in historical terms, to put the cart before the horse; Kirillov’s work was a primary inspiration for the philosophy expressed in §3.2.)

3.3.1. A key notion in Kirillov’s construction is that of *polarization*. Recall the discussion of coadjoint orbits in §3.2.2. Let  $G$  be a Lie group,  $\mathfrak{g} = \text{Lie}(G)$ ,  $\lambda \in \mathfrak{g}^*$ ,  $R_\lambda =$  the stabilizer of  $\lambda$  under  $\text{Ad}^*$ , and  $\mathfrak{r}_\lambda = \text{Lie}(R_\lambda)$ . By a *polarization* for  $\lambda$ , or *polarizing subalgebra*, or *maximal subordinate subalgebra* we mean a Lie subalgebra  $\mathfrak{p}$  of  $\mathfrak{g}$  such that

$$(3.3.1.1) \quad \mathfrak{p} \text{ is a maximal isotropic subspace for the form } \langle \ , \ \rangle_\lambda.$$

*Isotropic* means that  $\langle x, y \rangle_\lambda = 0$  for all  $x, y \in \mathfrak{p}$ . Maximal isotropic of course then means that there is no subspace of  $\mathfrak{g}$  which properly contains  $\mathfrak{p}$  and which also is isotropic for  $\langle \ , \ \rangle_\lambda$ . The duality theorems of basic linear algebra imply that if  $\mathfrak{p}$  is a polarization then

$$(3.3.1.2) \quad \begin{aligned} & \text{(i) } \mathfrak{r}_\lambda \subseteq \mathfrak{p}, \\ & \text{(ii) } \dim \mathfrak{p} = \frac{1}{2}(\dim \mathfrak{r}_\lambda + \dim \mathfrak{g}). \end{aligned}$$

Thus the single condition of (3.3.1.1) could be replaced by the two conditions: (i)  $\mathfrak{p} \supseteq \mathfrak{r}_\lambda$ , and (ii)  $\mathfrak{p}/\mathfrak{r}_\lambda$  is maximal isotropic for the symplectic form defined by  $\langle \cdot, \cdot \rangle_\lambda$  on  $\mathfrak{g}/\mathfrak{r}_\lambda$ .

Before stating Kirillov's results, we should note that for a connected, simply connected nilpotent group  $N$ , the exponential map  $\exp: \text{Lie}(N) \rightarrow N$  is a diffeomorphism [Malc, CoGr, Dixm2].

**THEOREM 3.3.1.3 (Kirillov).** *Let  $N$  be a connected and simply connected nilpotent Lie group. Set  $\text{Lie}(N) = \mathfrak{n}$ .*

(a) *There is a natural bijection between the unitary dual  $\hat{N}$  and the set  $\mathfrak{n}^*/\text{Ad}^*N$  of coadjoint orbits for  $N$ .*

(b) *Pick  $\lambda \in \mathfrak{n}^*$ . The representation  $\rho_\lambda$  corresponding to the coadjoint orbit  $\mathcal{O}_\lambda$  through  $\lambda$  may be realized as follows. Let  $\mathfrak{p} \subseteq \mathfrak{n}$  be a polarization for  $\lambda$ . (These exist.) Let  $P = \exp \mathfrak{p}$  be the connected subgroup of  $N$  with Lie algebra equal to  $\mathfrak{p}$ . It is a closed subgroup of  $N$ . Because  $\mathfrak{p}$  is isotropic for  $\langle \cdot, \cdot \rangle_\lambda$ , the formula*

$$(3.3.1.4) \quad \psi_\lambda(\exp x) = e^{2\pi i \lambda(x)}, \quad x \in \mathfrak{p},$$

*defines a unitary character of  $P$ . The unitarily induced representation (see §§A.1.14 and A.1.16)*

$$(3.3.1.5) \quad 2 - \text{ind}_P^G \psi_\lambda \approx \rho_\lambda$$

*is the representation we are looking for.*

(c) *Every element of  $\hat{N}$  is strongly trace class (see §A.1.18). For an  $\text{Ad}^*N$ -orbit  $\mathcal{O} \subseteq \mathfrak{n}^*$ , the character  $\mathcal{O}_{\rho_\mathcal{O}}$  of the corresponding representation  $\rho_\mathcal{O}$  can be computed as follows. On  $\mathcal{O}$ , there is an  $\text{Ad}^*N$ -invariant measure, unique up to multiples. Denote it by  $d_{\mathcal{O}\mu}$ . For  $f \in C_c^\infty(N)$ , let  $f \circ \exp \in C_c^\infty(\mathfrak{n})$  be the pullback to  $\mathfrak{n}$  of  $f$  via the exponential map. Define the Fourier transform from functions on  $\mathfrak{n}$  to functions on  $\mathfrak{n}^*$  in the usual way:*

$$(3.3.1.6) \quad \hat{\phi}(\lambda) = \int_{\mathfrak{n}} \phi(x) e^{-2\pi i \lambda(x)} dx, \quad \phi \in L^1(\mathfrak{n}),$$

*where  $dx$  is a Haar measure on  $\mathfrak{n}$ . Then for appropriate normalization of the invariant measure  $d_{\mathcal{O}\mu}$  on  $\mathcal{O}$ , we have*

$$(3.3.1.7) \quad \theta_{\rho_\mathcal{O}}(f) = \int_{\mathcal{O}} (f \circ \exp)^\wedge(\mu) d_{\mathcal{O}\mu}, \quad f \in C_c^\infty(N).$$

**3.3.2. REMARKS.** (a) The proof of Theorem 3.3.1.3 proceeds by induction on the dimension of  $N$ , using the tools of the “Mackey Machine” (see [FeDo, Mack4, Rief]) for computing representations of group extensions. In fact, the necessary computations are quite limited and depend mainly on understanding the Heisenberg group, the basic group of quantum mechanics, whose Lie algebra is described in formula (3.1.3.1) or (3.2.1.19).

(b) As well as being important for the proof of Theorem 3.3.1.3, the Heisenberg group provides a good illustration of it. Let  $V$  be a symplectic vector space. If we again use the isomorphism (3.2.1.1) between  $V$  and



$V^*$ , we can write

$$(3.3.2.1) \quad \mathfrak{h}(V)^* = (V \oplus \mathbf{R})^* \simeq V \oplus \mathbf{R}.$$

Using the expression (3.2.1.19b) for the Lie bracket in  $\mathfrak{h}(V)$ , and formula (2.4.8) for the adjoint action, we can compute that

$$(3.3.2.2) \quad \text{Ad}^* \exp(v, t)(v', t') = (v' + t'v, t'), \quad (v, t) \in \mathfrak{h}(V), (v', t') \in \mathfrak{h}(V)^*.$$

Denote the connected, simply connected group whose Lie algebra is  $\mathfrak{h}(V)$  by  $H(V)$ . From formula (3.3.2.2), we can easily verify the following description of  $\text{Ad}^*H(V)$  orbits.

The  $\text{Ad}^*H(V)$  orbits in  $\mathfrak{h}(V)$  are

$$(3.3.2.3) \quad \begin{aligned} & \text{(i) the points } (v, 0), \ v' \in V, \\ & \text{(ii) the hyperplanes } \{(v', t') : v' \in V\}, \ t' \in \mathbf{R} - 0. \end{aligned}$$

If we plug this data in Theorem 3.3.1.3 we obtain a complete description of the representations of  $H(V)$ . There are one-dimensional representations  $\chi_{v'}(\exp(v, t)) = e^{2\pi i \langle v, v' \rangle}$ ,  $v' \in V$ , which factor to the abelian quotient  $H(V)/ZH(V)$ . Here  $ZH(V)$  is the one-dimensional center of  $H(V)$ ; it is also the commutator subgroup. The non-one-dimensional representations correspond to the hyperplanes (3.3.2.3)(ii), and so Theorem 3.3.1.3 specializes to the following classical result (cf. [CoGr, Foll1, Howe2, Neum], etc.).

**THEOREM 3.3.2.4 (Stone-von Neumann).** *For each nontrivial character  $\chi$  of  $ZH(V)$  ( $\simeq \mathbf{R}$ ), there is up to unitary equivalence exactly one irreducible unitary representation  $\rho_\chi$  of  $H(V)$  with central character  $\chi$  (see §A.1.7.4). The representation  $\rho_\chi$  may be realized as an induced representation*

$$(3.3.2.5) \quad \rho_\chi \simeq 2 - \text{ind}_A^H \tilde{\chi},$$

where  $A \subseteq H(V)$  is any maximal abelian connected subgroup, and  $\tilde{\chi}$  is any extension of  $\chi$  from  $ZH(V)$  to  $A$ .

It is worthwhile to give a concrete description of the representations  $\rho_\chi$ , to emphasize how close we are here to the heart of classical harmonic analysis [FePh, Foll1, Howe2, 3]. For this we can first observe that for  $s \in \mathbf{R}^\times$ , the map

$$(3.3.2.6) \quad \begin{aligned} d_s \mathfrak{h}(V) &\rightarrow \mathfrak{h}(V), \\ d_s(v, t) &= (sv, s^2 t), \quad (v, t) \in \mathfrak{h}(V), \end{aligned}$$

is an automorphism of  $\mathfrak{h}(V)$ . The corresponding automorphisms of  $H(V)$  will permute almost transitively (there will be two orbits which are mutual complex conjugates) the characters of  $ZH(V)$ . Thus up to the action of the  $d_s$  and complex conjugation, there is only one (infinite dimensional unitary irreducible) representation of  $H(V)$ . So we only need to describe one such representation. But this is in fact given by the realization of  $\mathfrak{h}(\mathbf{R}^n \oplus (\mathbf{R}^n)^*)$

via the operators  $\frac{\partial}{\partial x_j}$  and  $ix_j$  on  $L^2(\mathbf{R}^n)$ , as described in §3.1.3. Thus, via this representation, the universal enveloping algebra  $\mathcal{U}(\mathfrak{h}(\mathbf{R}^n \oplus \mathbf{R}^{n*}))$  will be sent to the algebra of polynomial-coefficient differential operators on  $\mathbf{R}^n$ . (See §A.1.13 for an explanation of how to derive a representation of the enveloping algebra.)

**3.4. Solvable groups.** Kirillov's results appeared in 1960 [Kiri]. Through a lot of hard work since then, the basic principles embodied in Kirillov's theory have been extended to encompass a large portion of representation theory of Lie groups. The next class of groups to be analyzed was solvable groups. We briefly outline this development.

Inspection of Theorem 3.3.1.3 makes clear that the bijection of part (a) between orbits and representations is implemented in two quite distinct ways: first, by an explicit construction of the representations, and second by a description of the character of a representation in terms of the orbit. It might seem that a construction of the representation is very much to be preferred to just a description of the character. However, it should be noted that the construction of the representation involves a noncanonical intermediate construction between the orbit and the representation, namely a polarization. While it is always possible to find a polarization, there may in fact be many, and the choice of a particular one is arbitrary. However, one shows that all the representations one constructs by means of various polarizations are equivalent. (A key fact used to do this is the Stone-von Neumann Theorem). This "independence of polarization" allows the construction to succeed. On the other hand, the description of the character via formula (3.3.1.7) is canonical. There is even an a priori description of the proper normalization of the measure  $d_{\theta}\mu$  [Moor, Puka2]. Below we will discuss the generalizations of both parts (b) and (c) to other classes of Lie groups.

**3.4.1. For solvable Lie groups,** the situation is more complicated, but quite satisfactory. Kirillov's work was generalized almost immediately [Bert] to the class known as *exponential solvable groups*, which are characterized as those solvable groups  $G$  whose simply-connected cover  $\tilde{G}$  is such that the exponential map  $\exp: \text{Lie}(\tilde{G}) \rightarrow \tilde{G}$  is a diffeomorphism [Moor, Dixm3]. For exponential solvable groups the bijection between orbits and representations holds, and can be realized using induced representations by an explicit construction using polarizations, just as in the nilpotent case. However, two difficulties arise:

- (i) Not all polarizations yield the same representation, or even an irreducible representation;
- (ii) Not all representations are strongly trace class.

**3.4.1.1. Both difficulties** are already illustrated by the two-dimensional " $ax + b$  group"—the group of affine transformations of the line. This may

be realized as the set of  $2 \times 2$  matrices of the form

$$(3.4.1.1.1) \quad G = \left\{ \begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} : b \in \mathbf{R}, a \in \mathbf{R}^{+\times} \right\}.$$

(We restrict  $a$  to be positive in order to have a connected group.) The Lie algebra of this group is the space of matrices

$$(3.4.1.1.2) \quad \begin{bmatrix} \alpha & \beta \\ 0 & 0 \end{bmatrix}, \quad \alpha, \beta \in \mathbf{R},$$

and its dual may be realized as the space

$$(3.4.1.1.3) \quad \begin{bmatrix} \lambda & 0 \\ \mu & 0 \end{bmatrix}, \quad \lambda, \mu \in \mathbf{R}.$$

The pairing between the matrices (3.4.1.1.2) and (3.4.1.1.3) is given by taking the trace of products. The coadjoint orbits are

$$(3.4.1.1.4) \quad \mathcal{O}_\lambda = \left\{ \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix} \right\} \quad \text{for each } \lambda \in \mathbf{R},$$

$$\mathcal{O}^+ = \left\{ \begin{bmatrix} \lambda & 0 \\ \mu & 0 \end{bmatrix} : \lambda \in \mathbf{R}, \mu > 0 \right\}, \quad \mathcal{O}^- = \left\{ \begin{bmatrix} \lambda & 0 \\ \mu & 0 \end{bmatrix} : \lambda \in \mathbf{R}, \mu < 0 \right\}.$$

The representations corresponding to the one-point orbits are the linear characters (one-dimensional representations) of the group. These are trivial on the commutator subgroup  $G^1$ ,

$$(3.4.1.1.5) \quad G^1 = \left\{ \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} : b \in \mathbf{R} \right\}.$$

There are two non-one-dimensional irreducible representations, corresponding to the orbits  $\mathcal{O}^+$  and  $\mathcal{O}^-$ . The Lie algebra of the group  $G^1$  is a polarization for any element in either of these orbits, and we have

$$(3.4.1.1.6) \quad \rho_{\mathcal{O}^\pm} \simeq 2 - \text{ind}_{G^1}^G \chi_\mu, \quad \begin{bmatrix} 0 \\ \mu \end{bmatrix} \in \mathcal{O}^\pm,$$

where  $\chi_\mu \left( \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \right) = e^{2\pi i \mu b}$ . However, the group

$$(3.4.1.1.7) \quad A = \left\{ \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix} : a \in \mathbf{R}^{+\times} \right\}$$

also defines a polarization of any element of  $\mathcal{O}^\pm$ . If  $\chi$  is any character of  $A$ , then the unitary representation of  $G$  induced from  $\chi$  is equivalent to the sum  $\rho_{\mathcal{O}^+} \oplus \rho_{\mathcal{O}^-}$ .

The representations  $\rho_{\mathcal{O}^\pm}$  are also not strongly trace class. In fact, if  $f \in C_c^\infty(G)$  is such that  $\chi_\lambda(f) \neq 0$  for some  $\lambda$ , then  $\rho_{\mathcal{O}^\pm}(f)$  will not be trace class. Here  $\chi_\lambda$  indicates the character of  $G$  corresponding to the orbit  $\mathcal{O}_\lambda$  of (3.4.1.1.4). Precisely

$$\chi_\lambda \left( \begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} \right) = a^{2\pi i \lambda}.$$

3.4.1.2. Despite these two complications, the situation for exponential solvable groups is quite well understood. There is a simple criterion first formulated by Pukanszky [Moor, Puka5] to guarantee that a polarization will produce the appropriate irreducible representation. Further there is a clean description of the representation produced by any polarization [Moor, Verg1].

With regard to generalizing the trace formula (3.3.1.7), one must recognize that it will not generalize completely because not all representations are strongly trace class. Roughly speaking, it will be closed orbits which correspond to strongly trace class representations. (Observe that the orbits  $\mathcal{O}^\pm$  of (3.4.1.1.4) are not closed.) Even for orbits for which there is a trace formula, a new phenomenon enters: it is necessary to multiply a function by an appropriate normalizing factor, which depends on the orbit. Thus for an orbit  $\mathcal{O}$  for which it is valid, the trace formula takes the form

$$(3.4.1.2.1) \quad \theta_{\rho_{\mathcal{O}}}(f) = \int_{\mathcal{O}} ((f \circ \exp)(L_{\mathcal{O}}))^{\wedge}(\mu) d_{\mathcal{O}}\mu, \quad f \in C_c^\infty(G),$$

where  $L_{\mathcal{O}}$  is an analytic function on  $\text{Lie}(G)$  [Moor, Dufl1, Puka4]. The need to introduce  $L_{\mathcal{O}}$  stems from two sources:

- (i) For a general exponential solvable group  $G$ , the exponential map

$$\exp: \text{Lie}(G) \rightarrow G$$

will not take Haar measure on  $\text{Lie}(G)$  to Haar measure on  $G$ .

- (ii) The modular function of (the connected subgroup of  $G$  whose Lie algebra is) a polarization of  $\lambda \in \mathcal{O}$  may not agree with the restriction of the modular function of  $G$ .

3.4.2. For general solvable groups, one encounters several difficulties which did not arise in the exponential solvable case.

3.4.2.1. (i) The representations of solvable groups can be badly behaved: these groups need not be type I in the sense of  $C^*$ -algebras [Dixm1].

(ii) Not all representations are realizable as monomial representations, i.e., as induced representations from linear characters—in terms of the orbit method, this means there are elements  $\lambda \in \text{Lie}(G)^*$  for which there is no polarizing subalgebra in  $\text{Lie}(G)$ . An example is provided below.

(iii) Orbits  $\mathcal{O} \subseteq \text{Lie}(G)^*$  may not be simply connected—equivalently, their isotropy groups in  $\tilde{G}$ , the simply-connected cover of  $G$ , may not be connected (consider  $E_2$ , the isometry group of the Euclidean plane). Also, orbit structure may be highly irregular—orbits may not even be locally closed. (A semidirect product  $\mathbf{R} \ltimes \mathbf{R}^4$ , where  $\mathbf{R}$  acts on  $\mathbf{R}^4$  by a sum of mutually irrational rotations, provides the simplest example. It was first noted by Mautner.)

These phenomena force a substantial revision in the orbit method, and a naive one-to-one correspondence between coadjoint orbits and representations no longer exists. However, there still exists a highly satisfactory, detailed

theory which retains much of the flavor of Theorem 3.3.1.3 [AuKo, Moor, Puka3, Puka6]. We will describe how the new features of this theory solve the problems 3.4.2.1.

First one must lump coadjoint orbits into equivalence classes of “quasiorbits.” Two orbits define the same quasiorbit if their closures in  $\text{Lie}(G)^*$  are equal. It turns out that a quasiorbit in  $\text{Lie}(G)^*$  is an orbit for a slightly larger group  $G' \supseteq G$  such that the quotient  $G'/G$  is abelian [Puka3].

Second, one must seek to parametrize, not representations, but primitive ideals in  $C^*(G)$  (see [Dixm2]). Here  $C^*(G)$  is the group  $C^*$ -algebra (cf. §A.1.12, especially definition (A.1.12.6)). Recall that if the group  $G$  is type I, then there is a natural bijection between equivalence classes of irreducible unitary representations and primitive ideals in  $C^*(G)$  [Dixm2]. But if  $G$  is not type I, there may be infinitely many irreducible representations whose kernel in  $C^*(G)$  is a given primitive ideal. (This is in fact fairly typical behavior. The easiest examples may be induced representations of the rank 3 integral Heisenberg group.)

Third, the mapping to quasiorbits from primitive ideals is many-to-one. The fibers are quotients of the duals of subgroups of the component groups of the isotropy groups [Puka6].

Let us state the result precisely. Let  $G$  be a connected and simply connected solvable Lie group, with Lie algebra  $\text{Lie}(G)$ . Consider  $\lambda$  in  $\text{Lie}(G)^*$ . Let  $R_\lambda \subseteq G$  be the stabilizer of  $\lambda$  under the coadjoint action  $\text{Ad}^*G$ , and let  $R_\lambda^0$  be the identity component of  $R_\lambda$ . Recall that  $\mathfrak{r}_\lambda = \text{Lie}(R_\lambda^0)$  is the radical of the form  $\langle \cdot, \cdot \rangle_\lambda$  associated to  $\lambda$  (cf. (3.2.2.4)). There is a unique character  $\chi_\lambda$  on  $R_\lambda^0$  defined by

$$(3.4.2.2) \quad \chi_\lambda(\exp r) = e^{2\pi i \lambda(r)}, \quad r \in \mathfrak{r}_\lambda.$$

It is easy to see that the component group  $R_\lambda/R_\lambda^0$  is abelian. Also, since  $R_\lambda$  stabilizes  $\lambda$ , the character  $\chi_\lambda$  on  $R_\lambda^0$  is clearly invariant under conjugation by all of  $R_\lambda$ . Hence the quotient group  $R_\lambda/\ker \chi_\lambda$  is a central extension of  $R_\lambda/R_\lambda^0$  by the group  $R_\lambda^0/\ker \lambda$ , which we may identify to the unit circle  $\mathbf{T}$  by means of  $\chi$ . Thus  $R_\lambda/\ker \chi_\lambda$  is a two-step nilpotent group (or possibly abelian—the extension may split), and we have an exact sequence

$$(3.4.2.3) \quad 1 \rightarrow \mathbf{T} \rightarrow R_\lambda/\ker \chi_\lambda \rightarrow R_\lambda/R_\lambda^0 \rightarrow 1.$$

Let  $S_\lambda$  be the image in  $R_\lambda/R_\lambda^0$  of the center of  $R_\lambda/\ker \chi_\lambda$ .

**THEOREM 3.4.2.4 [Puka6].** *Let  $G$  be a connected, simply connected solvable Lie group. Let  $\mathbf{P}(C^*(G)) = \mathbf{P}(G)$  denote the space of primitive ideals of  $C^*(G)$ . Let  $(\text{Lie}(G)^*/\text{Ad}^*G)^\sim$  be the space of coadjoint quasiorbits. There is*

a mapping

$$(3.4.2.5) \quad \begin{array}{c} \mathbf{P}(G) \\ \kappa \downarrow \\ (\mathrm{Lie}(G)^* / \mathrm{Ad}^* G)^\sim \end{array}$$

such that the fiber  $\kappa^{-1}(\mathcal{O})$  above a quasisorbit  $\tilde{\mathcal{O}} \subseteq \mathrm{Lie}(G^*)$  can be identified with a quotient of  $\hat{S}_\lambda$  for any  $\lambda \in \tilde{\mathcal{O}}$ .

The most subtle aspect of this result is to understand which quotient of  $\hat{S}_\lambda$  gives the fiber. This is closely related to the group  $G'$  mentioned above for which the  $G$ -quasisorbit becomes an ordinary orbit. If the quasisorbit consists of a single  $G$ -orbit, then the fiber is all of  $\hat{S}_\lambda$ .

We can also use the notions just formulated to give the criterion of Auslander-Kostant that a solvable group be type I.

**THEOREM 3.4.2.6 [AuKo].** *The group  $G$  is type I if and only if*

- (i) *all coadjoint quasisorbits consist of a single coadjoint orbit, equivalently, the coadjoint orbits are locally closed, and*
- (ii) *for every  $\lambda$ ,  $S_\lambda = R_\lambda / R_\lambda^0$ , i.e., the extension (3.4.2.3) is trivial.*

3.4.3. The correspondence (3.4.2.5) is again described in the two ways indicated by Theorem 3.3.1.3—by explicit construction of induced representations, and by character formulas. However, both these constructions must be more sophisticated. The character formula is similar to the formula (3.4.1.2.1) for exponential groups, except one must restrict the functions  $f$  to have support in a certain neighborhood of the identity, and the formula does not distinguish between different elements in the fibers of the map  $\kappa$  of (3.4.2.5) [Puka3]. By considering integrals over quasisorbits rather than orbits, Pukanszky [Puka3] has formulated an extension of the character formula to the non-type I case.

Although polarizations no longer exist for an arbitrary  $\lambda \in \mathrm{Lie}(G)^*$ , there is still a fairly direct construction of the representation associated to an orbit as a representation induced from a special class of representations of subgroups. Here again the Heisenberg group, and somewhat more general two-step nilpotent groups, play a key role.

One can preserve the geometric flavor that polarizations give to the constructions by considering *complex polarizations*. In essence, a complex polarization is a complex Lie subalgebra of  $\mathrm{Lie}(G)_\mathbb{C}$ , the complexification of  $\mathrm{Lie}(G)$ , which satisfies condition (3.3.1.1), where  $\lambda$  now means the complex-linear extension of  $\lambda \in \mathrm{Lie}(G)^*$  to  $\mathrm{Lie}(G)_\mathbb{C}$ . In order for a complex polarization in the above sense to be usable for constructing representations, it should also satisfy some other technical conditions [Moor, p. 21; AuKo], which are usually incorporated into the definition of complex polarization. One can show that complex polarizations always exist. Indeed, Auslander-Kostant

establish the existence of complex polarizations satisfying an additional condition called *positivity*. The existence of positive complex polarizations is, once again, essentially a phenomenon associated with the Heisenberg group [AuKo, Moor].

Having a positive complex polarization for  $\lambda \in \text{Lie}(G)^*$  allows one to construct the representation associated to the coadjoint orbit through  $\lambda$  on a space of partially holomorphic sections of a complex line bundle. The basic example is the "Fock model" (cf. [Barg, Foll, Howe2, Segal], etc.), for the representations of the Heisenberg group. More recently, several authors [Carm, MoVe, Penn, Rose] have considered using nonpositive complex polarizations. This leads to the realization of representations on spaces of higher cohomology of the associated line bundles, rather than sections (= degree zero cohomology). Although these constructions using higher cohomology are not necessary to construct the representations of our solvable  $G$ , they establish a parallel between solvable groups and semisimple groups, for which realizations on cohomology are necessary (see §§3.5.5, 3.6.3, 3.6.5).

3.4.4. To conclude our discussion of solvable groups, we will give the basic example showing that polarizations may not exist for all  $\lambda \in \text{Lie}(G)^*$ , and, correspondingly, that representations of  $G$  may not be monomial (i.e., induced from one-dimensional representations of subgroups). The reason not all representations of solvable groups are monomial is related to the age-old fact that not all real matrices are diagonalizable, or even triangularizable, over the real numbers. The four-dimensional Lie algebra described in formula (3.1.1.4) typifies the problem. It may be realized as a Lie algebra of  $4 \times 4$  matrices:

$$(3.4.4.1) \quad \mathfrak{g} = \left\{ \begin{bmatrix} 0 & x & y & 2z \\ 0 & 0 & -t & y \\ 0 & t & 0 & -x \\ 0 & 0 & 0 & 0 \end{bmatrix} : t, x, y, z \in \mathbf{R} \right\}.$$

The three-dimensional subalgebra of elements of  $\mathfrak{g}$  with  $t = 0$  is a Heisenberg Lie algebra. Denote it by  $\mathfrak{h}$ . The center of  $\mathfrak{h}$  consists of the elements of  $\mathfrak{h}$  with  $x = 0 = y$ . Denote it by  $z(\mathfrak{h})$ . Then  $\mathfrak{h}/z(\mathfrak{h})$  is abelian, and it is easily seen that the adjoint action of  $\mathfrak{g}/\mathfrak{h}$  on  $\mathfrak{h}/z(\mathfrak{h})$  is irreducible (over  $\mathbf{R}$ —when complexified it will of course break up into a sum of two eigenlines). Consider any  $\lambda$  in  $\mathfrak{g}^*$  whose restriction to  $z(\mathfrak{h})$  is nonzero. Simple computations show that the coadjoint orbit  $\mathcal{O}_\lambda = \mathcal{O}$  through  $\lambda$  is two-dimensional. Thus the isotropy subalgebra  $\mathfrak{r}_\lambda$  of  $\lambda$  is also two-dimensional, and any polarization of  $\lambda$  must have dimension 3. However, the projection of  $\mathcal{O}_\lambda$  into  $\mathfrak{h}^*$  is also two-dimensional (see (3.3.2.3)), hence  $\dim(\mathfrak{r}_\lambda \cap \mathfrak{h}) = 1$ , so  $\mathfrak{r}_\lambda$  projects onto  $\mathfrak{g}/\mathfrak{h}$ . Since  $\mathfrak{g}/\mathfrak{h}$  acts irreducibly on  $\mathfrak{h}/z(\mathfrak{h})$ , there are no three-dimensional subalgebras of  $\mathfrak{g}$  containing  $\mathfrak{r}_\lambda$ . So  $\lambda$  has no polarizations.

On the other hand, the calculations of §3.1.1 produce a representation of  $\mathfrak{g}$ , acting on the same space as the canonical representation of  $\mathfrak{h}$ , de-

scribed by the Stone-von Neumann Theorem (Theorem 3.3.2.4). Using this extension of representations from  $\mathfrak{h}$  to  $\mathfrak{g}$ , one can verify a one-to-one correspondence between orbits and representations for the simply-connected group  $\tilde{G}$  associated to  $\mathfrak{g}$ . Similar, somewhat more general, constructions involving Heisenberg-like groups suffice to construct factor representations corresponding to arbitrary primitive ideals of  $C^*(G)$  for general solvable groups  $G$ .

**3.5. Compact groups.** The representation theory of compact Lie groups is equivalent, via the process of differentiating a representation (see §A.1.13), to the representation theory of complex semisimple (actually, reductive) Lie algebras. The bare essence of this is Cartan's theory of the highest weight, and is a key chapter in his foundational work on Lie theory [Crtn2]. (For an interesting account of some history of this, see [Hawk1].) Weyl [Weyl1, PeWe], provided the analytic apparatus to make the connection between the two theories, and provided important supplements (complete reducibility, character formula). Harish-Chandra [HaCh8] made a connection with the orbit method by providing an orbital interpretation of the Weyl character formula. It is interesting that this work, which is a key to Harish-Chandra's later construction of the discrete series for noncompact semisimple groups, precedes Kirillov's [Kiri] by several years. The other aspect of the orbit method, construction of representations via polarizations, is provided by the Borel-Weil-Bott Theorem [Bott, Warn, Voga 1], which is also a development of the 1950s. It too provided important guidance to the noncompact case. In the sections below, we will review these developments more closely.

**3.5.1.** To start, let us review the representations of  $\mathfrak{sl}_2$ , the unique simple Lie algebra over  $\mathbb{C}$  of minimal dimension, namely three. This is a simple and attractive topic, with numerous applications, both within Lie theory proper (cf. §2.8) and in many other parts of mathematics (cf. [Lang1, HoTa, Howe1, §4(b); Proc], etc.) and physics (cf. [BiLo1, 2, Hame, Jone, Shan], etc.).

Recall (see formulas (2.8.1)) that  $\mathfrak{sl}_2$  has a basis  $h, e^+, e^-$ , satisfying the commutation relations

$$(3.5.1.1) \quad [h, e^\pm] = \pm 2e^\pm, \quad [e^+, e^-] = h.$$

**REMARK.** We note that the compact group whose complexified Lie algebra is  $\mathfrak{sl}_2$  is  $SU_2$ , the special unitary group in two variables. A basis for  $\mathfrak{su}_2$  is provided by the famous *Pauli spin matrices* [Shan]

$$\sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

The basis  $h, e^\pm$  of  $\mathfrak{sl}_2$  is expressed in terms of the spin matrices as follows:

$$h = \sigma_z, \quad e^+ = \frac{1}{2}(\sigma_x + i\sigma_y), \quad e^- = \frac{1}{2}(\sigma_x - i\sigma_y).$$

The passage from  $SU_2$  to  $\mathfrak{su}_2$  to  $\mathfrak{sl}_2$  is typical of the flexibility permitted by Lie theory.



The first basic fact about representations of  $\mathfrak{sl}_2$  stems directly from the first of these relations. Suppose we have a triple of operators  $h, e^\pm$  on a vector space  $V$ , and suppose  $h, e^\pm$  satisfy the commutation relations (3.5.1.1). Suppose  $v$  is an eigenvector for  $h$ , with eigenvalue  $\lambda$ :

$$(3.5.1.2) \quad h(v) = \lambda v.$$

We compute

$$(3.5.1.3) \quad \begin{aligned} h(e^+(v)) &= ([h, e^+] + e^+h)(v) = 2e^+(v) + e^+(\lambda v) \\ &= (\lambda + 2)e^+(v). \end{aligned}$$

Thus  $e^+(v)$  is again an eigenvector for  $h$ , with eigenvalue  $\lambda + 2$ , the eigenvalue of  $v$  plus two. A similar computation shows  $e^-(v)$  is also an  $h$ -eigenvector, with eigenvalue  $\lambda - 2$ . Thus the effect of  $e^+$  is to shift the eigenspaces of  $h$  to higher eigenvalues;  $e^-$  shifts the  $h$ -eigenspaces toward lower eigenvalues. This phenomenon is commonly described by calling  $e^+$  a *raising operator* and  $e^-$  a *lowering operator*. We may summarize the above computation as follows.

**LEMMA 3.5.1.4.** *If  $V$  is a module for  $\mathfrak{sl}_2$ , and  $V_\lambda \subseteq V$  is the  $\lambda$ -eigenspace for  $h$ , then the sum  $\sum_{k \in \mathbb{Z}} V_{\lambda+2k}$  of  $h$ -eigenspaces is invariant under  $\mathfrak{sl}_2$ . More precisely, we have*

$$e^\pm(V_{\lambda+2k}) \subseteq V_{\lambda+2(k \pm 1)}.$$

The above discussion shows that the product  $e^-e^+$  preserves  $h$ -eigenspaces. For a sharper understanding of the structure of representations of  $\mathfrak{sl}_2$ , we investigate the structure of the operator  $e^-e^+$  (or  $e^+e^-$ ; but  $e^+e^- = e^-e^+ + h$ ).

To analyze  $e^-e^+$  we consider the *Casimir operator*

$$(3.5.1.5) \quad \mathcal{C} = h^2 + 2(e^+e^- + e^-e^+) = h^2 + 2h + 4e^-e^+ = h^2 - 2h + 4e^+e^-.$$

A straightforward computation shows that  $\mathcal{C}$  commutes with all of  $\mathfrak{sl}_2$ . Thus  $\mathcal{C}$  is in the center of the universal enveloping algebra of  $\mathfrak{sl}_2$ . In fact, it generates the center (cf. [Lang1, Hump], etc.).

Since  $\mathcal{C}$  commutes with  $\mathfrak{sl}_2$ , its eigenspaces will be invariant under  $\mathfrak{sl}_2$ . If  $V$  consists of a single eigenspace for  $\mathcal{C}$ , we will say the action of  $\mathfrak{sl}_2$  on  $V$  is *quasisimple*. Clearly all finite-dimensional irreducible representations are quasisimple by Schur's Lemma (cf. [HeRo, Lang3, Jaco2, Knap2] etc.).

Suppose the action of  $\mathfrak{sl}_2$  on  $V$  is quasisimple, so that  $\mathcal{C}$  acts on  $V$  by a scalar, which we will denote by  $\mu$ . Again let  $V_\lambda$  be the  $\lambda$ -eigenspace for  $h$ . Then if  $v \in V_\lambda$ , equation (3.5.1.5) says

$$(3.5.1.6) \quad e^-e^+(v) = \frac{1}{4}(\mu - \lambda^2 - 2\lambda)v, \quad e^+e^-(v) = \frac{1}{4}(\mu - \lambda^2 + 2\lambda)(v).$$

Thus, if  $V$  is quasisimple, the operator  $e^-e^+$  acts as a scalar on each  $V_\lambda$ , and this scalar will be nonzero unless we have the quadratic relation

$$(3.5.1.7a) \quad \mu + 1 = (\lambda + 1)^2.$$

Similarly,  $e^+e^-$  acts as a scalar, which is nonzero unless

$$(3.5.1.7b) \quad \mu + 1 = (\lambda - 1)^2.$$

Note that (3.5.1.7a) becomes (3.1.5.7b) under the translation  $\lambda \rightarrow \lambda + 2$ . These equations imply that in a sum  $\sum_{k \in \mathbb{Z}} V_{\lambda+2k}$ , there are at most two values of  $k$  for which either of the maps

$$e^+ : V_{\lambda+2k} \rightarrow V_{\lambda+2k+2}, \quad e^- : V_{\lambda+2k+2} \rightarrow V_{\lambda+2k}$$

fails to be an isomorphism.

Now suppose  $V$  is irreducible and finite dimensional. Then necessarily  $V = \sum_{k \in \mathbb{Z}} V_{\lambda+2k}$  for some fixed  $\lambda$ , and clearly  $V_{\lambda+2k} = \{0\}$  for  $k$  large enough. By replacing  $\lambda$  by  $\lambda + 2k$  for appropriate  $k$ , we can assume  $V_\lambda \neq \{0\}$  but  $V_{\lambda+2k} = \{0\}$  for  $k > 0$ . Choose  $v_0 \in V_\lambda$ , and set  $v_j = (e^-)^j v_0$ . Then we have the formulas

$$(3.5.1.8) \quad e^-(v_j) = v_{j+1}, \quad e^+v_j = j(\lambda - j + 1)v_{j-1}.$$

The first formula amounts to the definition of  $v_j$ , and the second follows from formulas (3.5.1.6). Since  $V$  is finite dimensional, we must have  $e^-(v_j) = 0$  for some  $j$ . Then also  $e^+e^-(v_j) = e^+(v_{j+1}) = 0$ . From the second of formulas (3.5.1.8), we see that necessarily  $j = \lambda$ . Hence  $\lambda$  must be a nonnegative integer. We then further see that formulas (3.5.1.8) define a unique  $\mathfrak{sl}_2$ -module structure on the span of the  $v_j$ ,  $0 \leq j \leq \lambda$ . This span must thus be all of  $V$ . We conclude  $\dim V = \lambda + 1$ , and that the  $v_j$ 's are a basis for  $V$ . The following result summarizes our analysis.

**PROPOSITION 3.5.1.9.** *For each positive integer  $n$ , there is up to isomorphism a unique irreducible representation of  $\mathfrak{sl}_2$  of dimension  $n$ . The space of this representation allows a basis  $\{v_j : 0 \leq j \leq n-1\}$  with respect to which the action of  $\mathfrak{sl}_2$  is described by (3.5.1.8), with  $\lambda = n-1$ . In particular, the eigenvalues of  $h$  are*

$$\{m : -(n-1) \leq m \leq n-1, m \equiv n-1 \pmod{2}\}$$

*and these eigenvalues all have multiplicity one.*

**REMARK.** The above arguments can easily be adapted to describe all irreducible representations, finite- or infinite-dimensional, of the group  $\mathrm{SL}_2(\mathbb{R})$  (cf. [Barg1, Lang1, HoTa], etc.).

**3.5.2.** The very precise picture presented by Proposition 3.5.1.9 has an analog for a general (semi-) simple complex Lie algebra  $\mathfrak{g}$ . The basic results are due to Cartan [Crtn2] but understanding of the structure behind them has been refined considerably since 1913. We will give a fairly modern account, based roughly on [HaCh1, Jaco1, Hump, BGG1-3].

To set the mood for this construction, consider the following description of the finite-dimensional representations of  $\mathfrak{sl}_2$ . If  $V$  is an  $\mathfrak{sl}_2$ -module and

$v \in V$ , call  $v$  a *highest weight vector* if  $e^+(v) = 0$  and  $h(v) = \lambda v$  for some number  $\lambda$ . The eigenvalue of  $\lambda$  is then called a *highest weight*. The vector  $v_0$  of the basis  $v_j$  in (3.5.1.8) is a highest weight vector with highest weight  $\lambda$ . An  $\mathfrak{sl}_2$ -module which is generated by a highest weight vector is called a *highest weight module*. It is easy to see that if we have a vector space  $V(\lambda)$  with basis  $\{v_j : 0 \leq j < \infty\}$ , and we define an action of  $\mathfrak{sl}_2$  on  $V(\lambda)$  by formulas (3.5.1.8), then we obtain a highest weight module, with  $v_0$  a highest weight vector of weight  $\lambda$ . Further, an easy argument, again based on formulas (3.5.1.6), shows  $V(\lambda)$  is the universal highest weight module with highest weight  $\lambda$  in the sense that if  $U(\lambda)$  is any highest weight with highest weight  $\lambda$ , there is a surjective  $\mathfrak{sl}_2$ -module morphism from  $V(\lambda)$  to  $U(\lambda)$ .

We can do this for any number  $\lambda$ . Typically  $V(\lambda)$  is irreducible. However, if  $\lambda$  is a nonnegative integer, then the quantity  $j(\lambda - j + 1)$  will be zero not only for  $j = 0$ , but also when  $j = \lambda + 1$ . In this case,  $v_{\lambda+1}$  will be a highest weight vector, with highest weight  $-\lambda - 2$ . Thus, when  $\lambda$  is a nonnegative integer, the module  $V(-\lambda - 2)$  is a submodule of  $V(\lambda)$ . One sees that the quotient  $V(\lambda)/V(-\lambda - 2)$  is the finite-dimensional irreducible representation of dimension  $\lambda + 1$ . Another way of saying this is to observe that we have an exact sequence

$$(3.5.2.1) \quad 0 \rightarrow V(-\lambda - 2) \rightarrow V(\lambda) \rightarrow F(\lambda) \rightarrow 0,$$

where  $F(\lambda)$  is the finite-dimensional irreducible representation with highest weight  $\lambda$ .

3.5.3. The description (3.5.2.1) of the finite-dimensional irreducible representation of  $\mathfrak{sl}_2$  has a generalization to all complex semisimple (finite-dimensional) Lie algebras [Jaco1, Hump, BGG1–3]. We will describe it. Let  $\mathfrak{g}$  be a complex semisimple Lie algebra, and let  $\mathfrak{a} \subseteq \mathfrak{g}$  be a Cartan subalgebra. Consider the decomposition (2.8.6) of  $\mathfrak{g}$  into root spaces for  $\mathfrak{a}$ :

$$\mathfrak{g} = \mathfrak{a} \oplus \sum_{\alpha \in \Sigma} \mathfrak{g}_{\alpha}.$$

Here  $\Sigma$  denotes the set of roots of  $\mathfrak{a}$  acting on  $\mathfrak{g}$ . As we remarked in §2.8, if  $\mathfrak{g}_{\alpha}$  is a root space, then so is  $\mathfrak{g}_{-\alpha}$ , and  $\mathfrak{g}_{\alpha}$  and  $\mathfrak{g}_{-\alpha}$  together generate an algebra  $\mathfrak{s}_{\alpha}$  isomorphic to  $\mathfrak{sl}_2$ . Let  $h_{\alpha}$  in  $\mathfrak{s}_{\alpha} \cap \mathfrak{a}$  be the element corresponding to the element  $h$  as in formulas (2.8.1). In other words, the element  $h_{\alpha}$  is determined by the conditions

$$(3.5.3.1) \quad h_{\alpha} \in \mathfrak{a} \cap \mathfrak{s}_{\alpha}, \quad \alpha(h_{\alpha}) = 2.$$

The element  $h_{\alpha}$  is frequently called a *coroot*.

It follows from the description in Proposition 3.5.1.9 of the representations of  $\mathfrak{sl}_2$  that  $\beta(h_{\alpha}) \in \mathbb{Z}$  for all roots  $\beta$ . Thus if we denote by  $\mathfrak{a}_{\mathbb{R}}$  the real span of the  $h_{\alpha}$  for all roots  $\alpha$ , we see that the roots take real values on  $\mathfrak{a}_{\mathbb{R}}$ .

Choose any  $h_0 \in \mathfrak{a}_{\mathbf{R}}$  such that  $\alpha(h_0) \neq 0$  for all  $\alpha \in \Sigma$ . Set

$$(3.5.3.2) \quad \Sigma^+ = \{\alpha \in \Sigma : \alpha(h_0) > 0\}, \quad \Sigma^- = -\Sigma^+ = \{\alpha : \alpha(h_0) < 0\}.$$

The sets  $\Sigma^+$  and  $\Sigma^-$  are called, respectively, the *positive roots* and the *negative roots*. Further, set

$$(3.5.3.3) \quad \mathfrak{n}^+ = \sum_{\alpha \in \Sigma^+} \mathfrak{g}_{\alpha}, \quad \mathfrak{n}^- = \sum_{\alpha \in \Sigma^-} \mathfrak{g}_{\alpha}.$$

Then  $\mathfrak{n}^+$  and  $\mathfrak{n}^-$  are maximal nilpotent Lie subalgebras of  $\mathfrak{g}$ , and we have the decomposition

$$(3.5.3.4) \quad \mathfrak{g} = \mathfrak{a} \oplus \mathfrak{n}^+ \oplus \mathfrak{n}^-.$$

Further, the algebras

$$(3.5.3.5) \quad \mathfrak{b}^+ = \mathfrak{a} \oplus \mathfrak{n}^+, \quad \mathfrak{b}^- = \mathfrak{a} \oplus \mathfrak{n}^-$$

are maximal solvable subalgebras of  $\mathfrak{g}$ . They are called *Borel subalgebras*. The commutator subalgebra of  $\mathfrak{b}^{\pm}$  is  $\mathfrak{n}^{\pm}$ .

Suppose we have a representation of  $\mathfrak{g}$  on a vector space  $V$ . Since the algebra  $\mathfrak{a}$  is commutative, it is possible to have simultaneous eigenvectors for  $\mathfrak{a}$  in  $V$ . Suppose  $v$  is such a vector, i.e., suppose that for all  $a$  in  $\mathfrak{a}$  we have  $a(v) = \lambda(a)v$  for some number  $\lambda(a)$ . It is trivial to check that the function

$$(3.5.3.6) \quad \lambda : a \rightarrow \lambda(a)$$

depends linearly on  $a$ , so that  $\lambda$  belongs to  $\mathfrak{a}^*$ . The linear functional  $\lambda$  is called the *weight* of  $v$ , and  $v$  is called a *weight vector of weight  $\lambda$* . The span of all weight vectors of weight  $\lambda$  is called the  *$\lambda$  weight space*. Suppose  $v$  is not just an eigenvector for  $\mathfrak{a}$ , but for all of  $\mathfrak{b}^+$ ; that is, suppose  $v$  is a weight vector for  $\mathfrak{a}$ , and additionally  $n(v) = 0$  for  $n \in \mathfrak{n}^+$ . Then  $v$  is called a *highest weight vector*, and the weight  $\lambda$  of  $v$  is a *highest weight*. If  $V$  is generated as a  $\mathfrak{g}$ -module by a highest weight vector, then  $V$  is called a *highest weight module*. Just as for  $\mathfrak{sl}_2$  we can prove

**LEMMA 3.5.3.7.** *Every finite-dimensional irreducible representation of  $\mathfrak{g}$  is a highest weight module. More precisely, a finite-dimensional irreducible representation contains a unique highest weight vector.*

**PROOF.** Let  $V$  be the space of the representation. Since  $V$  is irreducible, to show it is a highest weight module it suffices to show it contains a highest weight vector. This is done in completely elementary fashion just as for  $\mathfrak{sl}_2$ . If  $h_0 \in \mathfrak{a}$  is the element used to construct  $\mathfrak{n}^{\pm}$ , observe that if  $n \in \mathfrak{g}_{\alpha} \subseteq \mathfrak{n}^+$ , then  $n$  transforms an eigenvector for  $h_0$  of eigenvalue  $\lambda$  into an eigenvector of eigenvalue  $\lambda + \alpha(h_0)$ , which has larger real part than does  $\lambda$ . Hence if  $\lambda$  has maximal real part among the eigenvalues of  $h_0$  acting on  $V$ , then any eigenvector for  $h_0$  with eigenvalue  $\lambda$  must be annihilated by  $\mathfrak{n}^+$ . Since

$\mathfrak{a}$  is commutative and  $V$  is finite dimensional, we may find within the  $\lambda$ -eigenspace for  $h_0$  a weight vector for  $\mathfrak{a}$ . It is necessarily then a highest weight vector.

To show there is only one highest weight vector, we appeal to the Poincaré-Birkhoff-Witt Theorem (cf. [Jaco1, Serr2], etc.). From equations (3.5.3.4) and (3.5.3.5) we see that

$$\mathfrak{g} = \mathfrak{b}^+ \oplus \mathfrak{n}^-.$$

Let  $\mathcal{U}(\mathfrak{g})$  be the universal enveloping algebra of  $\mathfrak{g}$  (cf. [Jaco1, Serr2], etc.), and similarly for  $\mathfrak{b}^+$ ,  $\mathfrak{n}^-$ . Multiplication inside  $\mathcal{U}(\mathfrak{g})$  induces a linear mapping

$$(3.5.3.8) \quad \mathcal{U}(\mathfrak{n}^-) \otimes \mathcal{U}(\mathfrak{b}^+) \rightarrow \mathcal{U}(\mathfrak{g}).$$

The PBW Theorem tells us that the mapping (3.5.3.8) is a linear isomorphism.

Let  $v \in V$  be a highest weight vector. Denote by  $Cv$  the line through  $v$ . Then using PBW we find

$$\mathcal{U}(\mathfrak{g})(Cv) = \mathcal{U}(\mathfrak{n}^-)\mathcal{U}(\mathfrak{b}^+)(Cv) = \mathcal{U}(\mathfrak{n}^-)(Cv).$$

For each  $\alpha \in \Sigma^-$ , choose a nonzero element  $n_\alpha \in \mathfrak{g}_\alpha$ . Then  $\mathcal{U}(\mathfrak{n}^-)$  is spanned by monomials in the  $n_\alpha$ , i.e., by products  $n_{\alpha_1} n_{\alpha_2} \cdots n_{\alpha_k}$ . An easy inductive calculation shows that, if  $v$  has weight  $\lambda$ , then  $n_{\alpha_1} \cdots n_{\alpha_k}(v)$  is also a weight vector, of weight  $\lambda + \sum_{i=1}^k \alpha_i$ . We note that since  $\alpha(h_0) < 0$  for all  $\alpha$  in  $\Sigma^-$ , no sum  $\sum_{i=1}^k \alpha_i$  can be zero unless  $k = 0$ . Thus we have the following result.

**LEMMA 3.5.3.9.** *If  $V$  is a highest weight module with highest weight  $\lambda$ , then:*

- (i)  $V$  is a direct sum of its weight spaces;
- (ii) all weights of  $V$  have the form  $\lambda + \sum_{\alpha \in \Sigma^-} n_\alpha \alpha$ , where the  $n_\alpha$  are nonnegative integers; and
- (iii) the  $\lambda$ -weight space is one-dimensional, that is, it is  $Cv$ , where  $v$  is the highest weight vector of weight  $\lambda$ .

Now suppose  $V$  is an irreducible highest weight module, with highest weight  $\lambda$ , and suppose  $V$  contains a highest weight vector  $v_1$  in addition to the highest weight vector  $v$  of weight  $\lambda$ . Then by Lemma 3.5.3.9(ii) and (iii), the weight of  $v_1$  is  $\lambda + \sum_{\alpha \in \Sigma^-} m_\alpha \alpha$ , with some of the  $m_\alpha$ 's positive. By Lemma 3.5.3.9(ii) the  $\mathfrak{g}$ -module  $\mathcal{U}(\mathfrak{g})(v_1)$  is the span of weight spaces with weights  $\lambda + \sum_{\alpha \in \Sigma^-} (m_\alpha + n_\alpha) \alpha$ , with the  $n_\alpha$ 's nonnegative. It follows that  $v$  cannot belong to  $\mathcal{U}(\mathfrak{g})(v_1)$ , contradicting the irreducibility of  $V$ . This proves Lemma 3.5.3.7.

In fact, during the argument, we showed a more general fact about highest weight modules, which we will state explicitly.

**COROLLARY 3.5.3.10.** *Let  $V$  be a highest weight module, generated by the highest weight vector  $v$  with highest weight  $\lambda$ . Then*

- (i)  *$V$  is irreducible if and only if  $V$  contains no other highest weight vector, and*
- (ii)  *$V$  contains a unique maximal proper submodule  $U$  such that  $V/U$  is irreducible and nontrivial. (In particular,  $v \notin U$ .)  $U$  is generated by all highest weight vectors other than  $v$ .*

Thus we have identified irreducible finite-dimensional representations as members of a larger family of irreducible highest weight modules. We will now proceed by describing this larger class, then identifying the subclass consisting of finite-dimensional subrepresentations.

First, we show that, as for  $\mathfrak{sl}_2$ , there is a highest weight module with highest weight  $\lambda$  for any  $\lambda \in \mathfrak{a}^*$ . Indeed, given  $\lambda \in \mathfrak{a}^*$ , consider the left ideal  $\mathcal{L}_\lambda$  in  $\mathcal{U}(\mathfrak{g})$  generated by  $\mathfrak{n}^+$  and by elements  $a - \lambda(a)$  for  $a \in \mathfrak{a}$ . Note that  $\lambda$  defines a character (a one-dimensional representation) of  $\mathcal{U}(\mathfrak{b}^+)$ , and that  $\mathfrak{n}^+$  and the elements  $a - \lambda(a)$  generate the kernel of the corresponding homomorphism from  $\mathcal{U}(\mathfrak{b}^+)$  to  $\mathbb{C}$ . Thus they generate a two-sided ideal  $\mathcal{J}_\lambda$  of codimension one in  $\mathcal{U}(\mathfrak{b}^+)$ , and  $\mathcal{L}_\lambda$  is the left ideal in  $\mathcal{U}(\mathfrak{g})$  generated by  $\mathcal{J}_\lambda$ . It follows from PBW that  $\mathcal{L}_\lambda = \mathcal{U}(\mathfrak{n}^-)\mathcal{J}_\lambda$ , and that the natural map

$$(3.5.3.11) \quad \mathcal{U}(\mathfrak{n}^-) \hookrightarrow \mathcal{U}(\mathfrak{g}) \rightarrow \mathcal{U}(\mathfrak{g})/\mathcal{L}_\lambda$$

is a linear isomorphism.

**COROLLARY 3.5.3.12.** (a) *The  $\mathfrak{g}$ -module*

$$(3.5.3.13) \quad V_\lambda = \mathcal{U}(\mathfrak{g})/\mathcal{L}_\lambda$$

*is a highest weight module, with highest weight  $\lambda$ , generated by the image  $v_\lambda$  of 1, the identity element of  $\mathcal{U}(\mathfrak{g})$ .*

(b)  *$V_\lambda$  is free as a  $\mathcal{U}(\mathfrak{n}^-)$  module.*

(c) *Any highest weight module with highest weight  $\lambda$  is a quotient of  $V_\lambda$ .*

(d) *Consequently, for every weight  $\lambda \in \mathfrak{a}^*$ , there exists a unique irreducible highest weight module  $M_\lambda$  with highest weight  $\lambda$ .*

The modules  $V_\lambda$  are usually called *Verma modules* [Hump, BGG3].

Thus we have an irreducible highest weight module  $M_\lambda$  for every  $\lambda \in \mathfrak{a}^*$ . It remains to decide when  $M_\lambda$  is finite dimensional. We can deduce some restrictions on  $\lambda$  from our knowledge of  $\mathfrak{sl}_2$ . Suppose  $M_\lambda$  is finite dimensional. For a positive root  $\alpha \in \Sigma^+$ , consider the copy  $\mathfrak{s}_\alpha$  of  $\mathfrak{sl}_2$  generated by  $\mathfrak{g}_{\pm\alpha}$ . The highest weight vector  $v_\lambda$  of  $M_\lambda$  generates a highest weight module for  $\mathfrak{s}_\alpha$ , and this highest weight module is necessarily finite dimensional. It follows from §3.5.2 that  $\lambda(h_\alpha)$  is a nonnegative integer. Let us say  $\lambda \in \mathfrak{a}^*$  is *integral* if  $\lambda(h_\alpha)$  is an integer for all  $\alpha \in \Sigma^+$ . Let us say  $\lambda \in \mathfrak{a}^*$  is *dominant* if  $\lambda(h_\alpha) \geq 0$  for all  $\alpha \in \Sigma^+$ . (This is equivalent to saying  $\lambda$  is in the positive or fundamental Weyl chamber, cf. §2.10.) Then, for  $\lambda \in \mathfrak{a}^*$  to be

the highest weight of an irreducible finite-dimensional representation of  $\mathfrak{g}$ , we can say it must be dominant and integral. The main result of Cartan's highest weight theory is that these conditions on  $\lambda$  suffice to guarantee  $M_\lambda$  is finite dimensional.

**THEOREM 3.5.3.14.** *The irreducible module  $M_\lambda$  of highest weight  $\lambda$  is finite dimensional if and only if  $\lambda$  is dominant and integral.*

**REMARKS.** (a) This theorem reminds us again of the strong control  $\mathfrak{sl}_2$  exerts over the phenomena of semisimple Lie algebras. This control is evident even more in the proof of the theorem given below.

(b) The dominant integral  $\lambda$  in  $\mathfrak{a}^*$  clearly forms a semigroup under addition—the intersection of a lattice with a cone. If  $M_\lambda$  and  $M_\mu$  are irreducible highest weight modules with highest weight vectors  $v_\lambda, v_\mu$ , then the tensor product  $v_\lambda \otimes v_\mu$  will generate a highest weight module, of highest weight  $\lambda + \mu$ , inside the tensor product module  $M_\lambda \otimes M_\mu$ . Hence, if  $M_\lambda, M_\mu$  are finite dimensional, so must  $M_{\lambda+\mu}$  be. Thus the set of highest weights of finite-dimensional representations is also a semigroup. To prove that all dominant integral  $\lambda$  define finite-dimensional highest weight modules, it suffices to exhibit finite-dimensional  $M_\lambda$  for a set of  $\lambda$  which generate the semigroup of dominant integral weights. This is essentially what Cartan did [Crtn2], and in fact the procedure, though heavily computational for the exceptional groups, is illuminating, and for the classical groups is quite elegant, involving the exterior powers of the standard representations. From general structure theory [Jaco1, Hump] one can show that the dominant integral weights actually form a free semigroup on a unique set of  $\text{rank}(\mathfrak{g}) = \dim \mathfrak{a}$  generators. The representations corresponding to these generators are called the *fundamental representations* of  $\mathfrak{g}$ . For  $\mathfrak{g} = \mathfrak{sl}_n$ , the fundamental representations are just the natural action on the  $\Lambda^j(\mathbb{C}^n)$ , the exterior powers of  $\mathbb{C}^n$ , for  $1 \leq j \leq n-1$ . For orthogonal and symplectic Lie algebras, the fundamental representations (except for the spin representations of the orthogonal algebras [Arti, BeTu, Jaco2]) are also constructed fairly easily from the exterior powers of the basic representation.

We will briefly sketch the approach of [HaCh1] (see also [Jaco1, Hump]) to showing that, if  $\lambda$  is dominant integral, then  $M_\lambda$  is finite dimensional. Consider the fundamental positive roots in  $\Sigma^+$  (cf. §2.12). Let  $\alpha$  be a fundamental positive root. From the general structure theory, we know that  $\mathfrak{p}_\alpha$ , defined by

$$(3.5.3.15) \quad \mathfrak{p}_\alpha = \mathfrak{b}^+ + \mathfrak{g}_{-\alpha} = \mathfrak{n}_{(\alpha)}^+ \oplus \ker \alpha \oplus \mathfrak{s}_\alpha,$$

where  $\mathfrak{n}_{(\alpha)}^+ = \sum_{\beta \in \Sigma^+, \beta \neq \alpha} \mathfrak{g}_\beta$  and  $\ker \alpha = \{h : \alpha(h) = 0\} \subseteq \mathfrak{a}$ , is a Lie subalgebra of  $\mathfrak{g}$ . It is called a *parabolic subalgebra*. The subspace  $\mathfrak{n}_{(\alpha)}^+$  is an ideal in  $\mathfrak{p}_\alpha$ . In particular, we have

$$(3.5.3.16) \quad [\mathfrak{n}_{(\alpha)}^+, \mathfrak{s}_\alpha] \subseteq \mathfrak{n}_{(\alpha)}^+.$$

Consider the Verma module  $V_\lambda$  with highest weight vector  $v_\lambda$ . Suppose that  $\lambda(h_\alpha)$  is a nonnegative integer. Then formulas (3.5.1.8) show that if  $e_{-\alpha}$  belongs to  $\mathfrak{g}_{-\alpha}$ , and  $e_\alpha$  belongs to  $\mathfrak{g}_\alpha$ , the vector

$$y = e_{-\alpha}^{\lambda(h_\alpha)+1}(v_\lambda)$$

is annihilated by  $e_\alpha$ . Also, the commutation relations (3.5.3.16) imply that  $y$  will be annihilated by  $\mathfrak{n}_{(\alpha)}^+$ . Since  $\mathfrak{n}^+ = \mathfrak{n}_{(\alpha)}^+ \oplus \mathfrak{g}_\alpha$ , it follows that  $y$  is a highest weight vector, of weight  $\lambda - (\lambda(h_\alpha) + 1)\alpha$ . It will generate a highest weight submodule of  $V_\lambda$ . Since  $\mathfrak{n}^-$  acts freely on  $V_\lambda$ , we see that  $y$  generates a module isomorphic to  $V_{\lambda - (\lambda(h_\alpha) + 1)\alpha}$ . In other words, under the hypothesis that  $\lambda(h_\alpha)$  is a nonnegative integer, we obtain an embedding of  $V_{\lambda - (\lambda(h_\alpha) + 1)\alpha}$  in  $V_\lambda$ .

If  $\lambda$  is dominant integral, then we get an embedding of  $V_{\lambda - (\lambda(h_\alpha) + 1)\alpha}$  in  $V_\lambda$  for every fundamental root  $\alpha$ . This already suffices to show that  $M_\lambda$ , the irreducible quotient of  $V_\lambda$ , must be finite dimensional. Indeed, for each fundamental root  $\alpha$ ,  $M_\lambda$  will be a quotient of  $V_\lambda / V_{\lambda - (\lambda(h_\alpha) + 1)\alpha} = V_\lambda(\alpha)$ . The image in  $V_\lambda(\alpha)$  of the highest weight vector  $v_\lambda$  generates a finite-dimensional  $\mathfrak{s}_\alpha$ -module. Since the adjoint action of  $\mathfrak{g}$  on  $\mathcal{U}(\mathfrak{g})$  is a sum of finite-dimensional  $\mathfrak{g}$ -modules, hence  $\mathfrak{s}_\alpha$ -modules, it follows that any element of  $V_\lambda(\alpha) = \mathcal{U}(\mathfrak{g})(v_\lambda)$  generates a finite-dimensional  $\mathfrak{s}_\alpha$ -module. It follows that  $S_\alpha = \exp \mathfrak{s}_\alpha$ , the group obtained by exponentiating  $\mathfrak{s}_\alpha$ , acts on  $V_\lambda(\alpha)$ . In particular, the Weyl group reflection  $w_\alpha$  contained in  $S_\alpha$  acts on  $V_\lambda(\alpha)$ . It is easy to see this fact remains true in any quotient  $\mathfrak{g}$ -module of  $V_\lambda(\alpha)$ . In particular,  $w_\alpha$  acts on  $M_\lambda$ . Since the  $w_\alpha$  generate the full Weyl group  $W$  (cf. §2.9), we see that  $W$  acts on  $M_\lambda$ .

Since  $W$  normalizes  $\mathfrak{a}$ , the effect of  $W$  on  $M_\lambda$  is to permute weight spaces. Precisely, for  $\mu \in \mathfrak{a}^*$ , let  $M_\lambda^\mu$  denote the  $\mu$  weight space of  $M_\lambda$ . Then for  $p \in W$ , we have

$$(3.5.3.17) \quad p(M_\lambda^\mu) = M_\lambda^{p(\mu)},$$

where  $p(\mu)$  denotes the standard action of  $p$  on  $\mu$  as an element of  $\mathfrak{a}^*$ . Thus, in particular, one sees that the set of weights  $\mu$  for which  $M_\lambda^\mu \neq \{0\}$  is invariant under  $W$ . Since also the weights of  $M_\lambda$ , being contained in the weights of  $V_\lambda$ , are bounded above, as described by Lemma 3.5.3.9(ii), it follows easily from the geometry of the action of  $W$  on  $\mathfrak{a}^*$  that the set of weights  $\mu$  for which  $M_\lambda^\mu \neq 0$  must be bounded, hence finite in number. Since each weight space of  $M_\lambda$  (indeed, of  $V_\lambda$ ) is finite dimensional, we conclude  $M_\lambda$  is finite dimensional.

Although the argument above gives us the desired finite dimensionality of  $M_\lambda$  when  $\lambda$  is dominant integral, it does not give us a very precise picture of  $M_\lambda$ . A refinement of the above considerations yields a description of  $M_\lambda$  analogous to (3.5.2.1) [BGG1–3, Dixm1].



Let  $\rho$  denote the element of  $\mathfrak{a}^*$  such that

$$(3.5.3.18) \quad \rho(h_\alpha) = 1$$

for each fundamental root  $\alpha$ . Then we may write

$$(3.5.3.19) \quad \lambda - (\lambda(h_\alpha) + 1)\alpha = w_\alpha(\lambda + \rho) - \rho,$$

where again  $w_\alpha$  is the Weyl group reflection corresponding to the fundamental root  $\alpha$ .

As we have noted, the Weyl group  $W$  is generated by the reflections  $w_\alpha$ . Let the *length* of  $p \in W$  be the shortest product of the  $w_\alpha$ 's equaling  $p$  [Hill, Bour]. Denote it by  $l(p)$ . If

$$p = w_{\alpha_l} w_{\alpha_{l-1}} \cdots w_{\alpha_1}, \quad l = l(p),$$

is a shortest possible product expressing  $p$ , then

$$p' = w_{\alpha_l} p = w_{\alpha_{l-1}} \cdots w_{\alpha_1}$$

has length  $l - 1$ . From a systematic study of the geometry of a root system and its Weyl group, one can see that if  $\lambda$  is dominant, then  $p'(\lambda)(h_{\alpha_l}) \geq 0$ . It follows by the argument given above that  $V_{p(\lambda+\rho)-\rho}$  embeds in  $V_{p'(\lambda+\rho)-\rho}$ . The embedding is unique up to multiples.

By induction, we find that when  $\lambda$  is dominant integral, we can embed  $V_{w(\lambda+\rho)-\rho}$  in  $V_\lambda$  for every element  $w$  of the Weyl group  $W$ . It is shown in [BGG2] that these embeddings can be organized into an exact sequence, as follows. For  $k \geq 0$ , set

$$(3.5.3.20) \quad V_\lambda^{(k)} = \sum_{l(w)=k} V_{w(\lambda+\rho)-\rho}.$$

We have seen that whenever  $w$  has length  $k - 1$  and  $w_\alpha w$  has length  $k$ , there is an embedding  $V_{w_\alpha w(\lambda+\rho)-\rho} \rightarrow V_{w(\lambda+\rho)-\rho}$ , defined up to multiples. By taking linear combinations of these embeddings, we can construct mappings from  $V_\lambda^{(k)}$  to  $V_\lambda^{(k-1)}$ . If we choose these mappings correctly, we will get an exact sequence

$$(3.5.3.21) \quad 0 \rightarrow V_\lambda^{(m)} \rightarrow V_\lambda^{(m-1)} \rightarrow \cdots \rightarrow V_\lambda^{(2)} \rightarrow V_\lambda^{(1)} \rightarrow V_\lambda \rightarrow M_\lambda \rightarrow 0,$$

where  $m$  is the largest possible length of an element of  $W$ . In fact,  $m = \dim \mathfrak{n}^-$ . In the case of  $\mathfrak{sl}_2$ , this exact sequence is simply the sequence (3.5.2.1).

REMARKS. (a) The exact sequence (3.5.3.21) implies the Weyl character formula (cf. §3.5.4) by means of the Euler-Poincaré principle. The alternating sum of the highest weights of the  $V_{w(\lambda+\rho)-\rho}$  provides the numerator for the formula, while the character of  $V_0$  ( $\simeq \mathfrak{n}^-$  as an  $\mathfrak{a}$ -module) provides the celebrated "Weyl denominator."

(b) The multiplicities of the weight spaces of  $V_0 \simeq \mathfrak{n}^-$  are easily seen by PBW to be given by the *Kostant partition function* [Kost5, Jaco1, Hump],

$P(\lambda) = \#$  of ways of expressing  $\lambda$  as an integral linear combination of negative roots. Given this observation, Kostant's multiplicity formula [Kost5, Jaco1, Hump] for the multiplicities of weights of finite-dimensional representations follows immediately from (3.5.3.21). Indeed, Kostant's formula is basically a variant way of expressing the Weyl character formula, so when we can deduce one, we should be able to deduce the other.

(c) Also from the exact sequence (3.5.3.21), one can fairly directly deduce Kostant's description [Kost4, Warn, Knap1, Voga2, Arib] of the Lie algebra cohomology groups  $H^k(\mathfrak{n}^+, M_\lambda)$ -cohomology of  $\mathfrak{n}^+$  with coefficients in the module  $M_\lambda$ . We will discuss this in §3.5.5.

(d) The " $\rho$ -shift" seen in the highest weights of the  $V_{w(\lambda+\rho)-\rho}$ , and in the Weyl character formula, and elsewhere is in some sense explained by the Harish-Chandra homomorphism (cf. Theorem 3.5.5.23)).

**3.5.4. WEYL'S CHARACTER FORMULA.** In [Weyl1], (see also [Weyl2, Wall2, Knap]), Hermann Weyl gave a radically different approach to the representation theory of complex semisimple Lie algebras through the equivalent theory of representations of compact semisimple groups. (Part of his achievement was to make explicit the equivalence. This is the origin of the celebrated "unitary trick.") This approach yields not only the classification of irreducible representations but also a formula for their characters, the Weyl character formula. (We note that the Weyl character formula for  $U_n$  (and also for  $O_n$ ) is due to Schur [Schu].)

We will illustrate the method with the unitary group  $U_n$  in order not to become too involved with the notation necessary for the general case.

We think of  $U_n$  as a set of  $n \times n$  matrices. The subgroup (a Cartan subgroup)

$$(3.5.4.1) \quad A = \left\{ \begin{bmatrix} z_1 & & & \\ & z_2 & & 0 \\ & & \ddots & \\ & 0 & & z_n \end{bmatrix} : z_i \in \mathbb{C}, |z_i| = 1 \right\}$$

of unitary diagonal matrices is abelian and isomorphic to  $\mathbb{T}^n$ , the  $n$ -fold power of  $\mathbb{T}$ , the unit circle in  $\mathbb{C}$ . The unitary characters (irreducible representations) of  $A$  define a group isomorphic to  $\mathbb{Z}^n$ . They may be explicitly described by the formula

$$(3.5.4.2) \quad \chi_m(a) = \prod_{i=1}^n z_i^{m_i},$$

where  $m = (m_1, m_2, \dots, m_n)$  is an  $n$ -tuple of integers, and

$$a = a(z) = \text{diag}(z_1, z_2, \dots, z_n)$$

is the diagonal matrix with diagonal entries  $z_i \in \mathbb{T}$ . The characters  $\chi_m(a)$  form an orthonormal basis for  $L^2(A)$  with respect to Haar measure on  $A$

(assuming, as we will, that the Haar measure is normalized so that the total volume of  $A$  is 1).

Spectral theory for unitary matrices (cf. [Lang3, Stra], etc.) tells us that every unitary matrix is conjugate to a diagonal matrix. Thus the map

$$(3.5.4.3) \quad \Gamma: A \times U_n \rightarrow U_n, \quad \Gamma(a, g) = gag^{-1}$$

is surjective. It is clear that  $\Gamma(a, gb) = \Gamma(a, g)$  for  $b \in A$ . Hence the map  $\Gamma$  actually factors to

$$\tilde{\Gamma}: A \times (U_n/A) \rightarrow U_n.$$

The factored map  $\tilde{\Gamma}$  is generically finite-to-one. On the open dense set of matrices with  $n$  distinct eigenvalues, it is an  $n!$ -to-one covering map: two diagonal matrices define the same conjugacy class in  $U_n$  if and only if one can be turned into the other by permuting its diagonal entries. One can think of  $\tilde{\Gamma}$  as defining a system of "polar coordinates" on  $U_n$ .

Let  $dg$  denote Haar measure on  $U_n$ . Since  $\tilde{\gamma}$  is finite-to-one, up to sets of measure zero we can use it to lift  $dg$  up to  $A \times (U_n/A)$ . Thus we can find a unique measure  $d\mu(a, \dot{g})$  on  $A \times (U_n/A)$  such that the set where  $\tilde{\Gamma}$  is singular has measure zero and such that on the set where  $\tilde{\Gamma}$  is finite-to-one, we have the formula

$$(3.5.4.4) \quad \int_{A \times (U_n/A)} f(a, \dot{g}) d\mu(a, \dot{g}) = \int_{U_n} \left( \sum_{x \in \tilde{\Gamma}^{-1}(g)} f(x) \right) dg$$

for  $f$  a function on  $A \times (U_n/A)$ .

The coset space  $U_n/A$  also possesses a left-invariant measure  $d\dot{g}$ . Since Haar measure on  $U_n$  is also conjugation invariant, we see  $d\mu$  must be a product measure of the form

$$d\mu(a, \dot{g}) = d\nu(a) d\dot{g}.$$

Since we are in a context of smooth manifolds and smooth maps, we can easily believe that  $d\nu$  is absolutely continuous with respect to Haar measure  $da$  on  $A$ :

$$d\nu(a) = \nu(a) da$$

for an appropriate function  $\nu$  on  $A$ .

If we think of the map  $\tilde{\Gamma}$  as partitioning  $U_n$  into a family, parametrized (redundantly) by  $A$ , of fibers which are copies of  $U_n/A$ , then  $\nu(a)$  tells us the volume of the fiber through  $a$ . This volume can be computed, up to a constant factor, as the determinant of an appropriate Jacobian mapping, which can be identified with the action  $(1 - \text{Ad } a)_{|\mathfrak{a}^\perp}$  of  $a$  acting by conjugation on  $\mathfrak{a}^\perp$ , the orthogonal complement to  $\mathfrak{a}$  in  $\mathfrak{u}_n$ . (Note that, concretely,  $\mathfrak{a}^\perp$  is the space of skew-adjoint  $n \times n$  matrices with zeros on the diagonal.) It is easy to compute that [Wall2, Knap2, HoTa]

$$\nu(a) = c |\det(1 - \text{ad } a_{|\mathfrak{a}^\perp})| = c \prod_{1 \leq i < j \leq n} |z_i - z_j|^2$$

for an appropriate constant  $c$ . Here the  $z_i$  are the diagonal entries of  $a$ , as in formula (3.5.4.2). We will write

$$(3.5.4.5) \quad D(a) = \prod_{1 \leq i < j \leq n} (z_i - z_j).$$

(The function  $D$  is known as the discriminant; in our context it will play the role of the denominator in Weyl's character formula for  $U_n$ .) Then our formula for  $\nu(a)$  can be written as

$$(3.5.4.6) \quad \nu(a) = cD(a)\overline{D(a)}.$$

Here  $\overline{D(a)}$  denotes the complex conjugate of  $D(a)$ .

In formula (3.5.4.4), let us take the function  $f$  to be a pull-back from  $U_n$  by  $\tilde{\Gamma}$ :

$$f(a, \dot{g}) = \phi(\tilde{\Gamma}(a, \dot{g})) = \phi(\dot{g}a\dot{g}^{-1})$$

for some function  $\phi$  on  $U_n$ . Taking into account the discussion above, we see

$$\#(W) \int_{U_n} \phi(g) dg = \int_{A \times (U_n/A)} \phi(\dot{g}a\dot{g}^{-1})\nu(a) da d\dot{g},$$

where  $W \simeq S_n$  here indicates the group of permutations—the Weyl group of  $U_n$ . Suppose further that  $\phi$  is invariant under conjugation. Then our formula simplifies to become

$$(3.5.4.7) \quad \int_{U_n} \phi(g) dg = \frac{c}{\#(W)} \int_A \phi(a)D(a)\overline{D(a)} da.$$

This formula has a nice interpretation in terms of  $L^2$ -spaces. Let  $\varphi_1, \varphi_2$  be two conjugation invariant functions. Setting  $\varphi = \varphi_1\overline{\varphi_2}$  gives

$$(3.5.4.8) \quad \int_{U_n} \varphi_1(g)\overline{\varphi_2(g)} dg = \frac{c}{\#(W)} \int_A (\varphi_1 D)(a)\overline{(\varphi_2 D)(a)} da.$$

Let  $L^2(U_n)^{\text{Ad } U_n}$  be the space of conjugation-invariant  $L^2$ -functions on  $U_n$ . The restriction of  $\varphi \in L^2(U_n)^{\text{Ad } U_n}$  to  $A$  will be invariant under the Weyl group  $W = S_n$  of permutations of the diagonal coordinates. On the other hand, the discriminant function  $D$  is easily seen to be completely antisymmetric in the  $z_i$ ; more precisely we have

$$(3.5.4.9) \quad D(p(a)) = \text{sgn}(p)D(a), \quad a \in A, p \in S_n,$$

where  $\text{sgn}: S_n \rightarrow \pm 1$  is the sign character:  $\text{sgn}(p)$  is 1 if  $p$  is an even permutation, and  $\text{sgn}(p)$  is  $-1$  if  $p$  is odd. Thus the mapping

$$(3.5.4.10) \quad M_D: \varphi \rightarrow D\varphi$$

will take  $W$ -invariant or “symmetric” functions to “skew-symmetric” functions, i.e., functions transforming under  $W$  by the sign character. Let  $L^2(A)^{W, \text{sgn}}$  denote the subspace of skew-symmetric functions in  $L^2(A)$ . Let

$$(3.5.4.11) \quad \text{res}_A: f \rightarrow f|_A$$

denote the restriction map from functions on  $U_n$  to functions on  $A$ . With this notation, we may express formula (3.5.4.8) as follows:

$$(3.5.4.12) \quad \text{The map } M_D \circ \text{res}_A : L^2(U_n)^{\text{Ad } U_n} \rightarrow L^2(A)^{W, \text{sgn}} \text{ is, up to a scalar factor, a unitary isomorphism.}$$

The discriminant function  $D$  is distinguished among all skew-symmetric functions by the property that it divides any one of them. More precisely, if  $f$  is a smooth skew-symmetric function on  $A$ , then we can write  $f = D\varphi$  and the quotient  $\varphi$ , which will obviously be a symmetric function, will also be smooth. To see this, it suffices to consider two variables at a time: to show, say, that if  $f$  changes sign when  $z_1$  and  $z_2$  are interchanged, then  $f$  is divisible by  $z_1 - z_2$ . This can be done, for example, in terms of Fourier series. The basic formula is

$$z_1^n - z_2^n = (z_1 - z_2)(z_1^{n-1} + z_1^{n-2}z_2 + \cdots + z_2^{n-1}).$$

This argument shows that, in fact, if  $f$  has a finite Fourier series, then  $\varphi$  will also.

One way to create skew-symmetric functions is to take an arbitrary function  $f$  on  $A$  and skew-symmetrize it. Thus given  $f$ , we define

$$(3.5.4.13) \quad \text{skew}(f)(a) = \sum_{p \in W} \text{sgn}(p)f(p(a)).$$

It is simple to check that  $\text{skew}(f)$  is skew-symmetric, and if  $f$  is already skew-symmetric, then  $\text{skew}(f) = \#(W)f$ .

Consider  $\text{skew}(\chi_m)$  for some character  $\chi_m$  of  $A$ , as in formula (3.5.4.2). The Weyl group also acts naturally on characters, by permuting the coordinates of the  $n$ -tuple  $m$  labeling  $\chi_m$ . Specifically we have

$$w(m) = (m_{w^{-1}(1)}, m_{w^{-1}(2)}, \dots, m_{w^{-1}(n)})$$

and

$$\chi_{w(m)}(w(a)) = \chi_a.$$

From these formulas, it is clear that

$$(3.5.4.14) \quad \text{skew}(\chi_{w(m)}) = \text{sgn}(w) \text{skew}(\chi_m).$$

Thus in constructing the functions  $\text{skew}(\chi_m)$ , we need only consider  $m$  modulo the action of  $W$ . Thus let us define

$$(3.5.4.15) \quad \hat{A}^+ = \{\chi_m : m_1 \geq m_2 \geq \cdots \geq m_n\}.$$

It is easy to check that any character can be transformed by some element of  $W$  to a unique element of  $\hat{A}^+$ . Thus we need only consider  $\text{skew}(\chi_m)$  for  $\chi_m$  in  $\hat{A}^+$ .

We can also see from equation (3.5.4.14) that if any two coordinates of  $m$  are equal, then  $\text{skew}(\chi_m) = 0$ . Thus in fact it is sufficient to consider  $\text{skew}(\chi_m)$  for  $\chi_m$  belonging to

$$(3.5.4.16) \quad \hat{A}^{++} = \{\chi_m : m_1 > m_2 > \cdots > m_n\}.$$

Using elementary facts about Fourier series, we can see

$$(3.5.4.17) \quad \text{The functions } \#(W)^{-(1/2)} \text{skew}(\chi_m), \quad m \in \widehat{A}^{++}, \text{ define an orthonormal basis for } L^2(A)^{W, \text{sgn}}.$$

Let us remark that  $\widehat{A}^{++}$  has a minimal element. That is, if we define

$$(3.5.4.18) \quad \rho = (n-1, n-2, \dots, 1, 0)$$

then

$$(3.5.4.19) \quad \widehat{A}^{++} = \chi_\rho \widehat{A}^+ = \{\chi_\rho \chi_m = \chi_{\rho+m} : \chi_m \in \widehat{A}^+\}.$$

Since the functions  $\text{skew}(\chi_m)$  span the skew-symmetric functions, we must be able to express  $D$  (cf. (3.5.4.5)) as a linear combination of the  $\text{skew}(\chi_m)$ . In fact, by considering which characters could possibly occur in the expansion of the product defining  $D$ , we can conclude

$$(3.5.4.20) \quad D = \text{skew}(\chi_\rho)$$

with  $\rho$  as in (3.5.4.18). This identity, which is equivalent to the evaluation of the Vandermond determinant

$$\det \begin{vmatrix} 1 & z_1 & z_1^2 & \cdots & z_1^{n-1} \\ 1 & z_2 & z_2^2 & \cdots & z_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_n & z_n^2 & \cdots & z_n^{n-1} \end{vmatrix} = \prod_{1 \leq i < j \leq n} (z_j - z_i) = (-1)^{n(n-1)/2} D,$$

is one of the most fertile in mathematics. In [Macd2], I. G. Macdonald discovered a class of identities attached to affine root systems that turned out to be analogs of (3.5.4.20) for affine Kac-Moody Lie algebras [Kac1]. The developments of this theme are still proceeding at a rapid pace (cf. [KaPe, Lepo1, 2, LeMi, Macd3, Gust, Heck 1-3, Morr, HeOp, Opda1-3, Zeil], etc.).

Before continuing, let us note one consequence of the identity (3.5.4.20): it allows us to explicitly determine the constant  $c$  in formulas (3.5.4.7) and (3.5.4.8). Indeed, formula (3.5.4.20) tells us that  $D$  is the sum of  $n!$  characters of  $A$ , with coefficients  $\pm 1$ . Since characters are orthonormal in  $L^2(A)$ , we conclude

$$\int_A |D|^2(a) da = n! = \#(W).$$

Using this and  $\varphi = 1$ , the constant function, in (3.5.4.7) tells us that  $c = 1$ . Hence formula (3.5.4.8) reads simply

$$(3.5.4.21) \quad \int_{U_n} \varphi_1(g) \overline{\varphi_2(g)} dg = \frac{1}{\#(W)} \int_A (\varphi_1 D)(a) \overline{\varphi_2 D(a)} da,$$

for  $\varphi_1, \varphi_2 \in L^2(U_n)^{\text{Ad } U_n}$ .

Now turn to consideration of the irreducible characters of  $U_n$ . These are the functions

$$(3.5.4.22) \quad \text{ch}_\rho(g) = \text{trace } \rho(g),$$

where  $\rho : U_n \rightarrow \text{GL}(V)$  is an irreducible representation of  $U_n$  on a finite-dimensional vector space  $V$ . Evidently a similar definition can be made for any compact group. The significance of the characters for the representation theory of compact groups is summarized by the Peter-Weyl Theorem (cf. [HeRo, Knap2, PeWe, Loom], etc.). To state it we need some notation.

Let  $G$  be a compact group. Let  $dg$  be Haar measure on  $G$ , normalized so that  $G$  has total mass equal to 1. Let  $L^2(G)$  be the  $L^2$ -space of  $G$  with respect to  $dg$ , and let  $L^2(G)^{\text{Ad } G}$  denote the subspace of conjugation-invariant functions.

We can convolve elements of  $L^2(G)$  (cf. §A.1.12). It is easy to check, for any locally compact group, that the convolution of two  $L^2$ -functions is continuous. Since we have  $G$  compact, continuous functions are  $L^2$ , so  $L^2$  is an algebra under convolution. In fact,

$$\begin{aligned} \|f_1 * f_2\|_2 &= \left\| \int_G f_1(g)(L_g f_2) dg \right\|_2 \leq \int_G |f_1(g)| \|f_2\|_2 dg \\ &\leq \|f_1\|_1 \|f_2\|_2 \leq \|f_1\|_2 \|f_2\|_2. \end{aligned}$$

Here,  $\|f\|_p$  denotes the  $L_p$ -norm of a function on  $G$ . The last inequality follows since  $G$  has total mass 1. It is easy to check that  $L^2(G)^{\text{Ad } G}$  is the center of  $L^2(G)$ .

Let  $V$  be a finite-dimensional vector space, and  $\rho : G \rightarrow \text{GL}(V)$  a representation of  $G$  on  $V$ . We can define the character of  $\rho$ ,  $\text{ch}_\rho$ , by formula (3.5.4.22).

Recall that  $\hat{G}$  denotes the set of irreducible unitary representations of  $G$ .

**THEOREM 3.5.4.23 (Peter-Weyl).** *Let  $G$  be a compact group.*

(a) *Every continuous irreducible representation  $\sigma$  of  $G$  is finite dimensional and unitary (i.e., given irreducible  $\sigma$  acting on  $V$ , there is a  $G$ -invariant hermitian inner product on  $V$ ).*

(b) *Every irreducible representation of  $G$  can be realized as a subrepresentation of the left regular representation on  $L^2(G)$ .*

(c) *The irreducible characters  $\text{ch}_\sigma$ ,  $\sigma \in \hat{G}$ , constitute an orthonormal basis for  $L^2(G)^{\text{Ad } G}$ .*

(d) *The functions  $e_\sigma = (\dim \sigma) \text{ch}_\sigma$ ,  $\sigma \in \hat{G}$ , are idempotents for the convolution algebra structure on  $L^2(G)$ . They are precisely the minimal central idempotents in  $L^2(G)$ . Thus we have a decomposition*

$$\begin{aligned} L^2(G) &\simeq \sum_{\sigma \in \hat{G}} e_\sigma * L^2(G) = \sum_{\sigma \in \hat{G}} e_\sigma * L^2(G) * e_\sigma \\ &= \sum_{\sigma \in \hat{G}} \sigma \otimes \sigma^* \end{aligned}$$

*of  $L^2(G)$  into mutually orthogonal, minimal, two-sided ideals. Each ideal  $e_\sigma * L^2(G) * e_\sigma$  is isomorphic to a matrix algebra of rank  $\dim \sigma$ , and as a*

$G \times G$  module under left and right translation, is isomorphic to  $\sigma \otimes \sigma^*$ , where  $\sigma^*$  indicates the contragredient of  $\sigma$ .

REMARKS. (i) Part (b) is proved by considering matrix coefficients (see §A.1.11). Given part (b), part (a) is an application of the theory of integral operators and the spectral theorem for compact selfadjoint operators [Lang2, RiNa] (or for selfadjoint algebras of compact operators). Parts (c) and (d) are analogs of the Schur orthogonality relations for finite groups, and are proved in essentially the same way, via Schur's Lemma [HeRo, Knap2, Lang3].

(ii) To quote the Peter-Weyl Theorem in the derivation of the Weyl character formula is unhistorical as [PeWe] appeared several years after [Weyl1]. However, it is natural.

With these preparations, we are ready to state

(3.5.4.24) *Weyl character formula for  $U_n$* : The characters of the irreducible representations of  $U_n$  are the functions

$$\frac{\text{skew}(\chi_{m+\rho})}{\text{skew}(\chi_\rho)} = \frac{\text{skew}(\chi_{m+\rho})}{D}, \quad m \in \hat{A}^+.$$

Here  $\text{skew}(\chi_m)$  is defined in formula (3.5.4.13).

PROOF. Indeed, we know that if  $\sigma$  is a representation of  $U_n$ , then  $\sigma|_A$  will be a direct sum of irreducible representations, i.e., characters, of  $A$ . Thus  $(\text{ch}_\sigma)|_A = \text{ch}_{(\sigma|_A)}$  will be a *positive integer* linear combination of elements of  $\hat{A}$ . Also, of course,  $\text{ch}_\sigma$  is conjugation invariant, so  $\text{ch}_{(\sigma|_A)}$  is symmetric. Thus the product  $D(\text{ch}_{(\sigma|_A)})$  will be an integer linear combination of characters of  $A$ , and will be skew-symmetric. It follows easily that  $D(\text{ch}_{(\sigma|_A)})$  is an *integer* linear combination of the functions  $\text{skew}(\chi_m)$ ,  $m \in \hat{A}^{++}$ .

On the other hand, we know from Schur orthogonality, Theorem 3.5.4.23 (c), that the norm of  $\text{ch}_\sigma$  in  $L^2(U_n)$  is 1. It follows from (3.5.4.21) that the norm of  $D(\text{ch}_{(\sigma|_A)})$  in  $L^2(A)$  is  $\#(W)^{1/2}$ . Combining this with the previous paragraph forces  $D(\text{ch}_{(\sigma|_A)})$  to be  $\pm \text{skew}(\chi_m)$  for a single  $\chi_m$ . The sign can be checked by inspecting the coefficient of  $\chi_{m-\rho}$  in  $\text{skew}(\chi_m)/\text{skew}(\chi_\rho)$ , and seeing it is positive (in fact, it is 1). The fact that all  $\chi_m$  in  $\hat{A}^{++}$  are needed to express the characters follows from the completeness part of Theorem 3.5.4.23(c).

REMARKS. (a) To me, this proof is simply magical. If you attempt to analyze it, it dissolves into a few simple calculations and some general nonsense-airy nothing.

(b) One can recognize the same objects appearing here as in §3.5.3. The set  $\hat{A}^+$  is the collection of dominant integral weights,  $\rho$  is the half-sum of the positive roots, the alternating sum in (3.5.4.24) mirrors the Euler characteristic of the exact sequence (3.5.3.21), there is the same phenomenon of shifting by  $\rho$ , etc.



(c) A remarkable feature of this proof is that it simply identifies the irreducible characters. What the representations associated with these characters might be is, for the purposes of this argument, irrelevant and ignored. Of course, what the modules are was known well before this argument was given, from the highest weight theory described in §3.5.3. However, for non-compact groups the situation was reversed: Harish-Chandra [HaCh19, 20] gave a construction of discrete series characters using methods extending those explained in §3.5.6, well before these modules were constructed [OkOz, Schm1–3, Hott, Part1, Wall4]. Further, the early explicit constructions of discrete series modules all depended on knowing the character. It was not until [AtSc, FlJe, Wall2] that the existence of discrete series representations was established independently of character theory.

(d) Of course, the representation with character  $\text{skew}(\chi_{m+\rho})/\text{skew}(\chi_\rho)$  is the representation with highest weight  $\chi_m$ .

(e) By a l'Hopital's Rule argument as  $a \in A$  approaches the identity, one obtains from the character formula a formula for  $\dim \sigma$  (cf. [Weyl2, Knap2, Jaco1], etc.).

(f) In the case of  $U_n$ , the character formula is due to I. Schur, who used rather different arguments [Schu].

3.5.5. The highest weight theory (cf. Theorem 3.5.3.14) and the Weyl character formula (3.5.4.24) are the main constituents of our understanding of representations of compact Lie groups. Both were in place by the mid-1920s, well before the invention of the orbit method. However, both have been given interpretations consistent with the orbit picture. Even these interpretations, which date mainly from the 1950s, preceded the formulation of the orbit picture, and they provided guidance for the development of the representation theory of noncompact semisimple groups. In this section we will discuss the realization of representations by means of cohomology of line bundles over flag varieties. This is often called the Borel-Weil theory, but its full articulation is due to Borel-Weil [Serr3], Bott [Bott], and Kostant [Kost4].

An essential aspect of the BWBK theory is the double interpretation of flag manifolds as homogeneous spaces, either for compact groups, or for their complexifications. We will describe this in general terms and illustrate it for the special unitary group  $SU_n$ .

We will discuss the BWBK theory for a connected, simply-connected, semisimple compact Lie group  $K$ . This is the essential case; allowing  $K$  to be disconnected, non-simply-connected, or to be nonsemisimple (i.e., to have a positive-dimensional center) has mainly nuisance value: it complicates the discussion without requiring any essential ideas. It is for this reason that we use  $SU_n$  rather than  $U_n$  for our example.

Let  $K$  be a connected, simply-connected, semisimple compact Lie group with Lie algebra  $\mathfrak{k}$ , let  $\mathfrak{g} = \mathfrak{k}_{\mathbb{C}} \simeq \mathfrak{k} \otimes \mathbb{C}$  be the complexification of  $\mathfrak{k}$ , and let  $G$  be the simply-connected Lie group with Lie algebra  $\mathfrak{g}$ . Since  $\mathfrak{g}$

is complex,  $G$  likewise will carry a complex structure. We may think of  $K$  as the subgroup of  $G$  whose Lie algebra is  $\mathfrak{k} \subseteq \mathfrak{g}$ . If  $K = \mathrm{SU}_n$ , then  $G = \mathrm{SL}_n(\mathbb{C})$ . Let  $T \subseteq K$  be a maximal abelian subgroup (a Cartan subgroup, also called a maximal torus since it is a product of circles). This notation is inconsistent with that of the preceding and following sections; but here we have another use for  $A$ . Let  $\mathfrak{t}$  be the Lie algebra of  $T$ ,  $\mathfrak{a} = \mathfrak{t}_{\mathbb{C}}$  its complexification. For  $K = \mathrm{SU}_n$ , we may take  $\mathfrak{a}$  to be the complex diagonal matrices of trace zero, and  $\mathfrak{t}$  the pure imaginary ones. The algebra  $\mathfrak{a}$  is a Cartan subalgebra of  $\mathfrak{g}$ , and we have the root space decomposition of  $\mathfrak{g}$ , as described in formula (2.8.6). We may make a choice of positive roots, and the corresponding Borel subalgebra  $\mathfrak{b}^+$  (cf. formulas (3.5.3.2)–(3.5.3.5)). Let  $B \subseteq G$  be the connected subgroup whose Lie algebra is  $\mathfrak{b}^+$ . Since  $\mathfrak{b}^+$  is its own normalizer in  $\mathfrak{g}$ ,  $B$  is necessarily closed. For  $K = \mathrm{SU}_n$ , we may take  $B$  to be the complex upper triangular matrices of determinant 1.

The Iwasawa decomposition [Knap2, Wall2] and §A.2.3.5 for  $G$  says that

$$(3.5.5.1a) \quad G = KB, \quad B \cap K = T.$$

For  $G = \mathrm{SL}_n(\mathbb{C})$ , this amounts to the Gram-Schmidt orthonormalization procedure in  $\mathbb{C}^n$ . We also have the factorization

$$(3.5.5.1b) \quad B = AN^+,$$

where  $A$  and  $N^+$  are the connected subgroups of  $G$  whose Lie algebras are  $\mathfrak{a}$  and  $\mathfrak{n}^+$  (cf. formula (3.5.3.5)). The group  $A$  is the complexification of  $T$ ; it is called a Cartan subgroup or maximal torus of  $G$ . Every character  $\chi$  of  $T$  extends in a unique way to a holomorphic character (i.e., a group homomorphism which is holomorphic with respect to the complex structures on  $A$  and  $\mathbb{C}^\times$ ):

$$(3.5.5.2) \quad \chi : A \rightarrow \mathbb{C}^\times.$$

The complexification process described above also establishes, by a process of differentiation and analytic continuation just as discussed above for  $T$ ,  $\mathfrak{t}$ ,  $\mathfrak{a}$ , and  $A$ , bijections among the following sets:

$$\begin{aligned} &\{\text{irreducible unitary representations of } K\} \\ &\leftrightarrow \{\text{irreducible complex representations of } \mathfrak{k}\} \\ &\leftrightarrow \{\text{irreducible complex linear representations of } \mathfrak{g}\} \\ &\leftrightarrow \{\text{irreducible holomorphic representations of } G\}. \end{aligned}$$

By a complex linear representation of  $\mathfrak{g}$  we mean a complex linear homomorphism  $\sigma : \mathfrak{g} \rightarrow \mathrm{End}(V)$  of  $\mathfrak{g}$  into the endomorphisms of some complex vector space  $V$ . Complex linearity of  $\sigma$  guarantees that  $\sigma$  is determined by its restriction to the real form  $\mathfrak{k}$  of  $\mathfrak{g}$ . Similarly, a holomorphic representation  $\sigma : G \rightarrow \mathrm{GL}(V)$  is a representation which is holomorphic as a mapping of complex manifolds. It is easy to check that the representation  $\sigma$  of  $G$

is holomorphic if and only if the associated representation of  $\mathfrak{g}$  is complex linear.

Given a character  $\chi$  of  $T$ , consider the induced representation  $C_c^\infty(T \backslash K; \chi)$  (cf. §A.1.14). (Since  $K$  is compact, the subscript  $c$  in  $C_c^\infty$  is superfluous.) There is a geometric interpretation of  $C^\infty(T \backslash K; \chi)$  in terms of line bundles [FeDo, Huse, GrHa]; we will review it. The quotient mapping

$$(3.5.5.3) \quad \begin{array}{c} K \\ \downarrow \\ T \backslash K \end{array}$$

can be thought of as a principal fiber bundle [Huse] with fiber  $T$ . Given a representation  $\rho$  of  $T$  on a space  $V$ , we can form the associated vector bundle  $V \times_\rho K$ . If  $\rho = \chi$  is one-dimensional, then we simply have a line bundle. Comparison of the definition of  $V \times_\rho K$  with the definition of induced representation shows that the functions in  $C^\infty(T \backslash K; \rho)$  may be thought of as sections of the vector bundle  $V \times_\rho K$ .

The decomposition (3.5.5.1a) shows that

$$(3.5.5.4) \quad T \backslash K \simeq B \backslash G.$$

Since  $B \backslash G$ , being a quotient space of complex groups, is a complex manifold, we may use identification (3.5.5.4) to think of  $T \backslash K$  as a complex manifold. In the case  $G = \mathrm{SL}_n(\mathbb{C})$ , it is the set of all “complete flags” in  $\mathbb{C}^n$ : sequences of nested spaces  $\{D\} = V_0 \subseteq V_1 \subseteq \cdots \subseteq V_n = \mathbb{C}^n$ , with  $\dim V_j = j$ . In general,  $T \backslash K$  is called the (complete) *flag variety* of  $G$ . The action by right translations of  $K$  on  $T \backslash K$  extends holomorphically to an action of  $G$ . Further, given a character  $\chi \in \widehat{T}$ , we may extend  $\chi$  holomorphically to  $A$ , then to a character of  $B$  trivial on  $N^+$ . Having done that, we may consider the induced representation  $C^\infty(B \backslash G; \chi)$  (see §A.1.14). Decomposition (3.5.5.1a) then shows that by restricting elements of  $C^\infty(B \backslash G; \chi)$  to  $K$ , we obtain an isomorphism

$$(3.5.5.5) \quad C^\infty(B \backslash G; \chi) \simeq C^\infty(T \backslash K; \chi).$$

On the other hand, the line bundle  $\mathbb{C} \times_\chi G$  is a holomorphic line bundle over  $B \backslash G$ . Denote it by  $L_\chi$ . In these circumstances, it is natural to look at the space  $\Gamma(B \backslash G; L_\chi) = H^0(B \backslash G; L_\chi)$  of *holomorphic* sections of  $L_\chi$ ; and more generally, one can consider the (Dolbeault or, equivalently, sheaf) cohomology groups  $H^p(B \backslash G; L_\chi)$  [GrHa, Hart]. Since  $G$  acts holomorphically on  $L_\chi$ , the spaces  $H^p(B \backslash G; L_\chi)$  will all be  $G$ -modules. Note that  $H^0(B \backslash G; L_\chi)$  is a subspace of  $C^\infty(B \backslash G; \chi)$ —the kernel of the  $\bar{\partial}$  operator; however, the higher cohomology groups are not subspaces of  $C^\infty(B \backslash G; \chi)$ . The BWBK theory describes the spaces  $H^p(B \backslash G; L_\chi)$  as  $G$ -modules, and relates this to Lie algebra cohomology.

The group  $\widehat{T}$  of characters of  $T$  is a lattice, isomorphic to  $\mathbf{Z}^r$ ,  $r = \dim T$ . Given  $\chi \in \widehat{T}$ , let  $D\chi$  be the derivative of  $\chi$  at the identity. Then  $D\chi \in \mathfrak{t}^* \subseteq \mathfrak{a}^*$ , and the map  $\chi \rightarrow D\chi$  identifies  $\widehat{T}$  with a lattice inside  $\mathfrak{t}_{\mathbf{C}}^* = \mathfrak{a}^*$ . We will pass back and forth between  $\chi$  and  $D\chi$  without comment. Holomorphic extension to  $A$  further identifies  $\widehat{T}$  to a group of quasicharacters (i.e., homomorphisms into  $\mathbf{C}^\times$ ) of  $A$ . The differentials of these quasicharacters are elements of  $\mathfrak{a}^*$ , the same elements  $D\chi$ ,  $\chi \in \widehat{T}$ , previously obtained. Denote by  $\mathfrak{a}_{\mathbf{R}}^*$  the real linear span of the lattice of  $D\chi$ . Then  $\mathfrak{a}_{\mathbf{R}}^*$  is a real form of  $\mathfrak{a}$ , that is, we have the decomposition

$$(3.5.5.6) \quad \mathfrak{a}^* = \mathfrak{a}_{\mathbf{R}}^* \oplus i\mathfrak{a}_{\mathbf{R}}^*$$

of  $\mathfrak{a}^*$  as a real vector space. In the corresponding decomposition

$$(3.5.5.7) \quad \mathfrak{a} = \mathfrak{a}_{\mathbf{R}} \oplus i\mathfrak{a}_{\mathbf{R}}$$

we have  $\mathfrak{t} = i\mathfrak{a}_{\mathbf{R}}$ ; that is, the elements of  $\mathfrak{a}_{\mathbf{R}}$  are purely imaginary on  $\mathfrak{t}$ . For  $K = \mathrm{SU}_n$ , the Lie algebra  $\mathfrak{a}$  consists of complex diagonal matrices of trace zero,  $\mathfrak{a}_{\mathbf{R}}$  is the subspace of real diagonal matrices, and  $\mathfrak{t}$  is the space of purely imaginary diagonal matrices.

The Weyl group of  $A$  is the normalizer of  $A$  in  $G$ , modulo  $A$ :

$$(3.5.5.8) \quad W \simeq N(A)/A,$$

where  $N(A)$  is the normalizer of  $A$  in  $G$ . We may also describe  $W$  as the normalizer of  $T$  in  $K$ , modulo  $T$ :

$$(3.5.5.9) \quad W \simeq N(T)/T.$$

The group  $W$  acts on  $A$ , hence on  $\mathfrak{a}$ , by pullback via the exponential map—this action is via linear transformations. By duality,  $W$  acts on  $\mathfrak{a}^*$ . Under these related actions,  $T$ ,  $\mathfrak{t}$ ,  $\mathfrak{a}_{\mathbf{R}}$ ,  $\mathfrak{a}_{\mathbf{R}}^*$ , and the lattice  $\widehat{T} \subseteq \mathfrak{a}_{\mathbf{R}}^*$  are all preserved by  $W$ . The action of  $W$  on  $\mathfrak{t}$  or  $\mathfrak{a}_{\mathbf{R}}^*$  is generated by reflections in hyperplanes—these reflections are the elements of  $W$  contained in the copies of  $\mathrm{SL}_2$  generated by root subspaces  $\mathfrak{g}_\alpha$ ,  $\mathfrak{g}_{-\alpha}$ , as described in §2. Also as described there, the reflection hyperplanes divide  $\mathfrak{a}_{\mathbf{R}}$  and  $\mathfrak{a}_{\mathbf{R}}^*$  into convex cones, the Weyl chambers, which are permuted simply transitively by  $W$ . The positive Weyl chamber in  $\mathfrak{a}_{\mathbf{R}}$ , relative to  $\mathfrak{b}^+$ , or  $B$  is

$$(3.5.5.10) \quad \mathfrak{a}_{\mathbf{R}}^+ = \{a \in \mathfrak{a} : \alpha(a) \geq 0, \text{ all } \mathfrak{g}_\alpha \in \mathfrak{b}^+\},$$

and the corresponding positive chamber in  $\mathfrak{a}_{\mathbf{R}}^*$  is

$$(3.5.5.11) \quad (\mathfrak{a}_{\mathbf{R}}^*)^+ = \{\lambda \geq 0 \text{ on } \mathfrak{a}_{\mathbf{R}}^+\}.$$

For  $K = \mathrm{SU}_n$ , the Weyl group is  $S_n$ , the symmetric group, which acts by permuting the diagonal entries of elements of  $\mathfrak{a}_{\mathbf{R}}$ , which consists of traceless real diagonal matrices, and  $\mathfrak{a}_{\mathbf{R}}^+$  is the cone in  $\mathfrak{a}_{\mathbf{R}}$  consisting of matrices whose diagonal entries  $a_i$  decrease with  $i$ :  $a_i \geq a_{i+1}$ .

Write

$$(3.5.5.12) \quad \widehat{T}^+ = \widehat{T} \cap (\mathfrak{a}_{\mathbf{R}}^*)^+.$$

Then  $\widehat{T}^+$  consists of the dominant characters, or dominant integral weights—the highest weights of finite-dimensional representations, as described in §3.5.3.

The result of Borel and Weil describes the cohomology of line bundles defined by inverses of dominant characters [Serr, Warn, Knap2].

**THEOREM 3.5.5.13 (Borel-Weil).** (a) *Let  $\chi^{-1} \in \widehat{T}^+$  be a dominant character. (We then say  $\chi$  is antidominant.) Then the space  $H^0(B \backslash G; L_\chi)$  of global holomorphic sections of the line bundle  $L_\chi$  over  $B \backslash G$  defined by  $\chi$  is an irreducible  $G$ - (or  $K$ -) module, isomorphic to the dual of the representation whose highest weight is  $\chi^{-1}$ .*

(b) *For  $p > 0$ ,  $H^p(B \backslash G; L_\chi) = 0$ .*

**REMARKS.** (a) Theorem 3.5.5.13 is connected to the orbit method through a double interpretation of the complexification of the Lie algebra of  $\mathbf{k}$ : one as the (real) Lie algebra of the complexified group  $G$ , i.e., *real* right-invariant vector fields on  $G$ , and one as *complex* right-invariant vector fields on  $K$ . For functions which are holomorphic on  $G$ , these two interpretations coincide. Thus, for a holomorphic section  $f$  of  $L_\chi$ , left invariance of  $f$  by  $N^+$ , as a function on  $G$ , can be interpreted in terms of  $f|_K$  as a condition of being annihilated by the complex vector fields on  $K$  defined by  $\mathfrak{n}^+ \subseteq \mathfrak{k}_{\mathbf{C}}$ . Put another way, a function in  $C^\infty(T \backslash K; \chi)$  will extend to a *holomorphic*  $N^+$ -left-invariant function on  $G$  if and only if it is annihilated by the vector fields from  $\mathfrak{n}^+ \subseteq \mathfrak{k}_{\mathbf{C}}$ , which may be seen to define a system of Cauchy-Riemann type equations. Thus, interpreted on  $K$ , the holomorphy condition becomes a condition of being an eigenfunction for the algebra  $\mathfrak{b}^+ = \mathfrak{n}^+ + \mathfrak{a} \subseteq \mathfrak{k}_{\mathbf{C}}$ . The algebra  $\mathfrak{b}^+$  is seen to be a complex polarization for the element  $i\chi \in \mathfrak{t}^* \subseteq \mathfrak{k}^*$ , and this use of complex polarizations is closely analogous to the way they are used in [AuKo] to produce representations of solvable groups. Although the Auslander-Kostant construction can be replaced by a construction involving only the real group, but inducing from representations of Heisenberg groups, not just from characters, there does not seem to be any escape from infinitesimal constructions involving complex polarizations in the case of semisimple groups.

(b) The presence of inverses and duals in this result makes it somewhat confusing. Perhaps the quickest way to verify the proper formulation is to consider the element of  $H^0(B \backslash G; L_\chi)^*$  defined by the Dirac  $\delta$  at the identity in  $G$ . If  $\rho$  denotes the action of  $G$  on  $H^0(B \backslash G; L_\chi)$  by right translations,

then

$$\begin{aligned}\rho^*(b)(\delta)(f) &= \delta(\rho(b)^{-1}f) \\ &= \rho(b)^{-1}(f)(1) = f(b^{-1}) = \chi^{-1}(b)f(1) = \chi^{-1}(b)\delta(f), \\ &\quad (b \in B, f \in H^0)\end{aligned}$$

whence

$$\rho^*(b)(\delta) = \chi^{-1}(b)\delta.$$

In other words,  $\delta$  is a highest weight vector for  $H^0(B \backslash G; L_\chi)^*$ , with weight  $\chi^{-1}$ .

Part (a), the positive part of this result, is essentially a restatement of the highest weight theory. The main observation is that  $H^0(B \backslash G; L_\chi)$  can contain at most one  $N^+$ -invariant function. (Indeed, all of  $C^\infty(B \backslash G; \chi)$  contains only one  $N^+$ -invariant function.) This is because  $N^+$  has a dense orbit on  $B \backslash G$ . This follows from the Bruhat decomposition (see §1.1 and [HaCh1, Knap2, Wall2]) which says  $G = BWN^+$  so that, in particular, there are only finitely many  $N^+$  orbits on  $B \backslash G$ , one of which is open and dense; but the fact that there is an open  $N^+$ -orbit in  $B \backslash G$  is more elementary than the Bruhat decomposition.

The fact that  $H^0(B \backslash G; L_\chi)$  consists of holomorphic functions means that the  $G$ -invariant subspaces of  $H^0(B \backslash G; L_\chi)$  are the same as the  $K$ -invariant subspaces—in particular  $H^0(B \backslash G; L_\chi)$  must be a direct sum of irreducible finite-dimensional  $G$ -modules. The theorem of the highest weight means that any  $G$ -irreducible subspace of  $H^0(B \backslash G; L_\chi)$  must contain an  $N^+$ -invariant vector. Hence, there can be at most one subspace. On the other hand, if  $V_\chi$  is the irreducible  $\mathfrak{g}$ -module with highest weight  $-D\chi$  and highest weight vector  $v_\chi$ , then by exponentiating the action of  $\mathfrak{g}$  we obtain an action of  $G$  on  $V_{-D\chi}$ , and the matrix coefficients (cf. §A.1.11)

$$\varphi_{\lambda, v_\chi}(g^{-1}) = \lambda(g^{-1}(v_\chi)) \quad \lambda \in V_{-D\chi}^*$$

define a  $G$ -equivariant embedding of  $V_{-D\chi}^*$  into, hence an isomorphism with,  $H^0(B \backslash G; L_\chi)$ .

The complementary part (b) of Theorem 3.5.5.13 is a consequence of the Kodaira Vanishing Theorem [GrHa, Hart].

Theorem 3.5.5.13 provides a “geometric” realization of the irreducible representations of  $K$  (or  $\mathfrak{g}$ ). However, it also raises an issue that it only partially resolves: although we can form the line bundles  $L_\chi$  over  $B \backslash G$  for all  $\chi$  in  $\widehat{T}$ , Theorem 3.5.5.13 only describes the cohomology groups  $H^p(B \backslash G; L_\chi)$  for  $\chi$  antidominant. The highest weight theory guarantees that  $H^0(B \backslash G; L_\chi) = \{0\}$  if  $\chi$  is not antidominant, but it is silent about

higher cohomology. The structure of the higher cohomology groups was clarified by Bott [Bott]. To state Bott's result (conjectured by Borel and Hirzebruch), we need to introduce the character

$$(3.5.5.14) \quad \delta(a) = \det(\text{Ad}(a)|_{\mathfrak{n}^+}) = \prod_{\alpha > 0} \chi_{\alpha}(g).$$

Here  $\chi_{\alpha}$  is the character of  $A$  whose derivative  $D\chi_{\alpha}$  is equal to the positive root  $\alpha$ . Thus  $\delta$  is the character of  $A$  whose derivative at the identity is the sum of the positive roots. We observe that  $\delta$  is a holomorphic square root of the modular function of  $B$ ,

$$(3.5.5.15) \quad d(\text{Ad}(a)n) = |\delta(a)|^2 dn$$

if  $dn$  is Haar measure on  $N^+$ . It is a somewhat subtle point in the structure theory of compact groups that, under our assumption that  $K$  is simply connected,  $\delta$  itself has a square root in  $\hat{T}$ ; we will denote this by  $\delta^{1/2}$ . It is not hard to check that  $\delta$  and hence  $\delta^{1/2}$  is dominant. For example, if  $G = \text{SL}_n(\mathbb{C})$ , and we use the usual diagonal coordinates  $\{a_i\}$  on  $A$ , then

$$(3.5.5.16a) \quad \begin{aligned} \delta(a_1, \dots, a_n) &= \prod_{i < j} (a_i a_j^{-1}) = \prod_j a_j^{n+1-2j} \\ &= \left( \prod_j a_j \right)^{n+1} \left( \prod_j a_j^{-j} \right)^2. \end{aligned}$$

Thus for  $\text{SL}_n$ , we have

$$(3.5.5.16b) \quad \delta^{1/2}(a_1, \dots, a_n) = \prod_j a_j^{-j} = \prod_j a_j^{n-j}.$$

Note that  $\delta^{1/2}$  is essentially identical with the  $\chi_p$  used in §3.5.4 (cf. formula (3.5.4.18)).

**THEOREM 3.5.5.17 (Bott).** *Consider  $\chi \in \hat{T}$ , and form the associated holomorphic line bundle  $L_{\chi}$  over  $B \backslash G$ .*

(a) *If  $\chi \delta^{-1/2}$  is singular (i.e., fixed by a nontrivial element of  $W$ ), then  $H^p(B \backslash G; L_{\chi}) = 0$ .*

(b) *If  $\chi \delta^{-1/2}$  is not singular, then there is a unique  $w$  in  $W$  such that  $w(\chi^{-1} \delta^{1/2})$  is dominant, i.e.,  $w(\chi \delta^{-1/2})$  is antidominant. In this case, set*

$$(3.5.5.18) \quad \psi = w(\chi^{-1} \delta^{1/2}) \delta^{-1/2}.$$

*Let  $l(w) = l$  be the length of  $w$  as an element of  $W$  (cf. [Hilr, Bour]). Then  $H^p(B \backslash G; L_{\chi}) = 0$  for  $p \neq l$ , and*

$$(3.5.5.19) \quad H^l(B \backslash G; L_{\chi}) \simeq (V_{\psi})^*,$$

where  $V_\psi$  is the irreducible representation of  $G$  with highest weight  $\psi$ , with  $\psi$  given by (3.5.5.18).

Bott's proof of Theorem 3.5.5.17 used spectral sequences. However, Bott noted that, by some elementary yoga in sheaf cohomology, this theorem is equivalent to a statement about the Lie algebra cohomology of the Lie algebra  $\mathfrak{n}^+$  of  $N^+$  with coefficients in a  $\mathfrak{g}$ -module  $V_\psi$ ,  $\psi \in \widehat{T}^+$ . Since the Cartan subgroup  $A$  acts on  $\mathfrak{n}^+$  by automorphisms, one sees from the standard construction [BoWa, Jaco1, Knap1] of Lie algebra cohomology that each cohomology group  $H^p(\mathfrak{n}^+, V_\psi)$  naturally has the structure of an  $A$ -module. Kostant [Kost4] gave a direct explicit description of  $H^p(\mathfrak{n}^+, V_\psi)$  as an  $A$ -module, obtaining Bott's result as a corollary. To state Kostant's Theorem, we introduce the notation  $C_\chi$  for the one-dimensional irreducible representation of  $A$  whose associated character is  $\chi$ .

**THEOREM 3.5.5.20 (Kostant).** *Let  $\psi \in \widehat{T}^+$  be a dominant character of  $A$ , and let  $V_\psi$  be the associated finite-dimensional irreducible representation. Then there is an  $A$ -module isomorphism*

$$(3.5.5.21) \quad H^q(\mathfrak{n}^+, V_\psi) \simeq \sum_{l(w)=q} C_{\tilde{w}(\psi)},$$

where  $\tilde{w}(\psi) = w(\psi\delta^{1/2})\delta^{-1/2}$ .

The essential, and originally the most difficult, part of the proof is to show that only the  $C_{\tilde{w}(\psi)}$  can appear in the  $H^q(\mathfrak{n}^+, V_\psi)$ . (Aribaud [Arib] gave a simplified argument based on the Weyl character formula.) This is now understood to be an aspect of the Harish-Chandra homomorphism [Hump, Knap2, Wall2], which also accounts for the " $\rho$ -shifts" in the  $\tilde{w}(\psi)$ . This basic result gives a precise description of the center of the universal enveloping algebra of  $\mathfrak{k}$ , or  $\mathfrak{g}$ . The direct sum decomposition (cf. (3.5.3.4))

$$\mathfrak{g} = \mathfrak{n}^+ \oplus \mathfrak{a} \oplus \mathfrak{n}^+$$

of  $\mathfrak{g}$  leads via the Poincaré-Birkhoff-Witt Theorem (cf. [Hump, Jaco1, Serr2], etc.) to the decompositions

$$(3.5.5.22) \quad \begin{aligned} \mathcal{U}(\mathfrak{g}) &\simeq \mathcal{U}(\mathfrak{n}^-) \otimes \mathcal{U}(\mathfrak{a}) \otimes \mathcal{U}(\mathfrak{n}^+) \\ &\simeq \mathcal{U}(\mathfrak{n}^-) \otimes \mathcal{U}(\mathfrak{a}) \oplus \mathcal{U}(\mathfrak{g})\mathfrak{n}^+. \end{aligned}$$

For  $u \in \mathcal{U}(\mathfrak{g})$ , denote by  $p(u)$  the component of  $u$  in  $\mathcal{U}(\mathfrak{n}^-) \otimes \mathcal{U}(\mathfrak{a})$ , the first summand of decomposition (3.5.5.22).

**THEOREM 3.5.5.23 (Harish-Chandra homomorphism).** (a) *If  $u \in \mathcal{U}(\mathfrak{g})^{\text{Ad } A}$ , the subalgebra of  $\mathcal{U}(\mathfrak{g})$  of elements invariant under  $\text{Ad } A$ , then  $p(u) \in \mathcal{U}(\mathfrak{a})$ .*

(b) *The mapping  $u \rightarrow p(u)$  defines an algebra homomorphism from  $\mathcal{U}(\mathfrak{g})^{\text{Ad } A}$  to  $\mathcal{U}(\mathfrak{a})$ .*

(c) *If we make the identifications*

$$\mathcal{U}(\mathfrak{a}) \simeq S(\mathfrak{a}) \simeq \mathcal{P}(\mathfrak{a}^*),$$



where  $S(\mathfrak{a})$  denotes the symmetric algebra on  $\mathfrak{a}$  and  $\mathcal{P}(\mathfrak{a}^*)$  the algebra of polynomials on  $\mathfrak{a}^*$ , then the map

$$(3.5.5.24(i)) \quad \tilde{p} : \mathcal{Z}\mathcal{U}(\mathfrak{g}) \rightarrow \mathcal{P}(\mathfrak{a}^*)$$

defined by

$$(ii) \quad \tilde{p}(u)(\lambda) = p(u)(\lambda - \rho), \quad \lambda \in \mathfrak{a}^*,$$

where

$$(iii) \quad 2\rho = D_\delta,$$

i.e.,  $\rho$  is  $\frac{1}{2}$  of the sum of the positive roots, is an isomorphism

$$(3.5.5.25) \quad \tilde{p} : \mathcal{Z}\mathcal{U}(\mathfrak{g}) \simeq \mathcal{P}(\mathfrak{a}^*)^W$$

from the center of  $\mathcal{U}(\mathfrak{g})$  to the algebra of Weyl group invariant functions on  $\mathfrak{a}^*$ .

Parts (a) and (b) of this theorem are proved by easy computations, while part (c) may be seen using the Verma module approach to the highest weight theory, as described in §3.5.3.

In an illustration, and in some sense the crucial case, of Theorem 3.5.5.23, we recall the formula

$$\begin{aligned} \mathcal{E} &= h^2 + 2(e^+e^- + e^-e^+) = h^2 + 2h + 4e^-e^+ \\ &= (h+1)^2 - 1 + 4e^-e^+ \end{aligned}$$

for the Casimir element in  $\mathcal{U}(\mathfrak{sl}_2)$  (cf. §3.5.1). Note that  $1 = (\frac{1}{2})\alpha^+(h)$ , where  $\alpha^+$  is the positive root in  $\mathfrak{sl}_2$ , since  $[h, e^+] = \alpha^+(h)e^+ = 2e^+$ .

The Harish-Chandra homomorphism impinges on Theorem 3.5.5.20 as follows. Given a representation  $\rho$  of  $\mathfrak{g}$  on a vector space  $V$ , the images  $\rho(z)$ ,  $z \in \mathcal{Z}\mathcal{U}(\mathfrak{g})$ , are operators which commute with  $\rho(x)$ ,  $x \in \mathfrak{g}$ , and in particular with  $\rho(\mathfrak{n}^+)$ . It follows from the standard construction [Jaco1, BoWa, Knap1] of Lie algebra cohomology that  $\rho(\mathcal{Z}\mathcal{U}(\mathfrak{g}))$  will induce operators on the cohomology groups  $H^q(\mathfrak{n}^+, V)$ . Thus the  $\mathfrak{n}^+$  cohomology of a  $\mathfrak{g}$ -module may be regarded as a joint  $\mathcal{Z}\mathcal{U}(\mathfrak{g})$  and  $\mathcal{U}(\mathfrak{a})$ -module. (The  $\mathcal{U}(\mathfrak{a})$ -module structure is of course obtained as the infinitesimal version of the action of  $A$ .) Denote this action by  $\rho^+$ .

**THEOREM 3.5.5.26** (Casselman-Osborne [CaOs, Knap1, Voga2]). *The action  $\rho^+$  of  $\mathcal{Z}\mathcal{U}(\mathfrak{g})$  on  $H^*(\mathfrak{n}^+, V)$  factors through the Harish-Chandra homomorphism:*

$$(3.5.5.27) \quad \rho^+(u) = \rho^+(\tilde{p}(u))$$

with  $\tilde{p}$  as in formula (3.5.5.25).

This result follows from the general machinery of cohomology, if one observes that the standard resolution of  $V$  as a  $\mathfrak{g}$ -module [Jaco1, BoWa, Knap1] is also a resolution of  $V$  as an  $\mathfrak{n}^+$ -module, since  $\mathcal{U}(\mathfrak{g})$  is free as

a module over  $\mathcal{U}(\mathfrak{n}^+)$  by Poincaré-Birkhoff-Witt [Hump, Jaco1, Serr2]. Its relevance for Theorem 3.5.5.20 is that, if  $V$  is irreducible, then  $\rho(\mathcal{Z}\mathcal{U}(\mathfrak{g}))$  consists of scalars, and via  $\rho^+$  will obviously act by the same scalars. Thus formula (3.5.5.27) constrains the action of  $\mathcal{U}(\mathfrak{a}^*)$ . Indeed, it immediately implies that the only characters of  $\mathcal{A}$  which could possibly appear in formula (3.5.5.21) are the ones which do. As mentioned above, this is the essential step in the proof of Theorem 3.5.5.20, which in turn implies the “geometric realization” Theorems 3.5.5.13 and 3.5.5.17.

3.5.6. In this subsection we complete the geometric quantization version of the basic representation theory of compact groups by giving Harish-Chandra’s orbit method interpretation of the Weyl character formula [HaCh3]. With hindsight one can see in this remarkable paper the seeds of a large fraction of nonabelian harmonic analysis as it has developed in the ensuing 30 years. Besides Harish-Chandra’s own work on the construction of the discrete series, it foreshadows the whole orbit method and also implicitly uses the oscillator representation [Foll1, Howe3, Shal, Weil1]. Our account will make this last connection explicit. (The first explicit use of the connection is [Verg2]).

As in §3.5.4, we will present only the example of the unitary group  $U_n$ , to save notation and preparation. The Lie algebra  $\mathfrak{u}_n$  of  $U_n$  is the space of skew-adjoint  $n \times n$  complex matrices. To be definite we recall

$$(3.5.6.1) \quad \mathfrak{u}_n = \{T \in M_n(\mathbb{C}) : T = [t_{jk}]; t_{kj} = -\bar{t}_{jk}\},$$

where the overbar denotes complex conjugation, and  $\{t_{jk}\}$ ,  $1 \leq j, k \leq n$ , are the entries of the  $n \times n$  matrix  $T$ . The unitary group  $U_n$  acts on  $\mathfrak{u}_n$  by conjugation. As usual, we denote this action by  $\text{Ad}$ :

$$\text{Ad } g(T) = gTg^{-1}, \quad T \in \mathfrak{u}_n, g \in U_n.$$

On  $\mathfrak{u}_n$  we can define a positive definite inner product  $(\ , \ )$  by the formula

$$(3.5.6.2) \quad \begin{aligned} (S, T) &= \text{trace}(ST^*) = -\text{trace}(ST) \quad (S, T \in \mathfrak{u}_n) \\ &= -\sum_{1 \leq j, k \leq n} s_{jk} \bar{t}_{jk} \\ &= -\sum s_{jj} t_{jj} + 2 \sum_{1 \leq j < k \leq n} (\text{Re } s_{jk} \text{Re } t_{jk} + \text{Im } s_{jk} \text{Im } t_{jk}). \end{aligned}$$

This inner product is easily seen to be invariant under  $\text{Ad } U_n$ . Using  $(\ , \ )$  we can define a Fourier transform on functions on  $\mathfrak{u}_n$  by one of the usual recipes

$$(3.5.6.3) \quad \hat{f}(S) = \int_{\mathfrak{u}_n} f(T) e^{-2\pi i(S, T)} dT.$$

Here  $dT$  is Lebesgue measure defined by coordinates with respect to any orthonormal basis for  $(\ , \ )$ . For example, we could take the coordinates  $it_{jj}$ ,  $2^{-1/2} \text{Re } t_{jk}$ , and  $2^{-1/2} \text{Im } t_{jk}$ ,  $1 \leq j < k \leq n$ . With this normalization of Lebesgue measure, the Fourier transform is unitary.

Let  $\mathfrak{a} \subseteq \mathfrak{u}_n$  be the subspace of diagonal matrices

$$(3.5.6.4) \quad \mathfrak{a} = \left\{ \begin{bmatrix} ib_{11} & & & \\ & ib_{22} & & \\ & & \ddots & \\ 0 & & & ib_{nn} \end{bmatrix} : b_{jj} \in \mathbf{R} \right\}.$$

We denote the typical element of  $\mathfrak{a}$  by  $B$ , and the typical entries of  $B$  will be  $ib_{jj}$ . The restriction of the inner product  $(\ , \ )$  of formula (3.5.6.2) defines an inner product on  $\mathfrak{a}$ ; in fact, it is just the standard Euclidean inner product with respect to the coordinates  $b_{jj}$ . The orthogonal complement  $\mathfrak{a}^\perp$  of  $\mathfrak{a}$  with respect to  $(\ , \ )$  is the space of skew-adjoint matrices with zeros on the diagonal. We can define the Fourier transform for functions on  $\mathfrak{a}$  by an analog of formula (3.5.6.3).

We know by spectral theory for self-adjoint matrices [Lang3] that every  $T$  in  $\mathfrak{u}_n$  is conjugate by  $U_n$  to an element of  $\mathfrak{a}$ . Thus we have a surjective mapping

$$(3.5.6.5) \quad \begin{aligned} \gamma : \mathfrak{a} \times U_n &\rightarrow \mathfrak{u}_n, \\ \gamma(B, g) &= gBg^{-1}, \quad B \in \mathfrak{a}, \ g \in U_n. \end{aligned}$$

This map is an infinitesimal analog of the map  $\Gamma$  of formula (3.5.4.3), and it has a basic theory parallel to the theory for  $\Gamma$ . First, it factors to a “polar coordinates” map

$$\tilde{\gamma} : \mathfrak{a} \times (U_n/A) \rightarrow \mathfrak{u}_n$$

which is generically  $n!$ -to-one. Second, there is a polar-coordinates integration formula analogous to formula (3.5.4.21) :

$$(3.5.6.6a) \quad \int_{U_n} \varphi_1(T) \overline{\varphi_2(T)} dT = c_o \int_{\mathfrak{a}} \varphi_1(B) \overline{\varphi_2(B)} dB.$$

Here  $c_o$  is an appropriate constant and  $D$  is, as before, the discriminant function

$$(3.5.6.6b) \quad D(B) = \prod_{1 \leq i < j \leq n} (b_{ii} - b_{jj}), \quad B \in \mathfrak{a}.$$

The constant  $c_o$  can be determined explicitly (see [HaCh3, HoTa]). As opposed to the situation in §3.5.4, here  $D(B)$  is not to be thought of as a sum of characters, but as a polynomial function. Its structural interpretation is that it is the product of the positive roots for  $\mathfrak{a}$ . The proof of formula (3.5.6.7) is parallel to that for (3.5.4.21). In particular the calculation of the volume  $|D^2(B)|$  for the orbit  $\text{Ad } U_n(B)$  is a Jacobian determinant computation slightly simpler than but quite similar to the volume factor  $\nu(a)$ . See the discussion preceding formula (3.5.4.5).

Also in parallel to §3.5.4, we may define the spaces  $L^2(\mathfrak{u}_n)^{\text{Ad } U_n}$  of conjugation invariant  $L^2$  functions on  $\mathfrak{u}_n$ , and  $L^2(\mathfrak{a})^{\mathcal{W}, \text{sgn}}$  of skew-symmetric functions on  $\mathfrak{a}$ . We may define a map

$$(3.5.6.7) \quad M_D \circ \text{res}_{\mathfrak{a}} : L^2(\mathfrak{u}_n)^{\text{Ad } U_n} \rightarrow L^2(\mathfrak{a})^{\mathcal{W}, \text{sgn}}$$

with notation parallel to statement (3.5.4.12), and it will again be true that this map (multiplied by  $c_o^{1/2}$ , with  $c_o$  as in (3.5.6.6a)) is a unitary isomorphism.

Since conjugation by  $U_n$  preserves the inner product  $(\ , \ )$ , it will commute with the Fourier transform. Consequently, the space  $L^2(\mathfrak{u}_n)^{\text{Ad } U_n}$  will be invariant under the Fourier transform on  $\mathfrak{u}_n$ . Similarly  $L^2(\mathfrak{a})^{\mathcal{W}, \text{sgn}}$  will be invariant under the Fourier transform on  $\mathfrak{a}$ . Since we will now be considering the Fourier transform on  $\mathfrak{u}_n$  and on  $\mathfrak{a}$  at the same time, we will use the notations  $\wedge^{\mathfrak{u}}$  and  $\wedge^{\mathfrak{a}}$  respectively for them in order to be definite about which one is meant.

Harish-Chandra's discovery about the map  $M_D \circ \text{res}_{\mathfrak{a}}$  was that it intertwines the two Fourier transforms.

**THEOREM 3.5.6.8 (Harish-Chandra Restriction Theorem).** *The mapping (3.5.6.7) satisfies*

$$M_D \circ \text{res}_{\mathfrak{a}} \circ \wedge^{\mathfrak{u}} = i^{-n(n-1)/2} \wedge^{\mathfrak{a}} \circ M_D \circ \text{res}_{\mathfrak{a}}.$$

*In other words,*

$$D(B)(\varphi^{\wedge^{\mathfrak{u}}})(B) = i^{-n(n-1)/2} (D\varphi|_{\mathfrak{a}})^{\wedge^{\mathfrak{a}}}(B).$$

Since the Fourier transform is a nonlocal operator, a result like Theorem 3.5.6.8 is quite surprising. We will see shortly how special the circumstances are which give rise to this phenomenon.

To appreciate the structure underlying the Harish-Chandra Restriction Theorem, consider the Laplace operator on  $\mathfrak{u}$  dual to the inner product (3.5.6.2). It is the second-order, constant coefficient operator  $\Delta$ , or  $\Delta_{\mathfrak{u}}$  when more specificity is needed, given by the formula

$$(3.5.6.9) \quad \Delta = \Delta_{\mathfrak{u}} = \sum_{j=1}^n \frac{\partial^2}{\partial s_{jj}^2} + \frac{1}{2} \sum_{1 \leq j < k \leq n} \left( \frac{\partial^2}{\partial r_{jk}^2} + \frac{\partial^2}{\partial s_{jk}^2} \right),$$

where we take  $t_{jk} = r_{jk} + is_{jk}$ , i.e.,  $r_{jk}$  and  $s_{jk}$  are respectively the real and imaginary parts of  $t_{jk}$ . The factor  $\frac{1}{2}$  occurs because, as noted above (see formulas (3.5.6.2) to (3.5.6.4)), the coordinates for which  $(\ , \ )$  looks like the standard Euclidean inner product are  $s_{jj}$ ,  $2^{-1/2} r_{jk}$  for  $j < k$ , and  $2^{-1/2} s_{jk}$  for  $j < k$ . The operator  $\Delta_{\mathfrak{u}}$  is the standard Laplacian with respect to these coordinates.

Let us write

$$(3.5.6.10) \quad r^2(T) = (T, T).$$

Also let  $r^2$  denote the operation of multiplication by  $r^2$ . It is easy to compute the commutator

$$(3.5.6.11) \quad [\Delta, r^2] = 4 \left( \sum_{j=1}^n s_{jj} \frac{\partial}{\partial s_{jj}} + \sum_{1 \leq j < k \leq n} r_{jk} \frac{\partial}{\partial r_{jk}} + s_{jk} \frac{\partial}{\partial s_{jk}} \right) + 2n^2 \\ = 4E + 2n^2,$$

where  $E$  is the standard Euler degree operator on  $\mathbf{u}_n$ , which multiplies polynomials of degree  $m$  by  $m$ .

For  $\mathbf{R}^n$ , consider the operators

$$(3.5.6.12) \quad e^+ = \pi i r^2, \quad e^- = \frac{i\Delta}{4\pi}, \quad h = E + \left(\frac{1}{2}\right) n^2.$$

Analogous operators may be defined for any space endowed with an inner product, as we have done above for  $\mathbf{u}_n$ , and the statements below will hold also for such spaces. Using formula (3.5.6.11) and some other simple calculations, we can check that  $e^\pm$  and  $h$  form a standard basis for a copy of the Lie algebra  $\mathfrak{sl}_2$ .

**THEOREM 3.5.6.13 (Shale [Shal]).** *There is a unique representation  $\omega$  of  $\widetilde{\mathrm{SL}}_2(\mathbf{R})$ , the two-fold cover of  $\mathrm{SL}_2(\mathbf{R})$ , on  $L^2(\mathbf{R}^n)$  such that the image of the associated representation of  $\mathfrak{sl}_2$  (see §A.1.13) is the span of the operators (3.5.6.12).*

**REMARKS.** (a) The operators (3.5.6.12) are a Lie subalgebra of the Lie algebra of all polynomial-coefficient differentials of total order (= polynomial degree + order of differentiation) two on  $\mathbf{R}^n$ . These operators are the span of

$$(3.5.6.14) \quad \pi i x_j x_k \frac{1}{2} \left( x_j \frac{\partial}{\partial x_k} + \frac{\partial}{\partial x_k} x_j \right) = x_j \frac{\partial}{\partial x_k} + \frac{\delta_{jk}}{2} \frac{i}{4\pi} \frac{\partial^2}{\partial x_j \partial x_k}.$$

This algebra is isomorphic to the symplectic Lie algebra in  $2n$  variables, denoted  $\mathfrak{sp}_{2n}$ . Shale actually showed there is a unitary representation of  $\widetilde{\mathrm{Sp}}_{2n}(\mathbf{R})$ , the two-fold cover of the real symplectic group in  $2n$  variables, such that the image of the associated representation of the Lie algebra is the span of the operators (3.5.6.14).

(b) Shale's interest was quantum field theory. Shortly after Shale, Weil [Weil1], motivated by Segal's work on automorphic forms, independently showed the existence of this representation. Weil also established the existence of an analogous representation for  $\mathrm{Sp}_{2n}(F)$ , the symplectic group in  $2n$  variables with values in a  $p$ -adic field  $F$ . Weil showed that this representation underlies the classical theory of  $\theta$ -series, one of the most widely used means for constructing automorphic forms (cf. [Igus, Shmz1, 2, KuMi1-3, ToWa1-2, Shim, Shin2, Niwa], etc.).

(c) I call this representation the *oscillator representation*, because of its close association with the quantum harmonic oscillator (see §3.1). Other

names in use are the Weil representation, the Segal-Shale-Weil representation, the harmonic representation, etc.

(d) In some sense the oscillator representation is the quintessential example of geometric quantization, and it is derelict not to present its construction in detail. On the other hand, the construction gets rather involved and involves some special ideas, and so would constitute a sizeable digression. Also, I have written quite a bit about it [Howe1–6], and do not wish to repeat myself here. Detailed accounts of it, from the viewpoint of geometric quantization, can be found for example in [Blat, LiVe]. My own account, which takes a somewhat different viewpoint, is [Howe3].

(e) In fact, we do not need the full Theorem 3.5.6.13 for this discussion. We only need to exponentiate the operator  $\frac{\Delta}{4\pi} - \pi r^2$ , which is a multivariable variant of the Hamiltonian for the quantum harmonic oscillator discussed in §3.1. It can be handled by the same techniques. Thus our discussion is more or less complete on this point. However, Theorem 3.5.6.13 seems to identify the natural relevant structure for this situation. This connection was pointed out by Vergne in [Verg2].

The relevance of Theorem 3.5.6.13 to the Harish-Chandra Restriction Theorem is that the Fourier transform is almost an element of  $\omega(\widetilde{\mathrm{SL}}_2(\mathbf{R}))$ . Consider the element

$$(3.5.6.15) \quad \mathbf{k} = e^+ - e^- = \frac{i}{2} \left( 2\pi r^2 - \frac{\Delta}{2\pi} \right)$$

in our copy of  $\mathfrak{sl}_2$ . An easy computation shows that  $\mathbf{k}$  generates the standard maximal compact subgroup  $\mathrm{SO}_2$  inside  $\mathrm{SL}_2$ . On the other hand, from calculations just like those of §3.1, we know the eigenvalues and eigenvectors of  $\mathbf{k}$ . From the standard formulas for the Fourier transform on  $\mathbf{R}^n$ , viz.,

$$(2\pi i x_j f)^\wedge = -\frac{\partial}{\partial x_j}(\hat{f}), \quad \left( \frac{\partial}{\partial x_j} f \right)^\wedge = 2\pi i x_j \hat{f}, \quad (e^{-\pi r^2})^\wedge = e^{-\pi r^2},$$

we can deduce that the eigenvectors for  $\mathbf{k}$  are also eigenvectors for the Fourier transform, and further that the Fourier transform can be written as

$$(3.5.6.16) \quad \wedge = i^{-n/2} \exp(\pi \mathbf{k}/2).$$

See for example [HoTa, Howe3].

REMARK. The scalar factor  $i^{-n/2}$  in equation (3.5.6.16) comes from the fact that the smallest eigenvalue of  $\mathbf{k}$  is  $\frac{1}{2}$  rather than zero. This fact is interpreted in quantum mechanics as the Uncertainty Principle [Shan], and in quantum electrodynamics as the zero-point energy, or energy of the vacuum [Thir]. It also reflects the fact that  $\omega$  is a representation of  $\widetilde{\mathrm{SL}}_2(\mathbf{R})$ , and not of  $\mathrm{SL}_2(\mathbf{R})$ .

In view of formula (3.5.6.16), Theorem 3.5.6.8 follows from

**THEOREM 3.5.6.17.** *The mapping  $M_D \circ \text{res}_{\mathfrak{a}}$  of formula (3.5.6.7) intertwines the restriction of the oscillator representations  $\omega_{\mathfrak{u}}$  and  $\omega_{\mathfrak{a}}$  of  $\widetilde{\text{SL}}_2(\mathbf{R})$  on  $L^2(\mathfrak{u}_n)^{\text{Ad } U_n}$  and  $L^2(\mathfrak{a})^{W, \text{sgn}}$  respectively. It defines a unitary (up to multiples) equivalence of  $\widetilde{\text{SL}}_2(\mathbf{R})$  modules.*

**REMARKS.** We should note that the operators (3.5.6.12) are all invariant under conjugation by orthogonal transformations, and therefore the oscillator representation, whose existence is asserted by Theorem 3.5.6.13, will commute with orthogonal transformations. Since both spaces  $L^2(\mathfrak{u}_n)^{\text{Ad } U_n}$  and  $L^2(\mathfrak{a})^{W, \text{sgn}}$  are defined by how their elements transform under certain orthogonal transformations, each is invariant under the relevant oscillator representation. Hence the assertion of Theorem 3.5.6.17 at least makes sense.

We will sketch a proof of this theorem.

Since in both representations the image of the operator  $\mathbf{k}$  has discrete spectrum with finite-dimensional eigenspaces, as is revealed by the computations of §3.1, easy technical arguments reveal it is enough to show that Theorem 3.5.6.17 is true infinitesimally, i.e., that the map  $M_D \circ \text{res}_{\mathfrak{a}}$  intertwines the operators (3.5.6.12) for  $\mathfrak{u}_n$  with their counterparts for  $\mathfrak{a}$ . To do this for the operator  $e^+$  is trivial: one needs only the facts that restriction is a homomorphism for pointwise multiplication, and that pointwise multiplication of complex-valued functions is commutative. To check it for  $\mathbf{h}$  is also very simple: it uses the fact that  $\mathfrak{a}$  is invariant under scalar multiplication in  $\mathfrak{u}_n$ , and that  $D$  is homogeneous of degree  $\frac{1}{2}(\dim \mathfrak{u}_n - \dim \mathfrak{a})$ .

Thus the crucial calculation is to show that the map  $M_D \circ \text{res}_{\mathfrak{a}}$  intertwines the two Laplacians  $\Delta_{\mathfrak{u}}$  and  $\Delta_{\mathfrak{a}}$ . We would like to perform this calculation in a moderately general context, to illustrate the issues involved. Related calculations are given in [Helg3, Helg4]. See also [HoTa].

Consider  $\mathbf{R}^n \subseteq \mathbf{R}^{n+m}$ . Use coordinates  $x_1, \dots, x_n$  on  $\mathbf{R}^n$ , and let  $y_1, y_2, \dots, y_m$  be the remaining coordinates on  $\mathbf{R}^{n+m}$ . Imagine we are giving a “nonlinear orthogonal projection”

$$(3.5.6.18) \quad \begin{aligned} \Phi: \mathbf{R}^{m+n} &\rightarrow \mathbf{R}^n, \\ \Phi(x, y) &= (\phi_1(x, y), \phi_2(x, y), \dots, \phi_n(x, y)). \end{aligned}$$

Precisely, the points of  $\mathbf{R}^n$  should be fixed by  $\Phi$ , and the fibers  $\Phi^{-1}(x)$ ,  $x \in \mathbf{R}^n$ , should intersect  $\mathbf{R}^n$  orthogonally. In formulas, these conditions are

$$(3.5.6.19a) \quad \Phi(x, 0) = x,$$

$$(3.5.6.19b) \quad \frac{\partial \Phi}{\partial y_j}(x, 0) = 0.$$

For the calculations below, we need only that  $\Phi$  be defined on some open set intersecting  $\mathbf{R}^n$ .

Our prototype for  $\Phi$  is of course the map from  $\mathfrak{u}_n$  to  $\mathfrak{a}$  which takes  $T$  to an element in  $\mathfrak{a}$  conjugate to  $T$ . Globally, this map is not well defined, but

in a neighborhood of any regular point of  $\mathbf{a}$  it is well defined, and satisfies conditions (3.5.6.19).

We want to take a function  $f$  on  $\mathbf{R}^n$ , pull it back by  $\Phi$  to a function on  $\mathbf{R}^{n+m}$ , apply the Laplacian on  $\mathbf{R}^{n+m}$ , then restrict the result to  $\mathbf{R}^n$ . Let  $\Delta_x$  be the Laplacian in the  $x$  variables, and  $\Delta_y$  the Laplacian in the  $y$  variables, so the full Laplacian on  $\mathbf{R}^{n+m}$  is  $\Delta_x + \Delta_y$ . We compute, for  $x \in \mathbf{R}^n$ ,

$$\begin{aligned}
 \Delta(f \circ \Phi)(x, 0) &= (\Delta_x + \Delta_y)(f \circ \Phi)(x, 0) \\
 &= \Delta_x(f)(x) + \left( \sum_{l,j} \frac{\partial}{\partial y_l} \left( \frac{\partial f}{\partial x_j} \frac{\partial \phi_j}{\partial y_l} \right) \right) (x, 0) \\
 (3.5.6.20) \quad &= \Delta_x(f)(x) + \left( \sum_{l,j,k} \frac{\partial^2 f}{\partial x_k \partial x_j} \frac{\partial \phi_j}{\partial y_l} \frac{\partial \phi_k}{\partial y_l} + \frac{\partial f}{\partial x_j} \frac{\partial^2 \phi_j}{\partial y_l^2} \right) (x, 0) \\
 &= \Delta_x(f)(x) + \sum_j (\Delta_y \phi_j)(x, 0) \frac{\partial f}{\partial x_j}(x).
 \end{aligned}$$

We want to compare this with the result of conjugating the Laplacian on  $\mathbf{R}^n$  by a function. Thus we select a function  $\psi$  on  $\mathbf{R}^n$  and we compute (3.5.6.21)

$$\begin{aligned}
 (\psi^{-1} \Delta_x \psi) f(x) &= \psi^{-1}(x) \Delta_x(\psi f)(x) \\
 &= \psi^{-1}(\Delta_x(f)(x) \psi(x) + 2 \psi^{-1}(x) \sum_j \frac{\partial \psi(x)}{\partial x_j} \frac{\partial f(x)}{\partial x_j} \\
 &\quad + \psi^{-1} f(x) \Delta_x(\psi)(x)) \\
 &= \Delta_x(f)(x) + 2 \sum_j \frac{1}{\psi} \frac{\partial \psi}{\partial x_j} \frac{\partial f}{\partial x_j} + f(x) \frac{\Delta_x(\psi)(x)}{\psi(x)}.
 \end{aligned}$$

Comparing formulas (3.5.6.20) and (3.5.6.21), we see that if these two operations are going to be equal, the equations

$$(3.5.6.22a) \quad \frac{2}{\psi} \frac{\partial \psi}{\partial x_j} = \Delta_y \phi_j(x, 0),$$

$$(3.5.6.22b) \quad \Delta_x(\psi) = 0$$

must hold. But equation (3.5.6.22a) already determines  $\psi$  up to a scalar multiple. It will only be by some lucky accident that we find the  $\psi$  so determined to be harmonic, i.e., that condition (3.5.6.22b) also holds. (In addition, the  $\Delta_y \phi_j$  need to satisfy an integrability condition in order for (3.5.6.22a) to have a solution.)

Let us compute the  $\psi$  satisfying (3.5.6.22a) for the case of the eigenvalue projection of the unitary group. Here  $\mathbf{R}^n = \mathbf{a}$ , and the orthogonal space is  $\mathbf{a}^\perp$ , on which we may take coordinates  $2^{-1/2} r_{jk}$  and  $2^{-1/2} s_{jk}$ ,  $1 \leq j < k \leq n$ , where  $t_{jk} = r_{jk} + i s_{jk}$  are the off-diagonal entries of a skew-adjoint matrix  $T$ . Let  $\tilde{E}_{jk}$  be the matrix with all entries zero except for ones in the  $(j, k)$ th



and  $(k, j)$ th places. Let  $B$  be a diagonal matrix, with eigenvalues  $ib_l$ . To compute the right-hand side of formula (3.5.6.22a), we need to compute

$$\frac{d^2}{d\varepsilon^2} b_l(B + \varepsilon \tilde{E}_{jk})|_{\varepsilon=0},$$

where here  $ib_l(B + \tilde{E}_{jk})$  indicates the  $l$ th eigenvalue of  $B + \tilde{E}_{jk}$ , not the  $l$ th diagonal entry, which of course does not depend on  $\varepsilon$ .

It is easy to see that  $b_l$  does not change unless  $l = j$  or  $l = k$ , and that the computation of  $b_j$  and  $b_k$  only involves the  $2 \times 2$  matrix formed from the entries in the  $j$ th and  $k$ th rows and columns of  $B + \varepsilon \tilde{E}_{jk}$ . Thus the computation comes down to a  $2 \times 2$  matrix problem, viz, to find the eigenvalues of

$$\begin{bmatrix} \lambda_1 & \varepsilon \\ \varepsilon & \lambda_2 \end{bmatrix}.$$

We find that they are

$$\begin{aligned} & \frac{1}{2} \left( \lambda_1 + \lambda_2 \pm \sqrt{(\lambda_1 - \lambda_2)^2 + 4\varepsilon^2} \right) \\ & \simeq \frac{1}{2} (\lambda_1 + \lambda_2 \pm |\lambda_1 - \lambda_2|) \left( 1 + \frac{2\varepsilon^2}{(\lambda_1 - \lambda_2)^2} + \cdots \right). \end{aligned}$$

From this result, we easily find that

$$\Delta_y b_j = 2 \sum_{k \neq j} \frac{1}{\lambda_j - \lambda_k}.$$

From this formula, we can see that  $\psi = D$ , the discriminant, will solve the system (3.5.6.22a). Further, it is well known that  $D$ , as the skew-symmetric polynomial of smallest possible degree, is necessarily harmonic. Thus  $D$  satisfies equations (3.5.6.22). Combined with our previous remarks, this establishes Theorem 3.5.6.17.

We now discuss the connection between Harish-Chandra's Restriction Theorem and the Weyl character formula. Consider the Schwartz spaces  $\mathcal{S}(\mathbf{u}_n)$  and  $\mathcal{S}(\mathbf{a})$  of rapidly decreasing smooth functions on  $\mathbf{u}_n$  and  $\mathbf{a}$ . (See [Foll, CoGr, Lang2], etc. for the basic facts on Schwartz spaces.) From Theorem 3.5.6.8 or Theorem 3.5.6.17 one can conclude that the unitary (up to scalars) map  $M_d \circ \text{res}_{\mathbf{a}}$  of formula (3.5.6.7) is also an isomorphism between the Schwartz spaces  $\mathcal{S}(\mathbf{u}_n)^{\text{Ad } U_n}$  and  $\mathcal{S}(\mathbf{a})^{W, \text{sgn}}$ . Dual to this map, we have a pullback map on tempered distributions:

$$(3.5.6.23) \quad (M_d \circ \text{res}_{\mathbf{a}})^* : \mathcal{S}^*(\mathbf{a})^{W, \text{sgn}} \rightarrow \mathcal{S}^*(\mathbf{u}_n)^{\text{Ad } U_n}.$$

Among the conjugation-invariant distributions on  $\mathbf{u}_n$ , probably the most important are the orbital integrals: given  $T \in \mathbf{u}_n$ , the *orbital integral* defined by the conjugacy class  $\text{Ad } U_n(T)$  is

$$(3.5.6.24) \quad \mathcal{I}_T(f) = \int_{U_n} f(\text{Ad } g(T)) dg.$$

Note that to get all possible orbital integrals, one need only consider  $T$  in  $\mathfrak{a}$ . Every conjugation-invariant distribution on  $\mathfrak{u}_n$  is expressible as some sort of superposition of orbital integrals.

The analog of orbital integrals in the space  $\mathcal{S}^*(\mathfrak{a})^{W, \text{sgn}}$  are the skew-symmetric sums

$$(3.5.6.25) \quad \text{skew}(\delta_B) = \sum_{w \in W} \text{sgn}(w) \delta_{w(B)}, \quad B \in \mathfrak{a},$$

where  $\delta_B$  indicates the Dirac delta at  $B$ . Note that  $\text{skew}(\delta_B) \neq 0$  if and only if  $D(B) \neq 0$ . From the elementary computation

$$\begin{aligned} (M_D \circ \text{res}_{\mathfrak{a}})^*(\text{skew}(\delta_B))(f) &= \text{skew}(\delta_B)(M_D \circ \text{res}_{\mathfrak{a}}(f)) \\ &= \text{skew} \delta_B(Df) = \#(W)D(B)f(B), \quad f \in \mathcal{S}(\mathfrak{u}_n)^{\text{Ad } U_n}, B \in \mathfrak{a}, \end{aligned}$$

we conclude

$$(3.5.6.26) \quad (M_D \circ \text{res}_{\mathfrak{a}})^*(\text{skew}(\delta_B)) = \#(W)D(B)\mathcal{F}_B.$$

We want to combine this formula with Theorem 3.5.6.8. Consider the Fourier transform of the orbital integral  $\mathcal{F}_B$ ,  $B \in \mathfrak{a}$ . Since  $\mathcal{F}_B$  has compact support, its Fourier transform has the form

$$(3.5.6.27) \quad \widehat{\mathcal{F}}_B = (\widehat{\mathcal{F}}_B)^0 dT$$

where  $dT$  is Lebesgue measure on  $U_n$  and  $(\widehat{\mathcal{F}}_B)^0$  is an analytic function which is  $\text{Ad } U_n$ -invariant, hence determined by its values on  $\mathfrak{a}$ . Combining formulas (3.5.6.26) and (3.5.6.6) with Theorem 3.5.6.8, we conclude

$$(3.5.6.28) \quad (\widehat{\mathcal{F}}_B)^0(B') = c_1(D(B)D(B'))^{-1} \text{skew} \chi_{-B}(B') \quad B' \in \mathfrak{a}$$

for an appropriate constant  $c_1$ . Here we have written  $\chi_B(B') = e^{2\pi i(B, B')}$ . (An extra computation shows that  $c_1 = (\prod_{k=1}^{n-1} k!)(\frac{i}{2\pi})^m$  with  $m = n(n-1)/2$ .)

Using this formula, we can give an orbit-theoretic interpretation of the Weyl character formula. Let  $\exp: T \rightarrow \exp(T)$  be the natural exponentiation map. The map  $\exp$  allows us to identify a lattice in  $\mathfrak{a}$  with the character group of the torus  $A = \exp \mathfrak{a}$ . Specifically, the restriction of  $\exp$  to  $\mathfrak{a}$  is a group homomorphism. If  $B \in \mathfrak{a}$  is such that  $\ker \chi_B \supseteq \ker \exp$ , then  $\chi_B$  may be pushed forward to  $A$ , where it will define a character. Let us call  $B \in \mathfrak{a}$  *integral* if  $\chi_B$  factors through  $\exp$  on  $\mathfrak{a}$ . In terms of coordinates, we can see that if  $B \in \mathfrak{a}$  has diagonal entries  $ib_j$ , then  $B$  is integral if and only if the  $b_j$ 's are integers. Further, if we identify  $B$  with its  $n$ -tuple of coordinates, then our notation  $\chi_B$  for characters is consistent with the notation of §3.5.4.

The Weyl group  $W$  of permutations acts on  $\mathfrak{a}$  in the obvious way, and this action commutes with  $\exp$ . We have the notion of positive Weyl chamber in  $\mathfrak{a}$  (cf. §2.10). In this case the positive Weyl chamber is

$$\mathfrak{a}^+ = \{B \in \mathfrak{a} : b_j \geq b_{j+1}\}.$$

Let us call  $B$  *dominant* if  $B \in \mathfrak{a}^+$ . Denote by  $\rho$  the element of  $\mathfrak{a}$  whose  $j$ th diagonal entry is  $i(n - j)$ . (The parallel with formulas (3.5.4.18) and (3.5.5.16) will be evident. The need to multiply by  $i$  here comes simply from the concrete form of the Cartan subalgebra  $\mathfrak{a}$ .)

With these notations, we can state a formula which combines the Weyl character formula with the Harish-Chandra restriction formula.

**THEOREM (3.5.6.29)** (Harish-Chandra-Weyl character formula). *The irreducible characters of  $U_n$ , as functions on the maximal torus  $A = \exp \mathfrak{a}$ , have the form:*

$$(3.5.6.30) \quad \text{ch}_\sigma(\exp B') = \frac{D(B_{\sigma+\rho})D(B')(\widehat{\mathcal{F}}_{B_{\sigma+\rho}})^0(-B')}{c_1 D(\exp B')},$$

where  $B_\sigma$  is an appropriate dominant integral element of  $\mathfrak{a}$ .

**REMARKS.** (a) This formula is the analog for compact groups of the Kirillov character formula (3.3.1.7). A parallel for solvable groups which involves multiplication of the Fourier transform of the orbital integral by a correction factor is formula (3.4.1.2.1).

(b) Here again, as in the Verma module description of finite-dimensional representations (§5.3.3), and in the Weyl character formula (§5.3.4), we see a “ $\rho$ -shift” between the highest weight of the representation and the parameter we attach to the representation. Thus the  $\text{Ad } U_n$ -orbit associated to the trivial representation of  $U_n$  is not the origin, but rather the orbit through  $\rho$ . This phenomenon of  $\rho$ -shifts pervades the orbit method for semisimple groups. It is bookkeeping forced on us by the Harish-Chandra homomorphism (cf. Theorem 3.5.5.23).

(c) The argument given here for the Harish-Chandra Restriction Theorem and formula (3.5.6.30) looks quite different from the ones based on [HaCh3] (cf. [Helg1, Wall2]). Harish-Chandra first studies radial components of invariant differential operators, then uses them to deduce formula (3.5.6.30), then finally proves the Restriction Theorem. We established the Restriction Theorem first, then deduced formula (3.5.6.30). We could easily also deduce the results on radial components from the Restriction Theorem. However, although the order of main results is different, the crucial step in both developments is the computation of the radial component of the Laplacian (formula (3.5.6.20) and the discussion following it). In [HaCh3], the oscillator representation appears only implicitly, in the use of taking commutators with the Laplacian to convert an invariant polynomial into the dual constant coefficient operator.

**3.6. Noncompact semisimple groups.** Noncompact semisimple groups have received the bulk of researchers’ attention in representation theory since World War II, beginning with the papers of Wigner [Wign], Bargmann [Barg1], Gelfand-Naimark [GeNa], and Harish-Chandra [HaCh0]. Until the late

1960s, Harish-Chandra was a fairly lonely pioneer, but since then the field has attracted a substantial number of workers. Fundamental progress has been made, including Harish-Chandra's Plancherel Formula [HaCh22], and the classifications of Bernstein-Beilinson [BeBe], Langlands [Lgl4], and Vogan [Voga4] of the nonunitary irreducible representations. But many interesting and even basic problems, such as the determination of the unitary dual, remain to be solved (see, however, [Voga5, Barb, Tadi]), and much of the work already done sits undigested and unapplied.

In this account, we can only summarize some of the high points. We try to emphasize analogies with the easier classes of groups already discussed, and in particular we try to formulate results in terms of the orbit method. However, we emphasize that the structure of geometric quantization is for the most part imposed a posteriori, and played little role in the original arguments. Nevertheless, David Vogan currently is trying to create an understanding of the unitary dual more or less explicitly based on an appropriate version of the orbit method [Voga6].

Due to the greater length and technical involvement of the arguments establishing results about noncompact semisimple groups, we must for the most part omit them, and be content with stating results. Two very useful books for learning a large portion of the theory in its current form are [Knap2] and [Wall2]. We also refer to [Voga1] for a nice overview.

**3.6.1. PRINCIPAL SERIES.** The main concrete objects of study in the representation theory of noncompact semisimple groups are the principal series. As with many things, the meaning of the term "principal series" can vary slightly with context. We begin by describing the most elementary case.

Let  $G$  be a semisimple Lie group, and let  $P_0 \subseteq G$  be a minimal parabolic subgroup (cf. §A.2.4). We have a decomposition

$$(3.6.1.1) \quad P_0 = M_0 A_0 N_0,$$

where  $N_0$  is the unipotent radical of  $P_0$  (cf. §A.2.4), a connected, simply connected nilpotent group;  $A_0$  is an abelian group, connected and simply connected (i.e., isomorphic to  $\mathbf{R}^m$  for  $m = \dim A_0$ ), and such that under the adjoint action,  $A_0$  acts by diagonalizable matrices with positive real eigenvalues; and  $M_0$  is compact. The group  $N_0$  is normal in  $P_0$ , and  $M_0$  and  $A_0$  centralize each other. If  $G = \mathrm{SL}_n(\mathbf{R})$ , then  $N_0$  consists of the unipotent upper triangular matrices,  $A_0$  is the group of diagonal matrices with positive entries and determinant one, and  $M_0$  is the group of diagonal matrices with entries  $\pm 1$  and determinant one.

Let  $\psi$  be a quasicharacter of  $A_0$  (a homomorphism from  $A_0$  to  $\mathbf{C}^\times$ ), and  $\sigma$  an irreducible representation of  $M_0$ . Note that  $\sigma$  is finite dimensional. If  $V$  is the space of  $\sigma$ , define the representation  $\sigma \otimes \psi$  of  $P_0$  on  $V$  by

$$(3.6.1.2) \quad \begin{aligned} \sigma \otimes \psi(man)(v) &= \psi(a)\sigma(m)(v), & m \in M_0, a \in A_0, \\ & n \in N_0, v \in V. \end{aligned}$$

Let  $\delta_{P_0}$  be the modular function of  $P_0$  (cf. formula (A.1.15.3)). Define the *principal series* representation associated to  $\sigma$  and  $\psi$  to be the induced representation (cf. §§A.1.14–16).

$$(3.6.1.3) \quad \text{P.S.}(\sigma, \psi) = \text{ind}_{P_0}^G \sigma \otimes (\psi \delta_{P_0}^{-1/2}).$$

The set of representations  $\text{P.S.}(1, \psi)$ , where 1 here denotes the trivial representation of  $M_0$ , is called the *spherical principal series*. (From an etymological viewpoint, this is a solecism: zonal principal series would be preferable.) The spherical principal series are slightly simpler than the  $\text{P.S.}(\sigma, \psi)$  with  $\sigma$  nontrivial, and they have some claim to a special place: they encompass all irreducible representations of  $G$  which contain a nonzero fixed vector for  $K$ , and consequently, they are the representations involved in the spectral analysis of functions on the symmetric space  $G/K$ . These topics are treated in detail in [Helg1] and [GaVa].

The quasicharacters of  $A_0$  form a complex vector space  $\hat{A}_0^{\mathbb{C}}$  of dimension  $\dim A_0$ . Thus if we let  $\psi$  vary in  $\hat{A}_0^{\mathbb{C}}$ , the representations  $\text{P.S.}(\sigma, \psi)$  form a family, which in some sense (which can be made precise) is continuous or even holomorphic, of similar-looking representations. This is what the “series” in “principal series” connotes. Of course,  $\delta_{P_0}^{-1/2}$  is a point in  $\hat{A}_0^{\mathbb{C}}$ , and so multiplying  $\psi$  by  $\delta_{P_0}^{-1/2}$  before forming the induced representation does not change the family of representations constructed, it only changes the way they are parametrized. The point of the chosen parametrization is that it takes unitary representations to unitary representations (cf. §§A.1.3 and A.1.16). For this reason multiplying by  $\delta_{P_0}^{-1/2}$  before inducing is called *normalized induction* or *unitary induction*. The representations  $\text{P.S.}(\sigma, \psi)$  for  $\psi$  unitary are called the *unitary principal series*, and, by way of contrast, the whole principal series is sometimes called the *nonunitary principal series*. (We remark, however, that even for some nonunitary  $\psi$ , the representation  $\text{P.S.}(\sigma, \psi)$  can be given the structure of a unitary representation, though not in straightforward fashion [Ste1, Knap2, p. 653].)

EXAMPLE. As an example, consider  $G = \text{SL}_2(\mathbf{R})$ . We may take

$$(3.6.1.4) \quad \begin{aligned} P_0 &= B = \left\{ \begin{bmatrix} a & x \\ 0 & a^{-1} \end{bmatrix} : a \in \mathbf{R}^\times, x \in \mathbf{R} \right\}, \\ M_0 &= \left\{ \pm \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\}, \quad A_0 = \left\{ \begin{bmatrix} a & 0 \\ 0 & a^{-1} \end{bmatrix} : a > 0 \right\}, \\ N_0 &= \left\{ \begin{bmatrix} 1 & x \\ 0 & 1 \end{bmatrix} : x \in \mathbf{R} \right\}. \end{aligned}$$

Consider the space  $\mathbf{C}^{\lambda, \varepsilon}(\mathbf{R}^2)$  of smooth functions on  $\mathbf{R}^2 - \{0\}$  which are homogeneous of degree  $\lambda$ ,  $\lambda \in \mathbf{C}$ , under positive dilations, and which are

odd or even under reflection in the origin:

$$(3.6.1.5) \quad \begin{aligned} \mathbf{C}^{\lambda, \varepsilon}(\mathbf{R}^2) &= \{f : (\mathbf{R}^2 - \{0\}) \rightarrow \mathbf{C}, f \text{ smooth}, \\ f(tx, ty) &= t^\lambda f(x, y) \text{ for } t > 0, \text{ and} \\ f(-x, -y) &= \varepsilon f(x, y)\} \end{aligned}$$

for  $\lambda \in \mathbf{C}$ ,  $\varepsilon = \pm 1$ . The action of  $\mathrm{SL}_2(\mathbf{R})$  on  $\mathbf{R}^2$  gives rise in a natural way to an action  $\rho$  on  $\mathbf{C}^\infty(\mathbf{R}^2 - \{0\})$  by the recipe

$$(3.6.1.6) \quad \rho(g)(f) \left( \begin{bmatrix} x \\ y \end{bmatrix} \right) = f \left( g^{-1} \begin{bmatrix} x \\ y \end{bmatrix} \right).$$

It is easy to check that the spaces  $\mathbf{C}^{\lambda, \varepsilon}(\mathbf{R}^2)$  are invariant under  $\rho$ , so we may restrict  $\rho$  to any one of the  $\mathbf{C}^{\lambda, \varepsilon}$ .

Define a mapping  $E$  from functions on  $\mathbf{R}^2 - \{0\}$  to functions on  $\mathrm{SL}_2(\mathbf{R})$  by the rule

$$(3.6.1.7) \quad E(f)(g) = f \left( g^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right) = f \left( \begin{bmatrix} d \\ -c \end{bmatrix} \right),$$

$$f \in C^\infty(\mathbf{R}^2 - \{0\}), \quad g = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2(\mathbf{R}).$$

A straightforward calculation reveals that the mapping  $E$  defines an equivalence of  $\mathrm{SL}_2(\mathbf{R})$  representations

$$(3.6.1.8) \quad \mathbf{C}^{-\lambda-1, \varepsilon}(\mathbf{R}^2) \simeq \mathrm{P.S.}(\tilde{\varepsilon}, \tilde{a}^\lambda),$$

where  $\tilde{\varepsilon} : M_0 \rightarrow \{\pm 1\}$  is defined by  $\tilde{\varepsilon}(\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}) = \varepsilon$  and

$$\tilde{a}^\lambda \left( \begin{bmatrix} a & 0 \\ 0 & a^{-1} \end{bmatrix} \right) = a^\lambda, \quad a > 0.$$

Thus the  $\mathbf{C}^{\lambda, \varepsilon}(\mathbf{R}^2)$  serve as models for the principal series of  $\mathrm{SL}_2(\mathbf{R})$ .

Because of the homogeneity conditions (3.6.1.5) defining  $\mathbf{C}^{\lambda, \varepsilon}(\mathbf{R}^2)$ , we see that a function in this space is determined by its restriction to the unit circle

$$S^1 = \{(x, y) \in \mathbf{R}^2 : x^2 + y^2 = 1\}$$

and this restriction must be either an even or an odd function according as  $\varepsilon$  is  $+1$  or  $-1$ . The circle  $S^1$  is an orbit for the maximal compact subgroup  $K = \mathrm{SO}_2$  of  $\mathrm{SL}_2(\mathbf{R})$ . Thus the Fourier series of  $f|_{S^1}$  describes the decomposition of  $f$  into irreducible subspaces (in this example, eigenspaces, since  $\mathrm{SO}_2$  is abelian) for  $K$ . In particular, we see that each representation of  $K$  occurs with multiplicity at most one. For general groups, the Iwasawa decomposition (cf. equation A.2.3.5) shows us that the restriction to  $K$  of  $\mathrm{P.S.}(\sigma, \psi)$  is also an induced representation

$$(3.6.1.9) \quad \mathrm{P.S.}(\sigma, \psi)|_K \simeq \mathrm{ind}_{M_0}^K \sigma.$$

By Frobenius reciprocity ([HeRo, Knap2, Jaco2] etc.), we may conclude that an irreducible representation  $\tau$  of  $K$  occurs in  $\text{P.S.}(\sigma, \psi)$  with multiplicity equal to the multiplicity with which  $\sigma$  occurs in the restriction  $\tau|_{M_0}$  of  $\tau$  to  $M_0$ . This is certainly not more than  $\dim \tau$ . Thus all representations of  $K$  (= “ $K$ -types”) occur in the principal series with finite multiplicity, which means that the principal series are admissible (cf. §3.6.5) representations [Knap2, Wall2].

The importance of the principal series is brought out by the following result.

**THEOREM 3.6.1.10.** (a) *The principal series representations  $\text{P.S.}(\sigma, \psi)$  all have finite composition series. The number of composition factors is bounded independently of  $\sigma$  and  $\psi$ . For fixed  $\sigma$ ,  $\text{P.S.}(\sigma, \psi)$  is irreducible for a dense open set of  $\psi$ .*

(b) *Let  $\rho$  be any t.c.i. (cf. §A.1.7) representation of  $G$ . Then  $\rho$  is infinitesimally equivalent (cf. §A.1.20) to a subrepresentation of  $\text{P.S.}(\sigma, \psi)$  for appropriate  $\sigma$  and  $\psi$ .*

A weaker version of part (b), only asserting that  $\sigma$  could be realized as a constituent, i.e., subquotient, of some  $\text{P.S.}(\sigma, \psi)$ , was proved by Harish-Chandra in early work [HaCh4], and later simplified by Lepowsky [Lepo3] and Rader (see also [Wall2]). The refinement giving  $\sigma$  as a subrepresentation was a long-standing problem, resolved by Casselman (see [CaMi]), using a refined version of Harish-Chandra’s study [HaCh13] of the asymptotics of matrix coefficients. This study was based on the observation that elements of the center of the enveloping algebra give rise to differential equations which the matrix coefficients must satisfy. The differential equations imply that the matrix coefficients of an irreducible representation have certain asymptotic behavior at  $\infty$  on  $G$ ; this asymptotic behavior identifies the principal series into which the representation may be embedded.

The generic irreducibility of  $\text{P.S.}(\sigma, \psi)$ , and finiteness of the composition series in general, has a fuzzier history. Generic irreducibility of the unitary principal series was proved by Bruhat [Bruh]. Finiteness of the composition series follows from Harish-Chandra’s Regularity Theorem for characters [HaCh14–18] (see also [Wall2, Vara]), but proved this way, it is a deep result. Wallach [Wall2] gives a proof using his “Jacquet module.” The composition series of  $P(\sigma, \psi)$  when at least one constituent is finite dimensional is described by Vogan’s extension of the Kazhdan-Lusztig formulas [KaLu1, Voga7]. Explicit examples and refinements have been given by Casian and Collingwood [CaCo, Coll]. However, there is still much to understand regarding the structure of these easily constructed representations.

Shortly after Casselman’s proof of part (b), Langlands [Lgld4] (see also [Knap2, Wall2]) showed, again on the basis of Harish-Chandra’s study of asymptotics of matrix coefficients, that by using a more general notion of principal series one can obtain a more or less canonical realization of a gen-

eral irreducible representation. This is the “Langlands classification,” which we will describe in §3.6.4. Here we describe the more general family of representations.

Let  $P \subseteq G$  be any parabolic subgroup (cf. §A.2.4). Let  $P = MAN$  be the Langlands decomposition ([Knap2, Wall2] and §A.2.4) of  $P$ , where  $N$  is the unipotent radical of  $P$ ,  $A$  is a connected, simply connected abelian group, and  $M$  is semisimple. The group  $MA$  is the centralizer of  $A$  in  $G$ , and is a Levi component (cf. [Jaco1] and §A.2.4) for  $P$ . Let  $\sigma$  be an irreducible t.c.i. representation of  $M$ , and  $\psi$  a quasicharacter of  $A$ . Let  $\delta_P$  be the modular function of  $P$ . We can define a representation  $\sigma \otimes \psi$  of  $P$  in direct analogy with formula (3.6.1.2). Then we define the *generalized principal series* representation associated to  $\sigma$  and  $\psi$  to be

$$(3.6.1.11) \quad \text{P.S.}(\sigma, \psi) = \text{ind}_P^G \sigma \otimes (\psi \delta_P^{-1/2}).$$

The parallel to formula (3.6.1.3) is patent, and sometimes one drops the adjective “generalized” and just calls the representations (3.6.1.11) “principal series.” Note, however, that recipe (3.6.1.11) is much more of a black box than is (3.6.1.3), because the  $\sigma$  in (3.6.1.3) is a representation of the compact group  $M_0$  and thus is to some extent understood, as described in §3.5. However, since  $M$  is noncompact, determination of the possible  $\sigma$  to stick in (3.6.1.11) is part of the problem under study, although for a smaller group. Further, we note that, by Harish-Chandra’s Subquotient Theorem mentioned above, if  $\sigma$  is irreducible, the representations (3.6.1.11) are constituents of the usual principal series (3.6.1.3). For the Langlands classification, we only have to stick in for  $\sigma$  a special class of representations, the “tempered representations,” to be described in §3.6.2.

We note again that (3.6.1.11) is a “normalized induction”: if  $\sigma$  and  $\psi$  are unitary, then  $\text{P.S.}(\sigma, \psi)$  is also unitary.

We should also note that the formation of principal series, also known as *parabolic induction*, is eminently compatible with the orbit method. Let

$$(3.6.1.12) \quad \mathfrak{g} = \mathfrak{n}^- \oplus \mathfrak{p} = \mathfrak{n}^- \oplus \mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}$$

be the decomposition of the Lie algebra of  $G$  associated to the parabolic  $P$ . If the quasicharacter  $\psi$  in definition (3.6.1.6) is unitary, then it is the exponential of some  $\lambda \in \mathfrak{a}^*$  in the usual way:

$$\psi(\exp a) = e^{2\pi i \lambda(a)}, \quad a \in \mathfrak{a}.$$

(If  $\psi$  is not unitary, we could still use this formula, but would have to take  $\lambda$  in  $\mathfrak{a}_\mathbb{C}^*$ , the complexified dual of  $\mathfrak{a}$ .) Suppose we have associated the representation  $\sigma$  of  $M$  to the coadjoint orbit through some  $\mu \in \mathfrak{m}^*$ , and to some polarization  $\mathfrak{q}$  of  $\mu$  (which may well be a complex polarization, i.e.,  $\mathfrak{q} \subseteq \mathfrak{m}_\mathbb{C}^*$ ). Then it is easy to check that, for generic  $\lambda \in \mathfrak{a}^*$ , the subalgebra  $\mathfrak{q} \oplus \mathfrak{a}_\mathbb{C} \oplus \mathfrak{n}_\mathbb{C}$  of  $\mathfrak{g}_\mathbb{C}$  will be a polarization of  $\mu \oplus \lambda \in \mathfrak{m}^* \oplus \mathfrak{a}^* \subseteq \mathfrak{g}^*$ , and definition (3.6.1.11) would be the representation associated to the  $\text{Ad } G$ -orbit



of  $\mu \oplus \lambda$  according to the usual orbit method yoga. In particular, if  $M_0$  is discrete (which is the case for split real groups like  $\mathrm{GL}_n(\mathbf{R})$ ,  $\mathrm{Sp}_{2n}(\mathbf{R})$ , etc.) or if  $M_0$  is abelian (which is the case for complex groups, or quasi-split groups like  $\mathrm{U}(n, n)$ ), then the unitary principal series of definition (3.6.1.3) are constructed via real polarizations. If  $M_0$  is nonabelian, then we can understand its representations in terms of complex polarizations via the Borel-Weil Theorem (3.5.5.13), so the unitary principal series at least can be given a place in the orbit method. The combination of the Langlands classification (§3.6.4) and Zuckerman's derived functor construction (§3.6.5) extend this understanding to a class of representations at least big enough to write the Plancherel formula. A guide for further progress is provided by some conjectures of J. Arthur [Arth2], and the understanding of a family of representations dubbed unipotent [BaVo, Voga1], in homage to Lusztig's theory of representations of finite Chevalley groups [Lus1]. Alternatively, the character theory of Harish-Chandra [HaCh17–20, Wall2, Vara], refined by Rossmann [Ross1, 2] provides the direct connection of representations with orbits exemplified by formula (3.3.1.7).

**3.6.2. TEMPERED REPRESENTATIONS.** One of Harish-Chandra's basic insights into harmonic analysis on semisimple groups was the key role of what he called *tempered* representations. For his main goal, the Plancherel formula, the tempered representations were essential because, as he showed constructively, they are precisely the representations needed to perform the spectral analysis of  $L^2(G)$ . (This fact is now understood a priori [CoHH, Bern3]. It marks a fundamental difference between harmonic analysis on abelian, or even solvable groups, and semisimple groups.) They have turned out to be a basic ingredient in several other problems in representation theory, particularly problems suggested by automorphic forms [Arth2, Sata], and the Langlands classification (see §3.6.4).

Tempered representations are defined in terms of the decay of their matrix coefficients at  $\infty$  on the group. The precise definition is in terms of a certain function  $\Xi$  introduced by Harish-Chandra [HaCh10] (see also [Wall2]). In fact  $\Xi$  is a natural function to consider: it is the matrix coefficient (cf. §A.1.11) associated to the (unique)  $K$ -invariant vector in P.S.(1, 1), the spherical principal series associated to the trivial character of  $A_0$ . We can give an integral formula for  $\Xi$  as follows. Extend the modular function  $\delta_{P_0}$  on  $P_0$  to all of  $G$  by requiring the extended function to be invariant under left translation by  $K$ . Thus define

$$(3.6.2.1) \quad \delta_{P_0}(kp) = \delta_{P_0}(p), \quad k \in K, p \in P_0.$$

Then  $\Xi$  is defined by

$$(3.6.2.2) \quad \Xi(g) = \int_K \delta_{P_0}^{-1/2}(gk) dk.$$

Here we take Haar measure on  $K$  to have total mass 1. The reader may wish to check that this is indeed the  $K$ -invariant matrix coefficient of P.S.(1, 1).

Relevant definitions and formulas are given in §§A.1.11 and A.1.14–16. In any case, it is straightforward to check that  $\Xi$  is both left- and right-invariant by  $K$ , that  $\Xi(g) \geq 0$ , and that  $\Xi(1) = 1$ . Harish-Chandra established the following asymptotic properties of  $\Xi$ . We observe that by the Cartan decomposition (cf. formula (A.2.3.2)) the  $K$ -bi-invariance of  $\Xi$  implies that it is determined by its values on  $A_0^+$ , the positive Weyl chamber in  $A_0$ .

**THEOREM 3.6.2.3.** (a) For  $a \in A_0^+$ , and any  $\varepsilon > 0$ , we have the estimates

$$\delta_{P_0}^{-1/2}(a) \leq \Xi(a) \leq C_\varepsilon \delta_{P_0}^{-1/2+\varepsilon}(a)$$

for an appropriate constant  $C_\varepsilon$ .

(b) For any  $\varepsilon > 0$ ,

$$\int_G \Xi^{2+\varepsilon}(g) dg < \infty.$$

Actually Harish-Chandra proves more refined estimates than these [HaCh10, Knap2, Wall2, Vara]; but these statements give the basic flavor of his results. Statement (b) is often phrased: “ $\Xi$  belongs to  $L^{2+\varepsilon}(G)$ .”

Having  $\Xi$  in hand, we may define the notion of tempered representation. A representation  $\rho$  on the space  $V$  is called *tempered* if all its smooth matrix coefficients  $\varphi_{u,\lambda}$  (cf. §A.1.11),  $u \in V^\infty$ ,  $\lambda \in (V^*)^\infty$ , satisfy

$$(3.6.2.4) \quad \varphi_{u,\lambda}(g) \leq C_{u,v} \Xi(g)$$

for an appropriate constant depending on  $u$  and  $v$ . This is not precisely Harish-Chandra’s definition, but is equivalent to it [CoHH]. It is also equivalent to requiring the smooth matrix coefficients to be  $L^{2+\varepsilon}$ .

**3.6.3. DISCRETE SERIES.** Just as tempered representations are essential to the spectral analysis of  $L^2(G)$  for semisimple groups, the discrete series are essential to understanding tempered representations. Beyond this, discrete series are a fascinating phenomenon of general harmonic analysis. Also, discrete series for semisimple groups play a prominent role in the theory of automorphic forms [BoWa, Lgld7, DeGW]. The definition of discrete series makes sense for a general unimodular locally compact group  $G$ . Let  $\rho$  be an irreducible unitary representation of  $G$  on a Hilbert space  $\mathcal{H}$ . We call  $\rho$  a *discrete series* (or *square integrable*) if there exist  $u, v$  in  $\mathcal{H}$  such that the matrix coefficient  $\varphi_{u,v}$  (cf. §A.1.11) belongs to  $L^2(G)$ .

This rather innocent sounding definition has striking implications, described in the next result. The proof is a pleasant exercise in functional analysis, originally done by Godement [Gode1] (see also [Knap2]).

**THEOREM 3.6.3.1.** Let  $\rho$  be a discrete series representation of the unimodular locally compact group  $G$ . Let  $\rho$  be realized on the Hilbert space  $\mathcal{H}$ . Then the following assertions are true:

- (i) Every matrix coefficient  $\varphi_{u,v}$ ,  $u, v \in \mathcal{H}$ , of  $\rho$  is in  $L^2(G)$ .

(ii) There is a constant  $d_\rho$  such that

$$(3.6.3.2) \quad \int_G \varphi_{u,v}(g) \overline{\varphi_{w,z}(g)} dg = \frac{(u, w)(z, v)}{d_\rho}$$

for any  $u, v, w, z \in \mathcal{H}$ . Here  $(\ , \ )$  indicates the inner product in  $\mathcal{H}$ .

(iii) In particular, for fixed  $v \in \mathcal{H}$ , the mapping

$$u \rightarrow \left( \frac{d_\rho}{(v, v)} \right)^{1/2} \varphi_{u,v}$$

defines an isometric  $G$ -intertwining from  $\mathcal{H}$  to a subspace of  $L^2(G)$ . Thus  $\rho$  is equivalent to a summand of  $L^2(G)$ .

REMARKS. (a) Equation (3.6.3.2) is a generalization of the Schur orthogonality relations for finite or compact groups (cf. [HeRo, Knap2, Jaco2], etc.). For those groups, if Haar measure is normalized to have total mass 1, the constant  $d_\rho$  is just the dimension of  $\mathcal{H}$ , also known of old as the degree of  $\rho$ . In the general case, when  $\mathcal{H}$  is infinite dimensional,  $d_\rho$  is called the *formal degree* of  $\rho$ .

(b) The equation (3.6.3.2) is reminiscent of a fixed point formula: if  $u = v = w = z$ , then it expresses the integral of  $|\varphi_{u,u}|^2$  as a multiple of  $(u, u)^2 = |\varphi_{u,u}|^2(1)$ , where 1 denotes the identity element of  $G$ . Indeed, formula (3.6.3.2) has a natural interpretation as a “trace formula.”

With this background on discrete series for general groups, let us explain for semisimple Lie groups the relation between discrete series and tempered representations.

THEOREM 3.6.3.3. (a) If  $G$  is a semisimple Lie group, and  $\rho$  is a discrete series representation of  $G$ , then  $\rho$  is tempered.

(b) Let  $P \subseteq G$  be a parabolic subgroup with decomposition  $P = MAN$  as in equation (3.6.1.5). If  $\sigma$  is a tempered representation of  $M$ , and  $\psi \in \hat{A}$  is a unitary character, then the principal series representation  $\text{P.S.}(\sigma, \psi)$  (cf. formula (3.6.1.11)) is tempered. (In brief: unitary parabolic induction preserves temperedness.)

(c) Every irreducible tempered representation of  $G$  is a summand of some  $\text{P.S.}(\sigma, \psi)$  as in part (b), where  $\sigma$  is a discrete series.

The complete classification of tempered representations, i.e., a description of the precise decomposition of the tempered  $\text{P.S.}(\sigma, \psi)$ , and the equivalences between the pieces, was given by Knapp and Zuckerman [KnZu]. Their results were given a nice orbit method interpretation by Rossmann [Ross2], as an adjunct to his character formula, to be discussed in Theorem 3.6.3.7. However, it was known already from results of Bruhat in the 1950s [Bruh] for the basic principal series that for any  $\sigma$  and generic (i.e., for an open dense set of)  $\psi$ , the representation  $\text{P.S.}(\sigma, \psi)$  is irreducible. (It was to prove this

that Bruhat studied the double coset decomposition now named after him.) These results were extended by Harish-Chandra to cover the case when  $P$  is nonminimal and  $\sigma$  is a discrete series. Hence Theorem 3.6.3.3 gives a description of “almost all” the tempered representations. In particular, the representations described by Theorem 3.6.3.3 are precisely the representations which enter into Harish-Chandra’s Plancherel formula for  $G$ .

In sum, Theorem 3.6.3.3 shows that an understanding of tempered representations can to a large extent be reduced to an understanding of the discrete series.

Harish-Chandra [HaCh19–20] gave a description of the discrete series of a semisimple Lie group. He did so by explicitly constructing their characters, which he expressed in terms of Fourier transforms of invariant measures on certain orbits in the dual of the Lie algebra of  $G$ , a procedure with obvious analogies to the orbit method sketched above for nilpotent and solvable groups, and with Harish-Chandra’s own formula for the characters of compact groups. That the parallel is essentially perfect, so that the characters of discrete series, and in fact of all tempered representations, can be described by a close cousin of the formula of Theorem 3.5.6.29, was established by W. Rossmann [Ross1, 2]. We will describe this in Theorem 3.6.3.7.

Since Harish-Chandra’s classification of discrete series did not actually produce representations, in the sense of providing some concretely described spaces with some concretely given  $G$ -actions on them, a clamor soon arose for a “geometrical realization” of the discrete series. A candidate for such a realization, using “ $L^2$ -cohomology” and bearing strong analogies to the Bott-Borel-Weil Theorem (cf. §3.5.5) for compact groups was proposed by Langlands and Kostant [Lgld7, Kost7]. A realization of this sort was established in stages by W. Schmid [Schm1–3]. Other authors used variations on this theme to produce similar models for the discrete series (cf. [Hott, OkOz, Part1, Wall4], etc.). However, all of these constructions depended on Harish-Chandra’s existence proof via character theory; they did not independently establish either existence or exhaustion of the discrete series.

In the 1970s a more algebraic approach to problems of representation theory arose, and several purely algebraic constructions of discrete series and analogous representations were given [EnVa, Part2, Zuck]. The construction given by G. Zuckerman (see [Voga, Wall2, Knap1]) has turned out to be in some sense the most natural and has been shown to have numerous pleasant technical properties. It is now more or less the standard construction, and has been developed to the point where it can be used to give an independent proof of the existence of the discrete series [Wall2]. Proof that the construction exhausts all discrete series, however, still requires character theory. We describe Zuckerman’s construction in §3.6.5.

Several other, rather different, ways to construct discrete series have also been developed. A construction of discrete series on the family of “semisimple symmetric spaces”—homogeneous spaces of the form  $G/H$ , where  $G$  is

semisimple and  $H$  is the identity component of an automorphism of order 2 of  $G$  (see [FlJe2, Berg])—was discovered by M. Flensted-Jensen [FlJe2, Knap2]. This proceeds by giving integral formulas for certain matrix coefficients of the representations, by means of a beautiful duality between pairs of semisimple symmetric spaces [FlJe2]. Since a semisimple group  $G$  can be thought of as the semisimple symmetric space  $G \times G/\Delta(G)$ , where

$$\Delta(G) = \{(g, g) : g \in G\}$$

is the diagonal in  $G \times G$ , Flensted-Jensen's construction yields discrete series for  $G$  as a special case. Flensted-Jensen's methods have been extended and to some extent completed by Oshima [Oshi1, 2], using the theory of hyperfunctions and holonomic systems.

Another, again quite different, construction of the discrete series using an index theorem for covering spaces was given by M. Atiyah and W. Schmid [AtSc].

Let us turn now to a concrete description of discrete series. First, we should note that not all semisimple groups have discrete series. A major insight of Harish-Chandra was that discrete series should be associated to compact Cartan subgroups. A Cartan subgroup of  $G$  is a subgroup whose complexified Lie algebra is a Cartan subalgebra of  $\mathfrak{g}_{\mathbb{C}}$ .

**THEOREM 3.6.3.4.** *Let  $G$  be a semisimple Lie group and  $K \subseteq G$  a maximal compact subgroup. Then  $H$  has discrete series if and only if  $\text{rank } K = \text{rank } G$ , i.e., iff a Cartan subgroup of  $K$  is also a Cartan subgroup in  $G$  iff  $G$  has a compact Cartan subgroup.*

Thus, for example, the rank of  $\text{SL}_n(\mathbb{R})$  is  $n - 1$ , and that of its maximal compact subgroup  $\text{SO}_n$  is  $[n/2]$ . These are equal only for  $n = 2$ , so for  $n \geq 3$ ,  $\text{SL}_n(\mathbb{R})$  has no discrete series. Since  $\text{SL}_n(\mathbb{R})$  occurs as a factor of the Levi component of maximal parabolics of many groups (e.g. of  $\text{O}_{p,q}$ ,  $\text{Sp}_{2n}(\mathbb{R})$ ), this very substantially cuts down on the number of parabolic subgroups one must worry about in the context of Theorem 3.6.3.3. Also, complex groups always have their compact real form as a maximal compact subgroup, and this always has rank equal to  $\frac{1}{2}$  the rank of the full group (considered as a real Lie group), so complex groups have no discrete series. As a result, the only tempered representations for complex groups are constituents of the standard principal series induced from characters of the minimal parabolic subgroup. With hindsight we may say that it was this circumstance that permitted the early determination by Gelfand-Naimark [GeNa] and Harish-Chandra [HaCh5] of the Plancherel formula for complex semisimple groups.

Other examples are:  $\text{Sp}_{2n}(\mathbb{R})$  has maximal compact  $\text{U}_n$ , and both have rank  $n$ , so  $\text{Sp}_{2n}(\mathbb{R})$  has discrete series;  $\text{O}_{p,q}$  has rank  $[\frac{p+q}{2}]$ , while its maximal compact  $\text{O}_p \times \text{O}_q$  has rank  $[\frac{p}{2}] + [\frac{q}{2}]$ , so  $\text{O}_{p,q}$  has discrete series if and only if at least one of  $p$  and  $q$  is even. In this connection, we may note that if  $p$  is odd, then  $\text{O}_{p,1}$ , like complex groups, has only the standard prin-

cipal series as tempered representations, and hence has a simple Plancherel formula, analogous to that for complex groups [Wall3].

Following Theorem 3.6.3.4, consider a semisimple Lie group  $G$  containing a compact Cartan subgroup  $T$ . The discrete series of  $G$  are then parametrized, more or less, by the characters of  $T$ , in a fashion similar to the description given in §§3.5.3–3.5.5 for the case of  $G$  compact. However, the “more or less” hides several tricky points, of which we will try to give some idea.

We will describe the discrete series by associating to them coadjoint orbits. This gives a formula for their characters, which is the original description given by Harish-Chandra [HaCh19, 20]. The refinement to the orbital description is due to Rossmann [Ross1]. For convenience, we will restrict our discussion to connected groups. The results can be extended to more general  $G$ , but at the expense of substantial technical fussing.

Let  $G$  be a semisimple group, connected and without center. Let  $K \subseteq G$  be a maximal compact subgroup of  $G$  and  $T \subseteq K$  a maximal torus (Cartan subgroup). We assume  $T$  is also a Cartan subgroup of  $G$ ; this means (since  $G$  is connected) that  $T$  is its own centralizer in  $G$ . Let  $\mathfrak{g}$ ,  $\mathfrak{k}$ , and  $\mathfrak{t}$  be the Lie algebras of  $G$ ,  $K$ , and  $T$  respectively, and let  $\mathfrak{g}^*$ ,  $\mathfrak{k}^*$ , and  $\mathfrak{t}^*$  be the duals of  $\mathfrak{g}$ ,  $\mathfrak{k}$ , and  $\mathfrak{t}$ . Via the Killing form (cf. equation (2.8.8)) on  $\mathfrak{g}$ , we can identify  $\mathfrak{g}$  with  $\mathfrak{g}^*$ ,  $\mathfrak{k}$  with  $\mathfrak{k}^*$ , and  $\mathfrak{t}$  with  $\mathfrak{t}^*$ . We will not take full advantage of this identification, however, but merely use it to consider  $\mathfrak{t}^*$ , which naturally is a quotient of  $\mathfrak{g}^*$ , as a subspace of  $\mathfrak{g}^*$ .

Let  $N(T)$  be the normalizer of  $T$  in  $G$ . In fact,  $N(T) \subseteq K$ . Under standard technical assumptions [GaVa; Vara, p. 192]) which always hold if  $G$  is connected, the action of  $N(T)$  on  $\mathfrak{t}$  (or  $\mathfrak{t}^*$ ) factors through the Weyl group  $W$  of  $\mathfrak{t}$  in  $K$ . We recall from §§2.9 and 2.10 that  $W$  is generated by reflection in certain hyperplanes  $H_\alpha \subseteq \mathfrak{t}$ . These are the hyperplanes orthogonal to the roots  $\alpha$  of  $\mathfrak{k}$  relative to  $\mathfrak{t}$ ; we will refer to them as *K-root hyperplanes* or *compact root hyperplanes*. The complement of the compact root hyperplanes is called the set of *K-regular elements*. Denote this set by  $\mathfrak{t}_{r,K}$ . The connected components of  $\mathfrak{t}_{r,K}$  are open convex cones. These are permuted simply transitively by  $W$ . The closure of any one of them is called a *K-Weyl chamber*.

By duality,  $W$  acts on  $\mathfrak{t}^*$  also, and we use similar terminology to describe this dual action. The hyperplanes in  $\mathfrak{t}^*$  dual to the  $H_\alpha$  will be denoted by  $H_\alpha^*$ .

Since  $\mathfrak{t}$  is a Cartan subalgebra of  $\mathfrak{g}$ , its complexification  $\mathfrak{t}_\mathbb{C}$  is likewise a Cartan subalgebra in  $\mathfrak{g}_\mathbb{C}$ , the complexification of  $\mathfrak{g}$ . In this context too, we have a Weyl group, which by rather flagrant abuse of notation we will indicate by  $W_\mathbb{C}$ . It is not hard to show that  $W_\mathbb{C}$  preserves  $\mathfrak{t}$ , considered as a real subspace of  $\mathfrak{t}_\mathbb{C} = \mathfrak{t} \oplus i\mathfrak{t}$ . This is because  $\mathfrak{t}$  is characterized as the real subspace of  $\mathfrak{t}_\mathbb{C}$  on which the roots of  $\mathfrak{t}_\mathbb{C}$  in  $\mathfrak{g}_\mathbb{C}$  take on pure imaginary values. As a group of linear transformations of  $\mathfrak{t}$ , the group  $W_\mathbb{C}$  contains

$W$ , and is also generated by reflections in hyperplanes. These hyperplanes will still be denoted by  $H_\alpha$ , where now  $\alpha$  is a root of  $\mathfrak{t}_C$  in  $\mathfrak{g}_C$ . If  $H_\alpha$  is a reflecting hyperplane of  $W_C$ , but not of  $W$ , we call  $H_\alpha$  a *noncompact root hyperplane*. The complement of all the  $H_\alpha$ , compact or noncompact, is the set of *G-regular elements*, or just regular elements, denoted  $\mathfrak{t}_{r,G}$ . We define a notion of Weyl chamber for  $W_C$  as for  $W$ . These are called *G-Weyl chambers*. Evidently each *K*-Weyl chamber contains several *G*-Weyl chambers. (To be precise,  $\#(W_C/W)$  of them.)

Again, we note we can dualize the above discussion to  $\mathfrak{t}^*$ .

We can consider elements of  $\mathfrak{t}^*$  as defining unitary characters on  $\mathfrak{t}$  in the usual way: if  $\lambda \in \mathfrak{t}^*$ , the associated character  $\chi_\lambda$  of  $\mathfrak{t}$  is given by

$$\chi_\lambda(t) = e^{2\pi i \lambda(t)}, \quad t \in \mathfrak{t}.$$

Consider the exponential map  $\exp: \mathfrak{t} \rightarrow T$ . Since  $T$  is commutative,  $\exp$  is a group homomorphism. We call  $\lambda \in \mathfrak{t}^*$  *integral*, or *T-integral* if we need to specify  $T$ , if  $\chi_\lambda$  factors through  $\exp$  to define a character of  $T$ . As in §3.5.5, the set of integral  $\lambda$  form a lattice in  $\mathfrak{t}^*$ , identified via the map  $\lambda \rightarrow \chi_\lambda \circ \exp^{-1}$  to the Pontrjagin dual  $\hat{T}$  of  $T$ . Hence we denote it by  $\hat{T}$ .

Let  $\Sigma$  be the set of roots of  $\mathfrak{t}_C$  in  $\mathfrak{g}_C$ . Elements of  $\Sigma$  are the linear forms defined by eigenvectors for  $\mathfrak{t}_C$  acting on  $\mathfrak{g}_C$  via  $\text{ad}$ . As such they define by exponentiation characters of  $T$ , which we can then identify with elements of  $\hat{T} \subseteq \mathfrak{t}^*$ . As we have mentioned, the roots are elements in  $\mathfrak{t}_C^*$  which most naturally take imaginary values on  $\mathfrak{t}$ . To get them to be elements of  $\mathfrak{t}^*$ , we have essentially multiplied them by  $i$ . (Note the  $i$  in the definition of  $\chi_\lambda$ . This is a different convention from §3.5.5, where we did not multiply by  $i$ , instead we considered that  $\hat{T} \subseteq i\mathfrak{t} \subseteq \mathfrak{t}_C^*$ .)

Let  $\mathcal{C} \subseteq \mathfrak{t}$  be a *G*-Weyl chamber, and let  $\Sigma_{\mathcal{C}}^+ \subseteq \Sigma$  denote the set of roots which take positive values on  $\mathcal{C}$ . We set

$$(3.6.3.5) \quad \rho_{\mathcal{C}} = \frac{1}{2} \sum_{\alpha \in \Sigma_{\mathcal{C}}^+} \alpha.$$

Although  $\rho_{\mathcal{C}}$  clearly depends on the choice of  $\mathcal{C}$ , the difference  $\rho_{\mathcal{C}_1} - \rho_{\mathcal{C}_2}$  for different chambers  $\mathcal{C}_1$  and  $\mathcal{C}_2$  is, by standard results [Bour, Serr1], a sum of elements of  $\Sigma$ . Hence the coset  $\hat{T} + \rho_{\mathcal{C}}$  of  $\hat{T}$  in  $\mathfrak{t}^*$  is independent of the choice of  $\mathcal{C}$ . We denote it by  $\hat{T} + \rho$ .

The final ingredient we need before giving a precise description of the discrete series is an open set  $U \subseteq \mathfrak{g}$ , which is connected, contains the origin, and is such that the exponential map  $\exp: U \rightarrow G$  is a diffeomorphism onto its image. There is a natural maximal choice for  $U$  [Wall2] but we will not try to describe it. Let  $J$  be the Jacobian relating Haar measure on  $G$  with the push-forward via  $\exp$  of Lebesgue measure on  $\mathfrak{g}$ :

$$(3.6.3.6) \quad \int_{\exp U} f(\exp x) d(\exp x) = \int_U f(x) J(x) dx, \quad f \in C_c^\infty(\exp U).$$

Here the measures  $dg = d(\exp x)$  and  $dx$  are the appropriate Haar measures. Clearly  $J$  is smooth, conjugation-invariant, and positive, so it has a well-defined positive square root  $J^{1/2}$ , which also is smooth and conjugation-invariant.

**THEOREM 3.6.3.7** (Harish-Chandra [HaCh20], Rossmann [Ross1]). (a) *Let  $\pi$  be a discrete series representation of the connected semisimple Lie group  $G$ . There is a coadjoint orbit  $\mathcal{O}_\pi \subseteq \mathfrak{g}^*$  such that the character  $\Theta_\pi$  of  $\pi$  may be computed from the Fourier transform of the invariant measure on  $\mathcal{O}_\pi$  by means of the following formula:*

$$\Theta_\pi(f) = \text{trace } \pi(f) = \int_{\mathcal{O}_\pi} \left( \int_U f(\exp x) J^{1/2}(x) \overline{\chi_\lambda(x)} dx \right) d\lambda.$$

Here  $U$  is the neighborhood of 0 in  $\mathfrak{g}$  selected just above,  $f \in C_c^\infty(\exp U)$ , and  $d\lambda$  is the appropriately normalized invariant measure on  $\mathcal{O}_\pi$ .

(b) *The orbit  $\mathcal{O}_\pi$  intersects  $\mathfrak{t}^*$ . The intersection  $\mathcal{O}_\pi \cap \mathfrak{t}^*$  consists of a  $W$ -orbit of points in  $\widehat{T} + \rho$ , and is contained in  $\mathfrak{t}_{r,G}^*$ . The mapping*

$$\pi \rightarrow \mathcal{O}_\pi \rightarrow \mathcal{O}_\pi \cap \mathfrak{t}^*$$

*establishes a bijection between the discrete series of  $G$  and the  $W$ -orbits in  $(\widehat{T} + \rho) \cap \mathfrak{t}_{r,G}^*$ .*

**REMARKS.** (a) The parallel with the compact case (cf. §3.5.6) should be clear.

(b) Harish-Chandra [HaCh14–18, Vara, Wall2] showed that the character  $\Theta_\pi$  of  $\pi$  (in fact of any irreducible representation of  $G$ ) was given by integration against a locally  $L^1$ , conjugation-invariant function, analytic on the regular set. Furthermore, he gave an explicit formula for this function on  $T$ . We will describe his formula. Fix an element  $\lambda \in \mathfrak{t}_{r,G}^*$ , and let  $\mathcal{C}$  be the  $G$ -Weyl chamber containing  $\lambda$ . Define the “denominator function”

$$(3.6.3.8) \quad D = \prod_{\alpha \in \Sigma_{\mathcal{C}}^+} (\chi_{\alpha/2} - \chi_{-\alpha/2}) = \sum_{s \in W_{\mathcal{C}}} \text{sgn}(s) \chi_{s(\rho_{\mathcal{C}})}.$$

Here  $\text{sgn}$  is the standard sign character of  $W_{\mathcal{C}}$ . Although  $D$  depends on  $\mathcal{C}$ , the dependence is weak: the  $D$  associated to a different chamber equals this  $D$  up to a  $\pm$  sign.

The character  $\Theta_\pi$  is expressed, as a function on  $T$ , by the formula

$$(3.6.3.9) \quad \Theta_\pi = \pm \frac{\sum_{s \in W} \text{sgn}(s) \chi_{s(\lambda)}}{D}.$$

Some words about interpreting this formula may be helpful. Because  $\lambda$  is in  $\widehat{T} + \rho$ , which may not equal  $\widehat{T}$ , it may happen that neither the numerator nor the denominator defines a function on  $T$  (i.e., factors through  $\exp: \mathfrak{t} \rightarrow T$ ). However, the quotient may also be written as

$$\Theta_\pi = \pm \frac{\sum_{s \in W} \text{sgn}(s) \chi_{s(\mu+\rho)-\rho}}{\sum_{s \in W_{\mathcal{C}}} \text{sgn}(s) \chi_{s(\rho)-\rho}}, \quad \mu + \rho = \lambda,$$



and both the numerator and denominator of this expression do factor to  $T$ . In particular,  $\Theta_\pi$  factors to  $T$ .

(c) The analogy between formula (3.6.3.9) and the Weyl character formula (cf. (3.5.4.24)) is clear. In fact, although the derivation of formula (3.6.3.9) is very substantially more difficult than that of (3.5.4.24), several key features of the argument are parallel. However, it should be noted that the denominator function  $D$  involves antisymmetrization over the “complex Weyl group”  $W_C$ , and so has zeros along all the  $H_\alpha$ ,  $\alpha \in \Sigma$ , whereas the numerator of formula (3.6.3.9) involves only a sum over the “real Weyl group”  $W$ , and cannot be expected to vanish on the noncompact root hyperplanes. Thus the character  $\Theta_\pi$  will have singularities.

(d) As in the case of compact groups (cf. §3.5.6), the numerator in formula (3.6.3.9) is provided by  $\hat{\mathcal{O}}_\lambda$ , the Fourier transform of the orbital integral, and the denominator  $D$  is provided by the Jacobian factor  $J^{1/2}$ .

(e) The explicit formula (3.6.3.9) is due to Harish-Chandra [HaCh19, 20]. In the course of his argument, he established a less precise version of the orbital integral formula for  $\Theta_\pi$  given in Theorem 3.6.3.7: the expression for  $\Theta_\pi$  was allowed to be a linear combination of the orbital integrals coming from coadjoint orbits through the full  $W_C$ -orbit  $W_C(\lambda)$ , rather than a single orbit. The more refined result, that only one orbit is involved in  $\Theta_\pi$ , was established by Rossmann [Ross1]. The key step in Rossmann’s analysis was an analog for noncompact  $G$  of Theorem 3.5.6.8.

There is a story that Harish-Chandra had considered whether a single-orbit expression like that of Theorem 3.6.3.7 might be valid, but was led to abandon such a hope by erroneous computations for the example of  $\mathrm{SL}_2(\mathbf{R})$ . If this is true, it gives a rare instance where Harish-Chandra’s intuition, which guided him so well through the deep forest of semisimple harmonic analysis, led him astray. Particularly for  $\mathrm{SL}_2(\mathbf{R})$ , the distinctive analytic features of the various types of discrete series representations are so well mirrored by the geometry of the different type of elliptic coadjoint orbits, that the one-representation/one-orbit hypothesis seems, at least with hindsight, very plausible.

(f) The correct normalization of the invariant measure  $d\lambda$  on  $\mathcal{O}_\pi$  is given by the same universal normalization as in the nilpotent case, defined intrinsically in terms of the symplectic structure on  $\mathcal{O}_\lambda$ , as Kirillov [Kiri] suggested should hold.

(g) Rossmann [Ross2] has also shown that an orbital integral formula like that of Theorem 3.6.3.7 is valid for all tempered representations. In this generality, the bijectivity of the correspondence between representations and orbits breaks down. Sometimes different representations have characters which agree near the identity in  $G$ , and so correspond to the same orbit, and sometimes several orbits are needed to give the character of one representation. (This latter situation is exceptional, however.) These phenomena can already be seen in  $\mathrm{SL}_2(\mathbf{R})$  [Ross2].

(h) Duflo and Vergne [DuVe1] have showed that the orbital picture yields a nice interpretation of the Plancherel formula, yielding a sort of “Poisson-Plancherel” formula for  $G$ , partaking of the nature of both the classical Poisson and Plancherel formulas.

(i) The character formula (3.6.3.9) suggests what the restriction to  $K$  of the discrete series representation  $\pi$  should look like. We have observed that the numerator of (3.6.3.9) looks like the numerator of the Weyl character formula for  $K$ , while the denominator is the Weyl denominator for  $G$ , which contains the Weyl denominator for  $K$  as a factor. Decompose

$$(3.6.3.10) \quad \Sigma_{\mathcal{G}}^+ = \Sigma_c^+ \cup \Sigma_n^+,$$

where the  $\Sigma_c^+$  are the roots of  $\mathfrak{t}_c$  acting on  $\mathfrak{k}_c$  (the compact roots), and  $\Sigma_n^+$  are the remaining roots (the noncompact roots). Then

$$(3.6.3.11) \quad D_K = \prod_{\alpha \in \Sigma_c^+} (\chi_{\alpha/2} - \chi_{-\alpha/2})$$

is the Weyl denominator for  $K$ , and

$$(3.6.3.12) \quad D_n = \prod_{\alpha \in \Sigma_n^+} (\chi_{\alpha/2} - \chi_{-\alpha/2}) = \pm \left( \prod_{\alpha \in \Sigma_n^+} \chi_{-\alpha/2} \right) \left( \prod_{\alpha \in \Sigma_n^+} (1 - \chi_{\alpha}) \right)$$

is the quotient  $D/D_K$ . In analogy with the geometric series

$$\frac{1}{1-r} = \sum_{m=0}^{\infty} r^m$$

we may formally expand

$$(3.6.3.13) \quad \left( \prod_{\alpha \in \Sigma_n^+} (1 - \chi_{\alpha}) \right)^{-1} = \sum p_n(\gamma) \chi_{\gamma},$$

where  $\gamma$  is any weight of the form  $\sum_{\alpha \in \Sigma_n^+} n_{\alpha} \alpha$ , with nonnegative integers  $n_{\alpha}$ , and  $p_n(\gamma)$ , the “partition function” for  $\Sigma_n^+$ , is the number of ways of expressing  $\gamma$  as such a sum. Note also that decomposition (3.6.3.10) induces a parallel decomposition (cf. formula (3.6.3.5))

$$(3.6.3.14) \quad \rho_{\mathcal{G}} = \rho_c + \rho_n.$$

With this notation, we can write  $\prod_{\alpha \in \Sigma_n^+} \chi_{-\alpha/2} = \chi_{-\rho_n}$ . It is easy to see that  $D_n$  is invariant under  $W$ , the Weyl group of  $K$ . Thus, for any  $s \in W$ , we may write formally

$$D_n = \chi_{-s\rho_n} \left( \sum p_n(\gamma) \chi_{s(\gamma)} \right)^{-1}.$$

Plugging this in to expression (3.6.3.9) gives us

$$(3.6.3.15) \quad (D_K)^{-1} \sum_{s \in W} \operatorname{sgn} s \sum_{\gamma} p_n(\gamma) \chi_{s(\lambda + \rho_n + \gamma)}.$$

If we count the number of times that a given  $K$ -dominant weight  $\mu$  occurs in this sum, we obtain

$$(3.6.3.16) \quad \left( \sum_{\mu \in \mathcal{E}_K} \left( \sum_{s \in W} (\operatorname{sgn} s) p_n(s\mu - \lambda - \rho_n) \right) \right) \operatorname{ch}_K(\mu),$$

where

$$\operatorname{ch}_K(\lambda) = \sum_{s \in W} \frac{\operatorname{sgn}(s) \chi_s(\mu)}{D_K}$$

is the character of the representation of  $K$ , of highest weight  $\mu - \rho_c$ . Expression (3.6.3.16) leads us to suspect that the multiplicity of the representation of  $K$  with highest weight  $\mu$  would occur in the discrete series representation  $\pi$  with multiplicity

$$(3.6.3.17) \quad \sum_{s \in W} (\operatorname{sgn} s) p_n(s(\mu + \rho_c) - \lambda - \rho_n).$$

This formula is indeed true [HeSc1, Knap2, Wall2]. It is known as Blattner's formula. A cohomological explanation and a generalization was found by Zuckerman in the context of his derived functor construction [Knap1, Voga1, 2, Wall2] (cf. §3.6.5).

(j) Study of the geometry of Weyl chambers shows that the norm (with respect to the Killing form) of  $\lambda + \rho_n + \gamma$  is always greater than the norm of  $\lambda + \rho_n$  if  $\gamma \neq 0$ . Hence the only term in the sum (3.6.3.15) which yields  $\chi_{\lambda + \rho_n}$  is  $\gamma = 0$ . It follows from Blattner's formula (3.6.3.17) that the representation of  $K$  with highest weight  $\lambda + \rho_n - \rho_c = \lambda + \rho_\varphi - 2\rho_c$  occurs in the discrete series  $\pi$  of formula (3.6.3.9) with multiplicity one. This  $K$ -representation is known as the *lowest  $K$ -type* of  $\pi$ . Vogan [Voga2, 4] has shown that every representation of  $G$  contains a  $K$ -type, which is lowest in a certain sense, with multiplicity one.

(k) Even more than in the Weyl character formula for compact groups, one sees the necessity for various " $\rho$  shifts" to make parameters match up, particularly for the lowest  $K$ -type. To make sense of these, one should keep in mind that what is controlling everything is the infinitesimal character, which is computed by a  $\rho$ -shift coming from the Harish-Chandra homomorphism (cf. Theorem 3.5.5.23). In discussing lowest  $K$ -types, we must deal with the  $\rho$ -shifts for both  $K$  and  $G$ .

3.6.4. CLASSIFICATION. In the 1970s two classifications of the irreducible admissible representations of a semisimple Lie group were given, one by R. Langlands [Lgld4], and another by D. Vogan [Voga4]. Somewhat earlier, a classification of the representations of complex groups had been given by Zhelobenko [Zhel2] and Parthasarathy-Rao-Varadarajan [PaRV]. Langlands' classification was similar in flavor. Although the classifications of Langlands and Vogan seem to be based on rather different principles—Langland's on the asymptotic behavior of matrix coefficients, and Vogan's on the behavior

under restriction to the maximal compact subgroup  $K$ , in particular the existence of a “lowest  $K$ -type”—they were successfully combined by Vogan in [Voga2]. The result is a description of a standard realization of a given irreducible admissible representation, supplemented by information about its restriction to  $K$ .

In [BeBe], J. Bernstein and A. Beilinson announced a classification based on another circle of ideas. In particular it relies heavily on the theory of “ $D$ -modules”—modules for the sheaf of differential operators on a manifold. Despite its exotic origins, the Bernstein-Beilinson classification has a strong geometric flavor that gives it considerable appeal. The project of comparing and coordinating the features of the three classification schemes has been pursued in recent years by a group including H. Hecht, D. Milicic, W. Schmid, and J. Wolf [HMSW].

In this section we will give a brief description of the Langlands classification, which essentially describes representations in terms of standard embeddings in the principal series.

Let  $G$  be a semisimple Lie group, and let  $P \subseteq G$  be a parabolic subgroup (cf. §A.2.4). Write  $P = MAN$  as in formula (3.6.1.5). The abelian group  $A$  acts on the Lie algebra  $\mathfrak{n}$  of  $N$  by the conjugation action  $\text{Ad } A$ . Under this action  $\mathfrak{n}$  decomposes into a direct sum of eigenspaces

$$(3.6.4.1) \quad \mathfrak{n} = \sum_{\alpha} \mathfrak{n}_{\alpha}, \quad \alpha \in \Sigma^{+}(A, n),$$

where the  $\alpha$  are the roots of  $A$  acting on  $\mathfrak{n}$ :

$$(3.6.4.2) \quad \text{Ad } a(n) = \alpha(a)n, \quad a \in A, n \in \mathfrak{n}_{\alpha}.$$

Evidently from their definition, the  $\alpha$  are homomorphisms from  $A$  to  $\mathbb{C}^{\times}$ . In fact  $\text{Ad } A$  is a real-diagonalizable action, so that the  $\alpha$  have images in  $\mathbb{R}^{+\times}$ . Write

$$(3.6.4.3) \quad A^{+} = \{a \in A : \alpha(a) \geq 1, \text{ all } \alpha\}.$$

Note that  $A^{+}$  is a closed semigroup of  $A$ , with nonempty interior  $A^{+0}$ . Dually, write

$$(3.6.4.4) \quad (\hat{A}^{\mathbb{C}})^{+} = \{\psi \in \hat{A}^{\mathbb{C}} : |\psi(a)| > 1, \text{ all } a \in A^{+0}\}.$$

Note that, in contrast to  $A^{+}$ , our definition makes  $(\hat{A}^{\mathbb{C}})^{+}$  an *open* semigroup in  $\hat{A}^{\mathbb{C}}$ . This slight inconsistency will reduce by a little the total amount of notation that we need.

**THEOREM 3.6.4.5 (Langlands classification).** *Fix a minimal parabolic subgroup  $P_0$  inside the connected semisimple Lie group  $G$ . Let  $P$  be a parabolic subgroup containing  $P_0$ , with Langlands decomposition  $P = MAN$ . Let  $\sigma$  be an irreducible tempered (in particular, unitary) representation of  $M$ , and  $\psi \in (\hat{A}^{\mathbb{C}})^{+}$ . Then the principal series representation  $\text{P.S.}(\sigma, \psi)$  has a unique*

*irreducible quotient*  $L.Q.(\sigma, \psi)$ . Each irreducible admissible representation of  $G$  is isomorphic to a unique  $L.Q.(\sigma, \psi)$ .

REMARKS. (a) For generic  $\psi$ , more specifically, for  $\psi$  not satisfying certain integrality conditions [SpVo], the representations  $P.S.(\sigma, \psi)$  are irreducible, i.e.,  $L.Q.(\sigma, \psi) = P.S.(\sigma, \psi)$ . However, the nature of  $L.Q.(\sigma, \psi)$  can change drastically as  $\psi$  varies. The difficulty of describing  $L.Q.(\sigma, \psi)$  more explicitly than is done in Theorem 3.6.4.5 is one reason why various outstanding problems, e.g., the classification of the unitary dual, remain unsolved. Thus while Theorem 3.6.4.5 gives us a place to put each representation, it does not provide us with a complete picture of the structure of the representations.

(b) A version of Theorem 3.6.4.5 first appeared in [Lgld4]. Its proof relied on results of an unpublished manuscript of Harish-Chandra on asymptotic expansions of matrix coefficients [HaCh13]. Harish-Chandra's results were refined and simplified by several authors; we refer in particular to [CaMi]. Relatively streamlined and complete accounts of these matters, as well as most of the rest of semisimple representation theory, are available in the texts [Knap2] and [Wall2].

(c) In fact, the term "Langlands classification" for Theorem 3.6.4.5 is something of a misnomer. Langlands' goal in [Lgld4] was a different classification (see §4.2).

(d) In remark (i) following Theorem 3.6.3.7, we noted that if  $\sigma$  is a discrete series representation, then the restriction of  $\sigma$  to the maximal compact subgroup  $K$  contains a certain minimal  $K$ -type with multiplicity one. Vogan [Voga4] generalized the notion of minimal  $K$ -type to apply to any irreducible representation, and showed that his minimal  $K$ -type always occurs with multiplicity one. The  $K$ -module structure of the representations  $P.S.(\sigma, \psi)$  is independent of  $\psi$ , so they all have the same minimal  $K$ -type  $\mu_\sigma$ . Vogan shows [Voga2] that  $\mu_\sigma$  survives in  $L.Q.(\sigma, \psi)$ . This additional information about  $L.Q.(\sigma, \psi)$  allows one to characterize it as a subquotient of a broader class of principal series representations.

EXAMPLE. Parametrize the spherical principal series of  $SL_2(\mathbf{R})$  by  $\lambda \in \mathbf{C}$ , as in formula (3.6.1.8). Then the unitary principal series, which are all tempered, lie on the imaginary axis. They are all irreducible, and the representations labeled by  $\lambda$  and by  $-\lambda$  are mutually equivalent. The representations associated to  $P_0$  in Theorem 3.6.4.5 are those in the right half plane. All are irreducible, except for  $\lambda = 1, 3, 5, 7, \dots$ . For these, the Langlands quotient is the finite-dimensional representation of dimension  $\lambda$  (cf. §3.1). The remaining constituents (there are two) of  $P.S.(1, 2m+1)$  are discrete series, hence get counted among the tempered representations. The story is similar for the nonunitary principal series  $P.S.(\tilde{\epsilon}, \tilde{\alpha}^\lambda)$  ( $\tilde{\epsilon}$  here being the nontrivial character of  $M_0$ ; cf. formula (3.6.1.8)), except that  $P.S.(\tilde{\epsilon}, 1)$ , the point of symmetry of the unitary principal series, reduces

into two pieces, and the points of reducibility in the right half plane are at the even integers  $\lambda = 2, 4, 6, 8, \dots$ , the Langlands quotient again being the finite-dimensional representation of dimension  $\lambda$ .

The representations  $P.S.(\tilde{e}, \tilde{a}^\lambda)$  for  $\operatorname{Re} \lambda < 0$  are contragredient to  $P.S.(\tilde{e}, \tilde{a}^{-\lambda})$  (cf. §A.1.10). If  $P.S.(\tilde{e}, \tilde{a}^\lambda)$  is irreducible, it is equivalent to  $P.S.(\tilde{e}, \tilde{a}^{-\lambda})$ . If it is reducible, it has the same constituents as  $P.S.(\tilde{e}, \tilde{a}^{-\lambda})$ , but the finite-dimensional representation is now a subrepresentation, and the two discrete series are quotients.

**3.6.5. DERIVED FUNCTOR MODULES.** Part of the fascination of semisimple harmonic analysis is the strong interaction of algebra and analysis that it affords. After his initial, heavily algebraic, papers on foundational issues in representation theory, Harish-Chandra's methods became more and more analytic, culminating in the construction of the discrete series via a deep study of character theory [HaCh13-20]. Then, beginning in the late 1960s, the work of Dixmier [Dixm1], Kostant [Kost8], and the Gelfand school (cf. [BGG1-3]) reemphasized the algebraic aspects of the theory.

In the late 1960s, there had been efforts, by W. Schmid [Schm1-3], and others [OkOz, Hott, Part1], to “realize the discrete series,” i.e., construct specific vector spaces on which a semisimple group could act in a natural way, and such that the resulting representation of  $G$  was a given discrete series representation (cf. §3.6.3). G. Zuckerman, considering Schmid's work from the more algebraic point of view, invented a flexible and fruitful purely algebraic method for constructing representations. Parthasarathy [Part2] had an idea for a similar construction, and Enright-Varadarajan [EnVa] also proposed an interesting, though conceptually less transparent, algebraic method for constructing representations. After several simplifications and refinements [Voga1, EnWa, DuVe2, KnVo, Wign2, Wall2], Zuckerman's method, which has become known as the “derived functor construction,” has become a standard tool for constructing representations. In particular, it can be used to give an a priori construction of the discrete series [Wall2]. (Proof of “exhaustion”—that all discrete series are so realized—still, however, requires analysis, especially character theory.) It also yields other interesting classes of nontempered unitary representations, e.g., the representations with nontrivial  $(\mathfrak{g}, \mathfrak{k})$ -cohomology [VoZu]. We will give a brief description of Zuckerman's idea. A nice overview is given in [Voga1], a leisurely introductory treatment is in [Knap2], and detailed expositions are given in [Voga2] and [Wall2].

The inspiration behind the derived functor construction is epistemologically interesting: it consists in taking seriously what might have seemed merely a technical convenience—the notion of  $(\mathfrak{g}, K)$ -module. The general goal of the algebraic approach is to replace a  $G$ -module by a suitable  $\mathcal{U}(\mathfrak{g})$ -module which can serve as its proxy, i.e., that will mirror its essential features. A first guess might be the subspace of smooth vectors (cf. §A.1.13) but this is of uncountable dimension, hence too large to be studied algebraically. At

the outset, however, Harish-Chandra had shown [HaCh2] that if  $\rho$  is a t.c.i. (§A.1.7) (or quasisimple) irreducible representation of a semisimple group  $G$  with maximal compact subgroup  $K$  on a Banach space  $V$ , then the subspace  $V_K$  of  $K$ -finite vectors (vectors contained in finite-dimensional,  $K$ -invariant subspaces) is invariant under the action of the Lie algebra  $\mathfrak{g}$ . Hence  $V_K$  is a module for  $\mathfrak{g}$ . It is also obviously a module for  $K$ , and the two module structures are compatible in some obvious ways:

- (3.6.5.1) (i) The differential of the action of  $K$  is obtained by restriction of the  $\mathfrak{g}$ -action to the Lie subalgebra  $\mathfrak{k} \subseteq \mathfrak{g}$  corresponding to  $K$ .  
 (ii)  $K$  normalizes  $\mathfrak{g}$  inside  $\text{End } V_K$ , and conjugation by  $K$  in  $\text{End } V_K$  yields the usual adjoint action  $\text{Ad } K$  on  $\mathfrak{g}$ .

These properties were enshrined by Gelfand [Gelf] in a formal definition of what is now usually called a  $(\mathfrak{g}, K)$ -module. Our space  $V_K$  also has the property that each  $K$ -isotypic subspace  $V_\mu$ ,  $\mu \in \hat{K}$ , is finite dimensional. A  $(\mathfrak{g}, K)$ -module with this property is called an *admissible*  $(\mathfrak{g}, K)$ -module or *Harish-Chandra module*.

For a given irreducible representation of  $G$  on  $V$ , the associated Harish-Chandra module  $V_K$  is convenient to work with. For example, for  $\text{SL}_2(\mathbf{R})$  it is an easy, pleasant exercise to determine all possible irreducible  $(\mathfrak{sl}_2(\mathbf{R}), \text{SO}_2)$  modules, and to check which ones could carry an invariant inner product. This was the method of Bargmann [Barg1] at the very beginning of semisimple representation theory. Further,  $V_K$  captures much of what we want to know about  $V$ : for example, Harish-Chandra showed [HaCh3] that  $V$  could be unitary if and only if  $V_K$  carries an invariant positive-definite Hermitian form, and, in that case,  $V_K$  determines  $V$  up to unitary equivalence.

Thus, one might propose Harish-Chandra modules as a technically convenient class of  $\mathfrak{g}$ -modules with which to work in order to study algebraic aspects of representation theory. However,  $(\mathfrak{g}, K)$ -modules also appear to be rather awkward from certain points of view. Some people might be put off by the loss of symmetry entailed by the choice of a particular maximal compact group  $K$ . A serious technical problem is presented by induction. Induction is a basic method for constructing representations, so we would like to have an algebraic version of it. There is in fact a standard notion of induction in the category of associative algebras, defined in terms of tensor products. Thus suppose  $\mathfrak{g}$  is a Lie algebra,  $\mathfrak{h} \subseteq \mathfrak{g}$  is a subalgebra, and  $V$  is an  $\mathfrak{h}$ -module. Define

$$(3.6.5.2) \quad \text{ind}_{\mathfrak{h}}^{\mathfrak{g}} V = \mathcal{U}(\mathfrak{g}) \otimes_{\mathcal{U}(\mathfrak{h})} V.$$

We recall [Jaco2] that  $\mathcal{U}(\mathfrak{g}) \otimes_{\mathcal{U}(\mathfrak{h})} V$  is the quotient of the usual tensor product  $\mathcal{U}(\mathfrak{g}) \otimes V$  by the subspace spanned by tensors of the form  $xy \otimes v - x \otimes y(v)$ ,  $x \in \mathcal{U}(\mathfrak{g})$ ,  $y \in \mathcal{U}(\mathfrak{h})$ ,  $v \in V$ . (In fact, it suffices to take  $y \in \mathfrak{h}$ .) The action of  $\mathcal{U}(\mathfrak{g})$  on  $\text{ind}_{\mathfrak{h}}^{\mathfrak{g}} V$  is the push-down of left multiplication on  $\mathcal{U}(\mathfrak{g})$ .

A variant notion, more or less dual to (3.6.5.2), also exists. It is sometimes called *production*, (though some feel the terminology for induction and production should be reversed) and is defined by

$$(3.6.5.3) \quad \text{pro}_{\mathfrak{h}}^{\mathfrak{g}} V = \text{Hom}_{\mathfrak{h}}(\mathcal{U}(\mathfrak{g}), V).$$

Here the action of  $\mathfrak{h}$  on  $\mathcal{U}(\mathfrak{g})$  is via multiplication by  $\mathfrak{h}$  on the left. The action of  $\mathfrak{g}$  on  $\text{pro}_{\mathfrak{h}}^{\mathfrak{g}}(V)$  is via multiplication on the right in  $\mathcal{U}(\mathfrak{g})$ . If  $M$  is a subgroup of  $G$ , whose Lie algebra  $\mathfrak{m}$  is contained in  $\mathfrak{h}$ , and  $V$  is an  $(\mathfrak{h}, M)$ -module, we let  $M$  act on  $\text{pro}_{\mathfrak{h}}^{\mathfrak{g}} V$  by the recipe

$$m(f)(u) = m(f(\text{Ad } m(u))), \quad m \in M, f \in \text{pro}_{\mathfrak{h}}^{\mathfrak{g}} V, u \in \mathcal{U}(\mathfrak{g}).$$

We then replace  $\text{pro}_{\mathfrak{h}}^{\mathfrak{g}} V$  by the subspace of its  $M$ -finite vectors and so obtain a  $(\mathfrak{g}, M)$ -module.

Both induction [Wall2, Knap1] and production [Voga1, 2, Knap1], are used in accounts of derived functor modules. Other conventions concerning the derived functor construction also vary from author to author, necessitating an annoying, if in principle straightforward, translation process to compare results. We follow the conventions of [Voga1]; in particular, we use production.

The production process (3.6.5.3) converts  $\mathfrak{h}$ -modules to  $\mathfrak{g}$ -modules, but it is unlikely that it will yield modules which are spanned by  $K$ -finite (or, what is more reasonable to discuss at this stage,  $\mathfrak{k}$ -finite) vectors, or which even contain any  $\mathfrak{k}$ -finite vectors at all. This would seem to be a serious drawback of the  $(\mathfrak{g}, K)$ -module formalism. However, Zuckerman saw how to make a virtue of necessity, and converted this seeming liability into a construction method that is subtler than production, but still is fairly manageable.

The most obvious way to associate a  $(\mathfrak{g}, K)$ -module to a  $\mathfrak{g}$ -module is to look at the submodule of  $\mathcal{U}(\mathfrak{k})$ -finite vectors, then exponentiate to get an action of  $K$  on this submodule. (There is further fussing necessary if  $K$  is not connected and simply connected [Voga1, 2, Knap1, Wall2]. We will ignore this fussing. Thus we will actually be discussing  $(\mathfrak{g}, \mathfrak{k})$ -modules (the definition of which is hoped to be obvious) rather than  $(\mathfrak{g}, K)$ -modules.) This procedure, however, may well result in the trivial module. Zuckerman observed, however, that the process of passing from a  $\mathfrak{g}$ -module to the  $(\mathfrak{g}, \mathfrak{k})$ -module of  $\mathfrak{k}$ -finite vectors is a functor, and moreover, it is a left-exact functor. Thus we have the possibility of taking its (right) derived functors; and even if the module we start with has no  $K$ -finite vectors, one of the higher derived functors may be nontrivial. This does in fact happen in interesting cases.

Thus if  $V$  is a  $\mathfrak{g}$ -module, we can define

$$(3.6.5.4) \quad \Gamma(V) = \{v \in V : \dim \mathcal{U}(\mathfrak{k})(v) < \infty\}.$$

(As noted above, one must give a slightly more complicated definition of  $\Gamma$  if  $K$  is not connected or not simply connected.) The derived functors of  $\Gamma$



will be denoted  $\Gamma^i$ . A curious point about  $\Gamma$  is that, while  $\Gamma$  is a functor on  $\mathfrak{g}$ -modules, it depends only on the  $\mathfrak{k}$ -module structure of these modules. Further, we can express  $\Gamma$  in terms of  $\mathfrak{k}$ -fixed vectors in certain auxiliary modules, viz.

$$(3.6.5.5) \quad \Gamma(V) = \sum_{\sigma \in \hat{\mathfrak{k}}} (V \otimes \sigma^*)^{\mathfrak{k}} \otimes \sigma.$$

Here  $\hat{\mathfrak{k}}$  is the collection of irreducible finite-dimensional representations of  $\mathfrak{k}$ —the same as  $\hat{K}$  if  $K$  is connected and simply connected—and  $(V \otimes \sigma^*)^{\mathfrak{k}}$  indicates the  $\mathfrak{k}$ -invariant vectors in  $V \otimes \sigma^*$ .

Since formula (3.6.5.5) expresses  $\Gamma(V)$  in terms of the functor of  $K$ -fixed vectors, it suggests the  $\Gamma^i$  should be expressible in terms of the derived functors of the  $\mathfrak{k}$ -fixed vector functor, which is Lie algebra cohomology [BoWa, Jaco1, Knap1]. This is not immediate, since the construction of derived functors depends on injective resolutions, and the notion of injectiveness depends on the category in which one is working. However, we can construct injective (or projective) resolutions of  $\mathfrak{g}$ -modules with injective (or projective) modules of the form  $\text{Hom}(\mathcal{U}(\mathfrak{g}), Y)$  (or  $\mathcal{U}(\mathfrak{g}) \otimes Y$ ), where  $Y$  is a  $\mathfrak{g}$ -module. Since  $\mathcal{U}(\mathfrak{g})$  is free as a  $\mathcal{U}(\mathfrak{k})$ -module, by the Poincaré-Birkhoff-Witt Theorem [Jaco1, Serr2, Hump], one can check that these are injective (or projective) as  $\mathfrak{k}$ -modules also. It follows that

$$(3.6.5.6) \quad \Gamma^i(V) \simeq \sum_{\sigma \in \hat{\mathfrak{k}}} H^i(\mathfrak{k}, V \otimes \sigma^*) \otimes \sigma,$$

where  $H^i(\mathfrak{k}, X)$  is the  $i$ th Lie algebra cohomology of the  $\mathfrak{k}$ -module  $X$ .

There is also a relative version of (3.6.5.6). Suppose we are given a subalgebra  $\mathfrak{m} \subseteq \mathfrak{k}$ , and we start with a  $\mathfrak{g}$ -module  $V$  which is already  $\mathfrak{m}$ -finite (a  $(\mathfrak{g}, \mathfrak{m})$ -module). Then it is appropriate to work inside the category of  $\mathfrak{m}$ -finite  $\mathfrak{g}$ -modules, and the relevant derived functors of the  $\mathfrak{k}$ -fixed vector functor are the relative Lie algebra cohomology groups [BoWa, Knap1]. Thus, if  $V$  is an  $\mathfrak{m}$ -finite  $\mathfrak{g}$ -module, or a  $(\mathfrak{g}, M)$ -module for some  $\mathfrak{m} \subseteq \mathfrak{k}$ , then

$$(3.6.5.7) \quad \Gamma^i(V) \simeq \sum_{\sigma \in \hat{\mathfrak{k}}} H^i(\mathfrak{k}, \mathfrak{m}, V \otimes \sigma^*) \otimes \sigma,$$

where  $H^i(\mathfrak{k}, \mathfrak{m}, X)$  is the  $i$ th  $(\mathfrak{k}, \mathfrak{m})$ -relative cohomology group of the  $\mathfrak{m}$ -finite  $\mathfrak{k}$ -module  $X$ .

This describes the structure of  $\Gamma^i(V)$  as a  $\mathfrak{k}$ -module, but of course our interest in it is as a  $\mathfrak{g}$ -module. A simple observation [EnWa, DuVe, KnVo, Wign2] allows this to be done simply. Note that, if  $V$  is a  $\mathfrak{k}$ -module, and  $X$  is a finite-dimensional  $\mathfrak{k}$ -module, then

$$(3.6.5.8) \quad \Gamma(X \otimes V) \simeq X \otimes \Gamma(V).$$

In fact, this isomorphism is a natural equivalence of functors. Further, tensoring with  $X$  is exact. It follows easily that

$$(3.6.5.9) \quad \Gamma^i(X \otimes V) \simeq X \otimes \Gamma^i(V),$$

and that this also is an equivalence of functors, i.e., this isomorphism is natural.

Now suppose  $V$  is a  $\mathfrak{g}$ -module. This means we have a mapping

$$(3.6.5.10) \quad \begin{aligned} \mathfrak{g} \otimes V &\xrightarrow{\mu} V, \\ \mu(x \otimes v) &= x(v), \quad x \in \mathfrak{g}, v \in V. \end{aligned}$$

Clearly the mapping  $\mu$  of (3.6.5.10) determines the  $\mathfrak{g}$ -module structure of  $V$ . Furthermore, the fact that  $\mu$  defines a  $\mathfrak{g}$ -module structure on  $V$  can be expressed solely in terms of  $\mu$ : we should have the identity

$$(3.6.5.11) \quad \mu(x_1 \otimes \mu(x_2 \otimes v)) - \mu(x_2 \otimes \mu(x_1 \otimes v)) = \mu([x_1, x_2] \otimes v)$$

as mappings from  $\mathfrak{g} \otimes \mathfrak{g} \otimes V$  to  $V$ . Combining these remarks with the previous paragraph, we see that the  $\mathfrak{g}$ -module structure (3.6.5.10) on  $V$  induces a  $\mathfrak{g}$ -module structure on  $\Gamma^i(V)$  for all  $i$ . It is very plausible that this  $\mathfrak{g}$ -module structure is the one that should be carried by  $\Gamma^i(V)$ , and it is indeed so (cf. references preceding (3.6.5.8)).

Now let us introduce the modules to which we wish to apply the  $\Gamma^i$ . Consider  $x \in \mathfrak{k}$ . Let  $\mathfrak{l}$  be the centralizer of  $x$  in  $\mathfrak{g}$ . The complexification  $\mathfrak{l}_{\mathbb{C}}$  is a Levi component of a parabolic subalgebra of  $\mathfrak{g}_{\mathbb{C}}$ . More precisely, since  $x$  is in  $\mathfrak{k}$ ,  $\text{ad } x$  will act on  $\mathfrak{g}_{\mathbb{C}}$  with purely imaginary eigenvalues. Let  $\mathfrak{n}^+ \subseteq \mathfrak{g}_{\mathbb{C}}$  be the sum of the  $\text{ad } x$  eigenspaces with eigenvalues with positive imaginary part, and let  $\mathfrak{n}^-$  be the complex conjugate of  $\mathfrak{n}^+$ , the sum of  $\text{ad } x$ -eigenspaces whose eigenvalues have negative imaginary part. Then

$$(3.6.5.12) \quad \mathfrak{g}_{\mathbb{C}} = \mathfrak{n}^- \oplus \mathfrak{l}_{\mathbb{C}} \oplus \mathfrak{n}^+$$

and  $\mathfrak{q} = \mathfrak{l}_{\mathbb{C}} \oplus \mathfrak{n}^+$  is a parabolic subgroup of  $\mathfrak{g}_{\mathbb{C}}$ . Let  $\theta$  be the Cartan involution on  $\mathfrak{g}$ : the automorphism of order 2 whose fixed point set is  $\mathfrak{k}$  (cf. §A.2). Since  $x \in \mathfrak{k}$ , it is clear that  $\mathfrak{l}$  and  $\mathfrak{n}^{\pm}$  are invariant under  $\theta$ . The parabolic subalgebra  $\mathfrak{q}$  is therefore called a  $\theta$ -stable parabolic subalgebra.

Let  $\mathfrak{g} = \mathfrak{l}_{\mathbb{C}} \oplus \mathfrak{n}^+$  be the  $\theta$ -stable parabolic subalgebra defined by  $x \in \mathfrak{k}$ , as above. Let  $m = \dim \mathfrak{n}^+$ . The adjoint action of  $\mathfrak{l}_{\mathbb{C}}$  on  $\mathfrak{n}^+$  gives rise to a one-dimensional action of  $\mathfrak{l}_{\mathbb{C}}$  on  $\Lambda^m(\mathfrak{n}^+)$ , the top exterior power of  $\mathfrak{n}^+$ . Denote by  $2\rho_{\mathfrak{q}}$  the weight in  $\mathfrak{l}_{\mathbb{C}}^*$  defined by  $\Lambda^m(\mathfrak{n}^+)$ . Note that  $2\rho_{\mathfrak{q}}$  is a formal analog of the modular function of a real parabolic subgroup (cf. §A.1.15 and formula (3.5.5.16)) and we use it in the same way: to normalize induction by twisting with  $\rho$  before induction (cf. equation (3.6.1.3)). Let  $C_{\rho_{\mathfrak{q}}}$  be a one-dimensional module on which  $\mathfrak{l}_{\mathbb{C}}$  acts by the weight  $\rho$ . (In some sense,  $C_{\rho_{\mathfrak{q}}} = (\Lambda^m(\mathfrak{n}^+))^{1/2}$ .)

Let  $Z$  be an  $(\mathfrak{l}, \mathfrak{l} \cap \mathfrak{k})$ -module. Extend by complex linearity the action of  $\mathfrak{l}$  to  $\mathfrak{l}_{\mathbb{C}}$ . Extend the action of  $\mathfrak{l}_{\mathbb{C}}$  to an action of  $\mathfrak{q}$  by letting  $\mathfrak{n}^+$  act trivially. Define

$$(3.6.5.13) \quad \mathcal{R}^j(Z) = \Gamma^j(\text{pro}_{\mathfrak{q}}^{\mathfrak{g}_{\mathbb{C}}}(Z \otimes C_{\rho_{\mathfrak{q}}})_{(\mathfrak{k} \cap \mathfrak{l})}), \quad j \geq 0.$$

Here the sub- $(\mathbf{k} \cap l)$  means the submodule of  $(\mathbf{k} \cap l)$ -finite vectors. The  $\mathcal{R}^j$  are functors which transform  $(l, l \cap \mathbf{k})$ -modules into  $(\mathbf{g}, \mathbf{k})$ -modules (or, more carefully done,  $(l, L \cap K)$ -modules into  $(\mathbf{g}, K)$ -modules). The  $\mathcal{R}^j(Z)$  are the *derived functor modules*, and the functors  $\mathcal{R}^j$  are often referred to as *cohomological induction*. The usual yoga about derived functors [Jaco2, Lang3] shows that if

$$(3.6.5.14a) \quad 0 \rightarrow Z' \rightarrow Z \rightarrow Z'' \rightarrow 0$$

is a short exact sequence of  $(l, l \cap \mathbf{k})$ -modules, then the  $\mathcal{R}^j(Z)$  can be organized into a long exact sequence

$$(3.6.5.14b) \quad \rightarrow \mathcal{R}^j(Z) \rightarrow \mathcal{R}^j(Z'') \rightarrow \mathcal{R}^{j+1}(Z') \rightarrow \mathcal{R}^{j+1}(Z) \rightarrow .$$

Further, the  $\mathcal{R}^j$  are compatible with the Harish-Chandra homomorphism (cf. §3.5.5). Suppose  $Z$  has an infinitesimal character; that is, suppose there is a homomorphism

$$\lambda : \mathcal{ZU}(l) \rightarrow \mathbb{C}$$

of the center of the enveloping algebras of  $l$ , such that  $u(z) = \lambda(u)z$  for  $u \in \mathcal{ZU}(l)$  and  $z \in Z$ ; or, in other words,  $\mathcal{ZU}(l)$  acts on  $Z$  by scalars. Then the  $\mathcal{R}^j(Z)$  will also have an infinitesimal character (independent of  $j$ ), and this character is determined by the Harish-Chandra homomorphism (cf. Theorem 3.5.5.23). Precisely, if  $\tilde{p} : \mathcal{ZU}(\mathbf{g}) \rightarrow \mathcal{ZU}(l)$  is the Harish-Chandra homomorphism, then the infinitesimal character of the  $\mathcal{R}^j(Z)$  is  $\lambda \circ \tilde{p}$ . This is obvious for  $\mathcal{R}^0(Z)$ , and follows for  $j > 0$  by an argument similar to that used in the current proofs of Theorem 3.5.5.20.

One can also show that  $\mathcal{R}^j$  takes finite-length modules to finite-length modules, and Harish-Chandra (i.e., admissible) modules to Harish-Chandra modules.

The most interesting question, of course, is when are the  $\mathcal{R}^j(Z)$  nonzero. Because the  $\mathcal{R}^j$  are computed, at least as  $\mathbf{k}$ -modules, in terms of relative  $(\mathbf{k}, l \cap \mathbf{k})$ -cohomology as per equation (3.6.5.7), and since the standard complex [BoWa, Knap1] for computing relative cohomology has length  $\dim(\mathbf{k}/\mathbf{k} \cap l)$ , a first conclusion is that  $\mathcal{R}^j(Z) = 0$  if  $j > \dim(\mathbf{k}/(\mathbf{k} \cap l))$ . However, a stronger result holds. Observe that

$$(\mathbf{k}/(\mathbf{k} \cap l))_{\mathbb{C}} \simeq (\mathbf{k}_{\mathbb{C}} \cap \mathbf{n}^{+}) \oplus (\mathbf{k}_{\mathbb{C}} \cap \mathbf{n}^{-}).$$

Thus

$$\dim(\mathbf{k}/(\mathbf{k} \cap l)) = 2 \dim(\mathbf{k}_{\mathbb{C}} \cap \mathbf{n}^{-}) = 2 \dim(\mathbf{k}_{\mathbb{C}}/(\mathbf{k}_{\mathbb{C}} \cap \mathbf{q})).$$

The modules we are dealing with are defined in terms of the production constructions as described in equation (3.6.5.3). For produced modules, one can construct a special resolution (see [Wall2, §6.A.14] for the analog for induction) which implies they have relative cohomology only in a restricted range. Precisely, the relative  $(\mathbf{k}_{\mathbb{C}}, (\mathbf{k} \cap l)_{\mathbb{C}})$ -cohomology of a  $(\mathbf{k}_{\mathbb{C}}, (\mathbf{k} \cap l)_{\mathbb{C}})$ -module produced from  $\mathbf{k}_{\mathbb{C}} \cap \mathbf{q}$  will vanish in degrees above

$$\dim \mathbf{k}_{\mathbb{C}}/(\mathbf{k}_{\mathbb{C}} \cap \mathbf{q}) = \frac{1}{2} \dim(\mathbf{k}/\mathbf{k} \cap l).$$

The modules with which we are working have the form  $\text{pro}_q^{\mathfrak{g}_c}(Z \otimes C_{\rho_q})$ . Their restrictions to  $\mathfrak{k}$  are not precisely of the form  $\text{pro}_{\mathfrak{k}_c \cap \mathfrak{q}}^{\mathfrak{k}} Y$ , but they have a filtration by submodules whose quotients are of this form, and this is sufficient to establish the vanishing.

**THEOREM 3.6.5.15.**  $\mathcal{R}^j(Z) = 0$  for  $j \geq \dim(\mathfrak{k}_c/(\mathfrak{k}_c \cap \mathfrak{q}))$ .

On the other hand, relative Lie algebra cohomology features a version of Poincaré duality [BoWa, KnVo, Wall2]. This allows one to show, under a certain positivity condition [Knap1, Voga1, Wall2] on the infinitesimal character of  $Z$ , that  $\mathcal{R}^j(Z) = 0$  also for  $j < \dim(\mathfrak{k}_c/(\mathfrak{k}_c \cap \mathfrak{q}))$ . Thus, under these hypotheses, we have  $\mathcal{R}^j(Z) = 0$  except for  $j = S = \frac{1}{2} \dim(\mathfrak{k}/(\mathfrak{k} \cap \mathfrak{l}))$ . The function, roughly, of the positivity condition is to make the  $((\mathfrak{k} \cap \mathfrak{l})$ -finite vectors in the) produced modules  $\text{pro}_q^{\mathfrak{g}_c}(Z \otimes C_{\rho_q})$  irreducible, or at least to guarantee that these modules carry nondegenerate Hermitian forms (the Shapovalov form [Voga1, 3]); this then guarantees that their Hermitian dual modules are again produced modules of the same sort, so that Theorem 3.6.5.15 applies to them too. Then Poincaré duality, which relates the Hermitian dual of  $\mathcal{R}^j(Z)$  to  $\Gamma^{2s-j}$  of a module constructed from the Hermitian dual of  $Z$ , guarantees vanishing of  $\mathcal{R}^j$  except for  $j = S$ .

On the other hand, it is clear that some kind of condition on  $Z$  is needed to guarantee vanishing of  $\mathcal{R}^j(Z)$  for  $j < S$ . For example, in the case when  $\mathfrak{l} = \mathfrak{k} \cap \mathfrak{l} = \mathfrak{t}$  is a Cartan subalgebra, and  $Z = C_\mu$  is one dimensional we see that  $\text{pro}(Z)$  is the dual of a Verma module (cf. equation (3.6.5.21) below), which will contain a finite-dimensional representation if  $\mu$  is negative.

**REMARK.** It may be instructive to make a few elementary observations about the structure of  $\text{pro}_q^{\mathfrak{g}_c}(Z)_{(\mathfrak{k} \cap \mathfrak{l})}$ . For this digression, we will abbreviate

$$\text{pro}_q^{\mathfrak{g}_c}(Z)_{(\mathfrak{k} \cap \mathfrak{l})} = \text{pro}(Z).$$

By the Poincaré-Birkhoff-Witt Theorem [Jaco1, Serr2, Hump], multiplication on  $\mathcal{U}(\mathfrak{g})$  induces a linear isomorphism

$$(3.6.5.16) \quad \mathcal{U}(\mathfrak{g}) \simeq \mathcal{U}(\mathfrak{n}^-) \otimes \mathcal{U}(\mathfrak{q}).$$

It follows that a mapping  $T \in \text{Hom}_q(\mathcal{U}(\mathfrak{g}), Z)$  is determined by its values on  $\mathcal{U}(\mathfrak{n}^-)$ ; and conversely, any mapping from  $\mathcal{U}(\mathfrak{n}^-)$  to  $Z$  extends to a  $\mathfrak{q}$ -module map from  $\mathcal{U}(\mathfrak{g})$  to  $Z$ . Thus, as linear spaces, we have the isomorphism

$$(3.6.5.17) \quad \text{pro}(Z) \simeq \text{Hom}(\mathcal{U}(\mathfrak{n}^-), Z)_{(\mathfrak{k} \cap \mathfrak{l})}.$$

Consider the action of  $\mathfrak{l}$  on the module. The definition of the  $\mathfrak{g}$ -action is by multiplication on the right:

$$l(T)(u) = T(ul), \quad l \in \mathfrak{l}, T \in \text{Hom}_q(\mathcal{U}(\mathfrak{g}), Z).$$

On the other hand, the condition that  $T$  be a  $\mathfrak{q}$ -module map is that

$$T(lu) = l(T(u)).$$

Thus we may write

$$\begin{aligned} l(T)(u) &= T(ul) - T(lu) + l(T(u)) \\ &= -T([l, u]) + l(T(u)). \end{aligned}$$

Since  $\mathfrak{u}^-$  is stable under  $\text{ad } \mathfrak{l}$ , this formula allows us to see the linear isomorphism (3.6.5.17) as an isomorphism of  $\mathfrak{l}$ -modules, if we define

$$l(T_1)(u) = -T_1(\text{ad } l(u)) + l(T_1(u)),$$

$$l \in \mathfrak{l}, u \in \mathcal{U}(\mathfrak{n}^-), T_1 \in \text{Hom}(\mathcal{U}(\mathfrak{n}^-), Z).$$

Note that this is the standard action of  $\mathfrak{l}$  on  $\text{Hom}(\mathcal{U}(\mathfrak{n}^-), Z)$ , constructed from the actions  $\text{ad}$  on  $\mathcal{U}(\mathfrak{n}^-)$ , and the given action on  $Z$  [Serr2, Jaco1, Hump].

Denote by  $\mathfrak{c}$  the center of  $\mathfrak{l}$ ; thus

$$(3.6.5.19) \quad \mathfrak{l} = \mathfrak{c} \oplus [\mathfrak{l}, \mathfrak{l}],$$

where  $[\mathfrak{l}, \mathfrak{l}]$  indicates the commutator ideal in  $\mathfrak{l}$ . Note that  $\mathfrak{c} \subseteq \mathfrak{k}$ . Let  $\Sigma_{\mathfrak{n}^-} \subseteq (\mathfrak{c}_{\mathfrak{C}})^*$  be the set of weights for the adjoint action of  $\mathfrak{c}$  on  $\mathfrak{n}^-$ . Then the weights of  $\mathfrak{c}$  acting on  $\mathcal{U}(\mathfrak{n}^-)$  have the form  $\sum n_{\beta} \beta$ , with  $\beta$  running through  $\Sigma_{\mathfrak{n}^-}$  and  $n_{\beta} \in \mathbb{Z}^+$ , the nonnegative integers. These all lie inside some proper cone in  $i\mathfrak{c}^*$ . In particular, the multiplicity of any given weight for  $\mathfrak{c}$  in  $\mathcal{U}(\mathfrak{n}^-)$  is finite. This multiplicity is known as the *partition function*. We denote it by  $p_{\mathfrak{n}^-}$ . Also, denote by  $C_{\mathfrak{n}^-}$  the set of all weights of  $\mathfrak{c}$  on  $\mathcal{U}(\mathfrak{n}^-)$ . Write

$$(3.6.5.20) \quad \mathcal{U}(\mathfrak{n}^-) = \sum_{\gamma \in C_{\mathfrak{n}^-}} \mathcal{U}(\mathfrak{n}^-)_{\gamma},$$

where  $\mathcal{U}(\mathfrak{n}^-)_{\gamma}$  is the  $\mathfrak{c}$ -eigenspace for the weight  $\gamma$ . Note that  $\dim \mathcal{U}(\mathfrak{n}^-)_{\gamma} = p_{\mathfrak{n}^-}(\gamma)$ .

Suppose that  $\mathfrak{c}$  acts on the  $(\mathfrak{l}, \mathfrak{k} \cap \mathfrak{l})$ -module  $Z$  by scalars; that is,  $Z$  consists of a single weight space for  $\mathfrak{c}$ . Since this is automatically true if  $Z$  has an infinitesimal character, and since any  $(\mathfrak{l}, \mathfrak{k} \cap \mathfrak{l})$ -module is a direct sum of its  $\mathfrak{c}$ -weight spaces, the assumption that  $Z$  is a single weight space is only a mild restriction on  $Z$ . Note also that if  $Z$  is a weight space for  $\mathfrak{c}$ , so is  $Z \otimes C_{\rho_q}$ . If the weight defined by  $Z$  is  $\mu_Z = \mu \in (\mathfrak{c}_{\mathfrak{C}})^*$ , then the weight defined by  $Z \otimes C_{\rho_q}$  is  $\mu + \rho_q$ . A very important special case is when  $Z = C_{\mu}$ , the one-dimensional module with weight  $\mu$ .

By means of the decomposition (3.6.5.20) above, we may regard  $\text{Hom}(\mathcal{U}(\mathfrak{n}^-)_{\gamma}, Z)$  as a subspace of  $\text{Hom}(\mathcal{U}(\mathfrak{n}^-), Z)$ . Under the hypotheses of the previous paragraph, it is a weight space for  $\mathfrak{c}$ , with weight  $\mu - \gamma$ . It follows that  $\text{pro}(Z)$  consists simply of the direct sum

$$(3.6.5.21) \quad \text{pro}(Z) \simeq \sum_{\gamma \in C_{\mathfrak{n}^-}} \text{Hom}(\mathcal{U}(\mathfrak{n}^-)_{\gamma}, Z),$$

and the  $\mathfrak{c}$ -weights of this module are the set  $\mu - C_{\mathfrak{n}^-}$ . In the case when  $\mathfrak{l} = \mathfrak{k} \cap \mathfrak{l} = \mathfrak{t}$  is a Cartan subalgebra, we see that, as a  $\mathfrak{t}$ -module,  $\text{pro}(Z)$  looks like a Verma module (cf. §3.5.3). However, it is not; the induced module (3.6.5.2) is a Verma module, and our  $\text{pro}(Z)$  is dual to a Verma module. But if the positivity condition mentioned above is satisfied, then the Verma module dual to  $\text{pro}(Z)$  is irreducible; hence,  $V$  itself is an irreducible Verma module, and one is in a position to use the vanishing Theorem 3.6.5.15.

In light of the discussion above, the question to focus on is, what is the structure of  $\mathcal{R}^S(Z)$ ? When all  $\mathcal{R}^j(Z)$  except  $\mathcal{R}^S(Z)$  vanish, the Euler-Poincaré principle of cohomology [Lang3, p. 124] allows one to calculate the  $K$ -structure of  $\mathcal{R}^S(Z)$ . The result is a formula for the multiplicities of  $K$ -types in  $\mathcal{R}^S(Z)$  in terms of an alternating sum over the Weyl group of the values of a partition function, not the partition function of all of  $\mathfrak{n}^-$ , but of the noncompact part of  $\mathfrak{n}^-$ . This is a generalization of the Blattner formula (cf. Remark (i) following Theorem 3.6.3.7). Despite its elegance, it is difficult to get specific information about general  $K$ -types from it. However, if  $Z$  is a character, then under the same positivity conditions that imply vanishing of  $\mathcal{R}^j(Z)$  except for  $j = S$ , one can find one particular  $K$ -type, the analog of the lowest  $K$ -type for the discrete series (cf. Remark (i) again); in fact it is the lowest  $K$ -type in the sense of Vogan [Voga2, 4], which occurs with multiplicity one. This implies in particular that  $\mathcal{R}^S(Z) \neq 0$ .

Thus for certain  $(\mathfrak{l}, \mathfrak{l} \cap \mathfrak{k})$ -modules  $Z$ , we obtain a nontrivial  $(\mathfrak{g}, \mathfrak{k})$ -module  $\mathcal{R}^S(Z)$ , with (in principle) known  $K$ -structure. Furthermore, again under the same positivity hypotheses, Poincaré duality considerations enable one to show that if  $Z$  has an invariant Hermitian inner product, then so does  $\mathcal{R}^S(Z)$ , and in fact [Voga3, Wall5], if  $Z$  is unitary (i.e., the invariant Hermitian structure is positive definite), then so is  $\mathcal{R}^S(Z)$ . Hence one has a construction of unitary representations.

Among these representations are the discrete series. Again for simplicity, we assume that  $G$ , hence  $K$  and  $T$ , is connected, and we use the notation of §3.6.3.

**THEOREM 3.6.5.22.** *Let  $\mathfrak{l} = \mathfrak{k} \cap \mathfrak{l} = \mathfrak{t}$  be a Cartan subalgebra of  $\mathfrak{g}$ , and let  $\mathfrak{b} \supseteq \mathfrak{t}$  be a  $(\theta$ -stable) Borel subalgebra containing  $\mathfrak{t}$ . Let  $T$  be the torus associated to  $\mathfrak{t}$ . Choose  $\mu \in \hat{T} + \rho$ ,  $\mu$  dominant for  $\mathfrak{b}$ . Then  $\mathcal{R}^S(C_\mu)$  is the discrete series representation attached to  $\mu$ .*

In Zuckerman's original formulation, this result was simply a recognition theorem, based on a characterization by Schmid [Schm2] of discrete series in terms of their  $K$ -spectrum and Casimir eigenvalues. However, subsequent work has enabled irreducibility, unitarity, and square-integrability to be established a priori, so that the  $\mathcal{R}^S(C_\mu)$  provide a construction of the discrete series independent of Harish-Chandra's character theory (see [Wall2] for a careful account). That the  $\mathcal{R}^S(C_\mu)$  yield all discrete series is still beyond

primarily algebraic methods. However, they do, and many other representations besides. For example, in [VoZu], all unitary representations with nonvanishing relative  $(\mathfrak{g}, \mathfrak{k})$ -cohomology are classified. These are the representations whose multiplicities in  $L^2(\Gamma \backslash G)$  determine the Betti numbers of locally symmetric spaces, via Matsushima's formula [BoWa, HoWa, Mats].

**4. Other directions and applications.** The “applications” of Lie theory are diverse and many; but often it is absurd to speak of “applications” when the role of Lie theory is so basic and pervasive: although  $\mathbf{R}^n$  is a Lie group, usually it is used in such a low-tech way that its Lie-theoretic properties are superfluous. But when we combine it with its character group to form the Heisenberg group acting on  $L^2(\mathbf{R}^n)$ , then its identity as a Lie group becomes more relevant. Similarly with linear algebra: it would be egregious to claim it completely as part of Lie theory, but as I hope was demonstrated in §1, the border beyond which one should definitely consider oneself in Lie-theoretic territory is easy to cross and not so far from the public entrance. There is a similar identity problem on the high-end: to what extent should one include the more exotic algebraic structures dreamed up in physics—Jordan algebras, Kac-Moody algebras, Lie “superalgebras”, quantum groups, vertex algebras—in “Lie theory”? Thus a comprehensive survey of “applications of Lie theory” is not simply impossible, it is fruitless. Below I offer instead an eclectic set of examples that I hope hit a few of the high points. For some original and stimulating discussions of many applications of representation theory, not exclusively of Lie groups, see various books and articles of Mackey [Mack1–3].

**4.1. Combinatorics.** Representation theory, even just the finite-dimensional theory of the general linear group, is rife with combinatorial quantities. We illustrate with a few examples.

**4.1.1.  $S$ -FUNCTIONS.** Much of the combinatorics of symmetric functions, developed in the nineteenth century, found natural interpretations in terms of representations of  $GL_n$  when that subject began to be understood through the work of Schur [Schu] around the turn of the century. The symmetric functions known as  $S$ -functions or Schur functions [Litt, Macd1] were introduced by Jacobi, but have been named after Schur because of his interpretation of them as the characters of the irreducible “polynomial” representations of  $GL_n$  or  $U_n$ . There is a famous identity due to Cauchy [Macd1, Weyl2] that, when interpreted in terms of representation theory, yields one of the most useful formulations of the fundamental theorems of classical invariant theory.

We will give a brief explanation of Cauchy's identity and its representation-theoretic interpretation. For an extensive discussion of symmetric functions with applications to representation theory, I recommend [Macd1]. In this discussion, we will follow the notation of [Macd1], although it differs from our earlier notation.

primarily algebraic methods. However, they do, and many other representations besides. For example, in [VoZu], all unitary representations with nonvanishing relative  $(\mathfrak{g}, \mathfrak{k})$ -cohomology are classified. These are the representations whose multiplicities in  $L^2(\Gamma \backslash G)$  determine the Betti numbers of locally symmetric spaces, via Matsushima's formula [BoWa, HoWa, Mats].

**4. Other directions and applications.** The “applications” of Lie theory are diverse and many; but often it is absurd to speak of “applications” when the role of Lie theory is so basic and pervasive: although  $\mathbf{R}^n$  is a Lie group, usually it is used in such a low-tech way that its Lie-theoretic properties are superfluous. But when we combine it with its character group to form the Heisenberg group acting on  $L^2(\mathbf{R}^n)$ , then its identity as a Lie group becomes more relevant. Similarly with linear algebra: it would be egregious to claim it completely as part of Lie theory, but as I hope was demonstrated in §1, the border beyond which one should definitely consider oneself in Lie-theoretic territory is easy to cross and not so far from the public entrance. There is a similar identity problem on the high-end: to what extent should one include the more exotic algebraic structures dreamed up in physics—Jordan algebras, Kac-Moody algebras, Lie “superalgebras”, quantum groups, vertex algebras—in “Lie theory”? Thus a comprehensive survey of “applications of Lie theory” is not simply impossible, it is fruitless. Below I offer instead an eclectic set of examples that I hope hit a few of the high points. For some original and stimulating discussions of many applications of representation theory, not exclusively of Lie groups, see various books and articles of Mackey [Mack1–3].

**4.1. Combinatorics.** Representation theory, even just the finite-dimensional theory of the general linear group, is rife with combinatorial quantities. We illustrate with a few examples.

**4.1.1.  $S$ -FUNCTIONS.** Much of the combinatorics of symmetric functions, developed in the nineteenth century, found natural interpretations in terms of representations of  $GL_n$  when that subject began to be understood through the work of Schur [Schu] around the turn of the century. The symmetric functions known as  $S$ -functions or Schur functions [Litt, Macd1] were introduced by Jacobi, but have been named after Schur because of his interpretation of them as the characters of the irreducible “polynomial” representations of  $GL_n$  or  $U_n$ . There is a famous identity due to Cauchy [Macd1, Weyl2] that, when interpreted in terms of representation theory, yields one of the most useful formulations of the fundamental theorems of classical invariant theory.

We will give a brief explanation of Cauchy's identity and its representation-theoretic interpretation. For an extensive discussion of symmetric functions with applications to representation theory, I recommend [Macd1]. In this discussion, we will follow the notation of [Macd1], although it differs from our earlier notation.



Let  $x_1, x_2, \dots, x_n$  be indeterminates. We will, when convenient, think of the  $x_i$  as coordinates of a point in  $\mathbb{C}^n$ . Given an  $n$ -tuple

$$(4.1.1.1) \quad \mu = (\mu_1, \mu_2, \dots, \mu_n)$$

of nonnegative integers, we write

$$(4.1.1.2) \quad x^\mu = \prod_j x_j^{\mu_j}.$$

We let the symmetric group  $S_n$  permute the variables  $x_i$  in the obvious way. This gives rise to a permutation action of  $S_n$  on the monomials  $x^\mu$ :

$$(4.1.1.3) \quad s(x^\mu) = \prod_j x_{s(j)}^{\mu_j}.$$

If an  $n$ -tuple  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$  of nonnegative integers is arranged in decreasing order, i.e.,  $\lambda_j \geq \lambda_{j+1}$ , we call  $\lambda$  a *partition*. We write

$$(4.1.1.4) \quad \delta = (n-1, n-2, \dots, 2, 1, 0)$$

for the smallest partition whose entries are strictly decreasing.

For a partition  $\alpha$ , define

$$(4.1.1.5) \quad a_\alpha = \sum_{w \in S_n} \varepsilon(w) w(x^\alpha).$$

Here  $\varepsilon$  is the sign character of  $S_n$ . The polynomial  $a_\alpha$  is easily checked to be skew-symmetric, i.e.,

$$w(a_\alpha) = \varepsilon(w) a_\alpha.$$

Hence  $a_\alpha = 0$  unless the entries  $\alpha_i$  of  $\alpha$  are strictly decreasing. In this case, we can write  $\alpha = \lambda + \delta$ , where  $\lambda$  is again a partition. The argument with the Vandermonde determinant (cf. §3.5.4) shows that  $a_{\lambda+\delta}$  is divisible by  $a_\delta$ . Set

$$(4.1.1.6) \quad s_\lambda = a_{\lambda+\delta} / a_\delta.$$

Comparison with formula (3.5.4.24), allowing for the differences in notation, shows that if the  $x_i$  are thought of as the coordinates of a unitary diagonal matrix, then  $s_\lambda$  is the character of the representation of  $U_n$  with highest weight  $\lambda$ . In this context we call the  $s_\lambda$  *Schur-functions* or *S-functions*.

Cauchy's identity says

$$(4.1.1.7) \quad a_\delta(x) a_\delta(y) \left( \prod_{i,j=1}^n (1 - x_i y_j) \right)^{-1} = \sum_\lambda a_{\lambda+\delta}(x) a_{\lambda+\delta}(y).$$

The sum is over all partitions  $\lambda$ . This identity arises by evaluating

$$\det((1 - x_i y_j)^{-1})$$

in two different ways (cf. [Macd1, pp. 38, 33]). Dividing by  $a_\delta(x)a_\delta(y)$  gives us

$$(4.1.1.8) \quad \prod_{i,j=1}^n (1 - x_i y_j)^{-1} = \sum_{\lambda} s_{\lambda}(x) s_{\lambda}(y).$$

The right-hand side of this has a clear representation-theoretic interpretation, at least formally: it is the character of the representation  $\sum_{\lambda} \rho_n^{\lambda} \otimes \rho_n^{\lambda}$  of  $U_n \times U_n$ , where  $\rho_n^{\lambda}$  denotes the representation of  $U_n$  with highest weight  $\chi_{\lambda}$  (notation as in equation (3.5.4.2)). This observation challenges us to find a representation-theoretic interpretation of the left-hand side of (4.1.1.8).

The challenge is met by the *Molien formula* [Bour, Moli, Stan]. This is a formula for the trace of an element of  $g \in GL_n$  acting on the space  $\mathcal{P}(\mathbb{C}^n)$  of polynomials on  $\mathbb{C}^n$ . Since  $\mathcal{P}(\mathbb{C}^n)$  is infinite dimensional, the notion of trace will require some interpretation. Consider the space  $\mathcal{P}^d(\mathbb{C}^n)$  of polynomials homogeneous of degree  $d$ . Let us compute the trace of  $g$  acting on  $\mathcal{P}^d(\mathbb{C}^n)$ . Up to conjugation of  $g$ , and ignoring the lower-dimensional subvariety consisting of nondiagonalizable  $g$ , we may assume  $g$  is diagonal, with eigenvalues  $c_j$ :

$$g(x_j) = c_j x_j.$$

Then

$$g(x^{\mu}) = \prod_j g(x_j)^{\mu_j} = \left( \prod_j c_j^{\mu_j} \right) x^{\mu} = c^{\mu} x^{\mu},$$

where  $c = (c_1, c_2, \dots, c_n)$  is the  $n$ -tuple of eigenvalues of  $g$ . Thus the eigenvalues of  $g$  acting on  $\mathcal{P}^d(\mathbb{C}^n)$  are  $\{c^{\mu}\}$ , where  $|\mu| = \sum \mu_j = d$ . Thus

$$\text{trace } g|_{\mathcal{P}^d(\mathbb{C}^n)} = \sum_{|\mu|=d} c^{\mu}.$$

If now we formally sum over all  $d$  we get

$$\begin{aligned} (4.1.1.9) \quad \text{trace } g|_{\mathcal{P}(\mathbb{C}^n)} &= \sum_{\mu} c^{\mu} = \sum_{\mu_j \geq 0} c_1^{\mu_1} c_2^{\mu_2} \cdots c_n^{\mu_n} \\ &= \left( \sum_{\mu_1 \geq 0} c_1^{\mu_1} \right) \left( \sum_{\mu_2 \geq 0} c_2^{\mu_2} \right) \cdots \left( \sum_{\mu_n \geq 0} c_n^{\mu_n} \right) \\ &= \prod_{j=1}^n \frac{1}{(1 - c_j)} = \left( \prod_{j=1}^n (1 - c_j) \right)^{-1} = (\det(1 - g))^{-1}. \end{aligned}$$

Here  $1$  indicates the  $n \times n$  identity matrix. Observe that if all the eigenvalues  $c_j$  of  $g$  have absolute value less than  $1$ , then the infinite series in (4.1.1.9) converges as  $d \rightarrow \infty$ . Thus equation (4.1.1.9) can be regarded as an equality of analytic functions on the appropriate open subset of  $GL_n(\mathbb{C})$ ; or since

the right-hand side of the equation is rational, it can be thought of as an interpretation of the left-hand side as a rational function.

If  $G \subseteq \text{GL}_n(\mathbf{C})$  is a subgroup, then  $G$  may not contain any elements whose eigenvalues are all of absolute value less than 1 (i.e.,  $G = \text{SL}_n(\mathbf{C})$ , etc). However, if we augment  $G$  by the group  $\mathbf{C}^\times$  of scalar operators, this larger group will clearly have such elements. By this device, formula (4.1.1.19) makes sense for elements of any subgroup of  $\text{GL}_n(\mathbf{C})$ . If we wish to emphasize the role of the scalar, we can write

$$(4.1.1.10) \quad \text{trace } tg|_{\mathcal{P}(\mathbf{C}^n)} = (\det(1 - tg))^{-1}.$$

This formula is valid for all  $g \in \text{GL}_n(\mathbf{C})$ , and all sufficiently small  $t$  (depending on  $g$ ) in  $\mathbf{C}$ . It can also be regarded as an identity in formal power series in  $t$ . Thus if  $\pi : G \rightarrow \text{GL}_n(\mathbf{C})$  is a representation of  $\mathbf{C}^n$ , we can write

$$(4.1.1.11a) \quad \text{trace}(t\pi(g)|_{\mathcal{P}(\mathbf{C}^n)}) = (\det(1 - t\pi(g)))^{-1}.$$

We can regard equation (4.1.1.11a) as the correct version of the equation

$$(4.1.1.11b) \quad \text{trace}(\pi(g)|_{\mathcal{P}(\mathbf{C}^n)}) = (\det(1 - \pi(g)))^{-1}$$

which will make sense as long as the set of  $g$  in  $G$  such that  $\pi(g)$  has 1 as an eigenvalue is nowhere dense in  $G$ .

Given this background, we see our challenge is to find a representation of  $U_n \times U_n$  such that the left-hand side of equation (4.1.1.8) is the right-hand side of equation (4.1.1.11b) (for  $g$  a product of diagonal matrices). We do not have far to seek. Consider the action  $\pi$  of  $U_n \times U_n$  on  $M_n(\mathbf{C})$ , the  $n \times n$  matrices, by left and right multiplication:

$$(4.1.1.12) \quad \pi(g_1, g_2)(T) = g_1 T g_2^t, \quad g_i \in U_n, T \in M_n.$$

Here  $g_2^t$  is the transpose of  $g_2$ ; we use  $g_2^t$  instead of the more standard  $g_2^{-1}$  in order to make things come out symmetrically in  $g_1$  and  $g_2$ . It is trivial to check that if  $g_1, g_2$  are diagonal matrices with eigenvalues  $\{x_j\}_{j=1}^n$  and  $\{y_k\}_{k=1}^n$  respectively, then the right-hand side of (4.1.1.11b) is exactly the left-hand side of (4.1.1.8). Thus (4.1.1.8) is revealed as an expansion of  $\text{trace}(\pi(g_1, g_2)|_{\mathcal{P}(M_n)})$  into a sum of characters of  $U_n \times U_n$ . Because representations are determined by their characters, we deduce the following corollary of the combination of the Cauchy identity and Molien's formula.

**THEOREM 4.1.1.13** (Fundamental theorem of invariant theory, polynomial duality version). *Under the action  $\pi$  (cf. (4.1.1.12)) of  $U_n \times U_n$  on  $M_n(\mathbf{C})$ , the decomposition of  $\mathcal{P}(M_n(\mathbf{C}))$  into irreducible representations is described by*

$$\mathcal{P}(M_n(\mathbf{C})) \simeq \sum_{\lambda} \rho_n^{\lambda} \otimes \rho_n^{\lambda}.$$

Here  $\lambda$  runs over all partitions (cf. (4.1.1.1)–(4.1.1.4)). The subspace  $\mathcal{P}^d(M_n(\mathbf{C}))$  is the sum over all  $\lambda$  with  $|\lambda| = \sum_{j=1}^n \lambda_j = d$ .

REMARKS. (a) A significant feature of this result is that, in the decomposition of  $\mathcal{P}(M_n(\mathbb{C}))$ , a given representation  $\rho_n^\lambda$  of the first factor of  $U_n \times U_n$  is combined with exactly one representation (which again happens to be  $\rho_n^\lambda$ ) of the second factor. In other words, the symmetry type under left multiplication of a polynomial on  $M_n(\mathbb{C})$  determines its symmetry type under right multiplication. This fact has many repercussions in invariant theory. In particular, it directly implies various reciprocity laws (cf. [Howe8]).

(b) In the discussion above, we have used a combinatorial formula to make a representation-theoretic conclusion. However, the flow could be reversed: we can prove Theorem 4.1.1.13 directly from general principles of representation theory. Computing the character of the action via Molien's formula (4.1.1.12) then would yield Cauchy's identity in the form (4.1.1.8). The main observation of the representation-theoretic approach is that Theorem 4.1.1.13 is essentially an example of the Peter-Weyl Theorem (cf. §3.5.4). To see this, we make the following observations.

(i) Polynomial functions on  $M_n(\mathbb{C})$  are determined by their restriction to  $U_n$ . This is because polynomials are in particular holomorphic functions on  $M_n(\mathbb{C})$ , and  $U_n$  is a "real form" of  $GL_n(\mathbb{C})$ , in the sense that its Lie algebra  $\mathfrak{u}_n$  is a real subspace of  $\mathfrak{gl}_n(\mathbb{C}) \simeq M_n(\mathbb{C})$  such that  $\mathfrak{gl}_n(\mathbb{C}) = \mathfrak{u}_n \oplus i\mathfrak{u}_n$  (cf. §3.5.5). It follows that if  $p$  is a polynomial on  $M_n(\mathbb{C})$ , then the Taylor series of  $p$  at 1, the identity matrix, is determined by the Taylor series of  $p|_{U_n}$ , the restriction of  $p$  to  $U_n$ . Since  $p$  is in turn determined by its Taylor series at 1, the observation follows.

(ii) The restriction of  $\mathcal{P}(M_n(\mathbb{C}))$  to  $U_n$  will yield a space of functions which is invariant under left and right multiplications on  $U_n$ . Hence the Peter-Weyl Theorem (cf. Theorem 3.5.4.23) implies that  $\mathcal{P}(M_n(\mathbb{C}))$  is a direct sum of modules of the form  $\rho_n^\lambda \otimes (\rho_n^\lambda)^*$ , where  $(\rho_n^\lambda)^*$  is the contra-gradient of  $\rho_n^\lambda$ . The automorphism  $g \rightarrow (g^t)^{-1}$ , used to create the action  $\pi$  of equation (4.1.1.12), transforms  $(\rho_n^\lambda)^*$  to  $\rho_n^\lambda$ .

(iii) From (i) and (ii) we conclude that  $\mathcal{P}(M_n(\mathbb{C}))$  is a sum of  $(U_n \times U_n)$ -modules of the form  $\rho_n^\lambda \otimes \rho_n^\lambda$ . It remains to determine which  $\lambda$  can occur. Since all the weights of the diagonal torus in  $U_n \times U_n$  are polynomial weights, in the sense that they involve only nonnegative powers of the diagonal entries, it follows that the highest weights  $\lambda$  must be restricted to be partitions, i.e., must also have all entries nonnegative. Conversely, it is possible to explicitly exhibit  $U_n \times U_n$  highest weight vectors of weight  $\lambda$ , for any partition  $\lambda$ . Let  $\{z_{ij} : 1 \leq i, j \leq n\}$  denote the standard matrix-entry coordinates on  $M_n(\mathbb{C})$ . Set

$$\gamma_k = \det \begin{vmatrix} z_{11} & z_{12} & \cdots & z_{1k} \\ z_{21} & & & \vdots \\ \vdots & \ddots & & \\ z_{k1} & & \cdots & z_{kk} \end{vmatrix}$$

for  $1 \leq k \leq n$ , and, for a partition  $\lambda$ ,

$$(4.1.1.14) \quad \gamma^\lambda = \prod_{k \geq 1} (\gamma_k)^{\lambda_k - \lambda_{k-1}}$$

(where we agree that  $\lambda_{n+1} = 0$ ). Then  $\gamma^\lambda$  is a  $U_n \times U_n$  highest weight vector of weight  $\lambda$ , showing that  $\rho_n^\lambda \otimes \rho_n^\lambda$  does occur as a  $(U_n \times U_n)$ -submodule of  $\mathcal{P}(M_n(\mathbb{C}))$ .

(c) The above argument is perhaps an expensive way to establish Cauchy's identity, but it does illustrate in a simple case the potential combinatorial import of the seemingly bland general structure theorems of representation theory. The combinatorial content, of course, is supplied by the structure of Lie groups, in this case  $U_n$ . Also, it is important to understand that Theorem 4.1.1.13 is a robust result, not dependent on clever manipulations of power series.

(d) For further details on how Theorem 4.1.1.13 leads to invariant theory, we refer to [Howe1, 8, 9]. For more examples of identities that arise in similar fashion, see [Proc2, Litt, Tera].

**4.1.2. MULTIPLICITIES.** A basic problem in the representation theory of  $U_n$  (or any other compact Lie group), is understanding how an irreducible representation of  $U_n$  decomposes under restriction to the Cartan subgroup (diagonal torus)  $A$ . This problem is of interest in itself; additionally, it has many ramifications, some quite surprising. Our goal in this section is to explain some of these ramifications.

Let  $\rho$  be a representation of  $U_n$  on a vector space  $V$ . We can decompose  $V$  into eigenspaces for the diagonal torus  $A$ :

$$(4.1.2.1) \quad V = \sum_{\chi \in \widehat{A}} V_\chi,$$

where

$$\rho(a)(v) = \chi(a)v, \quad a \in A, v \in V_\chi.$$

The space  $V_\chi$  is called the  $\chi$ -weight space. The collection of  $\{V_\chi\}$  are called the  $A$  weight spaces, or just the weight spaces. The dimension  $\dim V_\chi$  is called the *multiplicity* of  $\chi$  in  $\rho$ . We are particularly interested in the multiplicity of  $\chi$  when  $\rho = \rho_n^\lambda$  is an irreducible representation of  $U_n$ . We denote the multiplicity of  $\chi$  in  $\rho_n^\lambda$  by

$$(4.1.2.2) \quad m(\lambda, \chi).$$

The Weyl group  $S_n$  normalizes  $A$  inside  $U_n$ , and so it will act on the decomposition (4.1.2.1), permuting the  $V_\chi$ . Thus  $\dim V_\chi = \dim V_{w(\chi)}$  for  $\chi \in \widehat{A}$ . Hence in computing the  $m(\lambda, \chi)$ , one can restrict to  $\chi \in \widehat{A}^+$  (cf. definition (3.5.4.15)). In other words, if we take  $\chi = \chi_\alpha$ , where  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$  is in  $\mathbb{Z}^n$  (cf. equation (3.5.4.2)), we can assume that the

entries  $\alpha_j$  of  $\alpha$  are decreasing. If all the  $\alpha_j$  are nonnegative, then  $\alpha$  will be a partition. If  $\chi = \chi_\alpha$ , for  $\alpha$  a partition, we will write

$$(4.1.2.3) \quad m(\lambda, \chi_\alpha) = K_{\lambda\alpha}.$$

In some sense, the problem of computing the  $m(\lambda, \chi)$  is solved. There is a general formula due to Kostant [Hump, Jaco1, Kost5] for the multiplicity of a weight in an irreducible representation that applies to any semisimple Lie group. It amounts to a suitable rewriting of the Weyl character formula. There is also a recursive method for computing multiplicities given by Freudenthal [Jaco1], which has been made much more efficient by Moody and Patera [MoPa, BrMP]. For  $U_n$ , there is a combinatorial description of the multiplicities that goes back to Kostka [Kstk], and has its representation-theoretic interpretation in the Gelfand-Cetlin basis [BiLo2, Zhel1, Cher1]. There is an analog of this latter for more general groups in the standard monomial theory of Seshadri et al. [Sesh, LaSe1, 2, LaMS]. Thus, in some sense, the determination of the  $m(\lambda, \chi)$  is a very well-solved problem. On the other hand, there are problems for which these standard solutions are of little help, so there is still considerable room for a deeper understanding of the  $m(\lambda, \chi)$ .

Our purpose here is not to compute the  $m(\lambda, \chi)$ , but to show how these numbers, which arise in this context, reverberate through representation theory, echoing from topic to topic until they reach contexts seemingly quite removed from each other.

The first reflected image of the  $m(\lambda, \chi)$  is in the theory of symmetric functions. This is formed by the  $S$ -functions as the characters of the irreducible representations of  $U_n$ . The  $S$ -functions are defined (cf. equation (4.1.1.6)) in a relatively sophisticated and indirect way via a quotient. The most simple minded symmetric function to associate to a partition is the symmetrization of a monomial:

$$(4.1.2.4) \quad m_\alpha = \#(W_\alpha)^{-1} \sum_{w \in S_\alpha} w(x^\alpha).$$

Here  $\alpha$  is a partition and  $W_\alpha \subseteq S_n$  is the stabilizer of  $\alpha$ . It is obvious that the  $m_\alpha$  form a basis for the space of symmetric functions. In order to understand the  $S$ -functions  $s_\lambda$ , we might try to express  $s_\lambda$  as a linear combination of the  $m_\alpha$ . It is not hard to see that the desired expression is

$$(4.1.2.5) \quad s_\lambda = \sum K_{\lambda\alpha} m_\alpha.$$

Indeed, it is clear that the contribution of  $\sum_{w \in S_n} V_{\chi_{w(\alpha)}}$  to the trace of  $a \in A$  with entries  $(z_1, \dots, z_n)$  is just  $(\dim V_{\chi_\alpha}) m_\alpha(z)$ ; equation (4.1.2.5) is immediate from this and definition (4.1.2.3). It is in this role, as coefficients for expressing the  $s_\lambda$  in terms of the  $m_\alpha$ , that the  $K_{\lambda\alpha}$  first appeared. As such, they were studied by Kostka [Kstk], hence are known as *Kostka coefficients*.

Perhaps the most simple-minded occurrence of the Kostka coefficients outside their initial role as weight multiplicities is as  $K$ -multiplicities for principal series representations of  $\mathrm{GL}_n(\mathbb{C})$ . We refer to §3.6.1 for the context of these remarks. As explained in §3.6.5, the study of representations of a semisimple group  $G$  by means of their restrictions to the maximal compact subgroup  $K$  of  $G$  is a basic technique. Of course a representation of  $K$  is determined by the multiplicities with which the various irreducible representations of  $K$  occur. Since the principal series are such an important class of representations of  $G$ , we are curious to know the multiplicities of  $K$ -types (i.e., irreducible representations of  $K$ ) in principal series representations of  $G$ . Because of the Iwasawa decomposition  $G = KP$  for any parabolic subgroup  $P$  (cf. §A.2.3.5), the multiplicity problem for principal series reduces to a problem in the representation theory of  $K$ . Indeed, we can see directly from the Iwasawa decomposition and the definition (3.6.1.6), that

$$(4.1.2.6) \quad \mathrm{P.S.}(\sigma, \psi)|_K \simeq \mathrm{ind}_{K \cap M}^K(\sigma|_{K \cap M}).$$

For the minimal parabolic  $P_0$ , we have  $K \cap M_0 = M_0$ ; also  $\sigma$  is an irreducible representation of  $M_0$ , so formula (4.2.1.6) simplifies to

$$(4.1.2.7) \quad \mathrm{P.S.}(\sigma, \psi)|_K \simeq \mathrm{ind}_{M_0}^K \sigma.$$

Thus the multiplicities of  $K$ -types for these principal series are governed by Frobenius reciprocity (cf. [Gaal, HeRo, Knap2], etc.):

$$(4.1.2.8) \quad m(\tau, \mathrm{P.S.}(\sigma, \psi)|_K) = m(\sigma, \tau|_M), \quad \tau \in \widehat{K}, \sigma \in \widehat{M_0}.$$

Here the left-hand side of the equation indicates the multiplicity of the irreducible representation  $\tau$  of  $K$  in the principal series  $\mathrm{P.S.}(\sigma, \psi)$ ,  $\sigma \in \widehat{M_0}$ , and the right-hand side denotes the multiplicity of  $\sigma$  in the restriction of  $\tau$  to  $M_0$ .

Consider the example of  $G = \mathrm{GL}_n(\mathbb{C})$ . Then  $K = U_n$ , and  $M_0 = A$ , the diagonal torus. Hence  $\sigma$  is just a character of  $A$ , and the multiplicity of  $\tau$  in  $\mathrm{P.S.}(\sigma, \tau)$  is just the multiplicity of  $\sigma$  in  $\tau$ , i.e., is a Kostka coefficient.

The next appearance of the  $K_{\lambda_\alpha}$  is as decomposition numbers for tensor products. This is an example of the reciprocity law associated to Theorem 4.1.1.13. Let  $\mathcal{P}^a(\mathbb{C}^n)$  be the space of polynomials of degree  $a$  on  $\mathbb{C}^n$ . It defines an irreducible representation of  $U_n$ , of a relatively simple and comprehensible sort. In particular, all the weight spaces have dimension 1. One might hope to understand more complicated representations of  $U_n$  in terms of the  $\mathcal{P}^a(\mathbb{C}^n)$ . To that end, consider a tensor product

$$\mathcal{P}^{\alpha_1}(\mathbb{C}^n) \otimes \mathcal{P}^{\alpha_2}(\mathbb{C}^n) \otimes \cdots \otimes \mathcal{P}^{\alpha_n}(\mathbb{C}^n).$$

This defines a representation of  $U_n$ , probably not irreducible. The reason for considering a tensor product involving  $n$  factors is that this is what is needed to obtain an arbitrary polynomial representation of  $U_n$  as a constituent. If we want to consider fewer than  $n$  factors, we can just let some of the  $\alpha_j$

equal zero. Without loss of generality, we can arrange for the  $\alpha_j$  to decrease in  $j$ , so that  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$  is a partition. Let us then consider the decomposition

$$(4.1.2.9) \quad \mathcal{P}^{\alpha_1}(\mathbf{C}^n) \otimes \mathcal{P}^{\alpha_2}(\mathbf{C}^n) \otimes \dots \otimes \mathcal{P}^{\alpha_n}(\mathbf{C}^n) \simeq \sum n(\alpha, \lambda) \rho_n^\lambda$$

of the tensor product of the  $\mathcal{P}^{\alpha_j}(\mathbf{C}^n)$ . It turns out that

$$(4.1.2.10) \quad n(\alpha, \lambda) = K_{\lambda\alpha}.$$

This can be deduced as an example of the reciprocity laws which follow from Theorem 4.1.1.13. One can identify the tensor product on the left side of (4.1.2.9) as a subspace of  $\mathcal{P}(M_n(\mathbf{C}))$  by regarding the variables of the  $k$ th factor  $\mathcal{P}^{\alpha_k}(\mathbf{C}^n)$  to be the entries  $\{z_{jk} : 1 \leq j \leq n\}$  of the  $k$ th column of the matrix  $M = \{z_{jk} : 1 \leq j, k \leq n\}$ . Since the homogeneous pieces  $\mathcal{P}^a(\mathbf{C}^n) \subseteq \mathcal{P}(\mathbf{C}^n)$  are precisely the eigenspaces for the center of  $U_n$ , it is not hard to convince oneself that the tensor product in (4.1.2.9) is precisely the  $\chi_\alpha$ -eigenspace for the diagonal torus of  $U_n$  acting on  $M_n(\mathbf{C})$  by multiplication on the right, i.e., the torus in the second factor in the action  $\pi$  of formula (4.1.1.12). Combining this observation with Theorem 4.1.1.13, we see that the  $\rho_n^\lambda$ -isotypic component of the tensor product in (4.1.2.9) is equal to the  $\chi_\alpha$ -eigenspace of  $A$  in the second factor of  $U_n \times U_n$ , in the representation  $\rho_n^\lambda \otimes \rho_n^\lambda$ . It follows that the multiplicity of  $\rho_n^\lambda$  in the tensor product is the multiplicity of  $\chi_\alpha$  in  $\rho_n^\lambda$ , whence formula (4.1.2.10). In fact, the symmetric function equivalent of (4.1.2.10) was already known to Kostka.

The  $K_{\lambda\alpha}$  also have an interpretation in terms of the representation theory of the symmetric group. This results from Schur duality [Howe8; Weyl2, Chapter 4] in the same way that (4.1.2.10) followed from the  $(GL_n, GL_n)$ -Duality Theorem 4.1.1.13.

We recall just enough representation theory of  $S_n$  to state the result. The irreducible representations of  $S_n$  can be parametrized in a reasonable way by partitions of size  $n$  (cf. [Weyl2, Jaco2, Litt], etc.). If  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$  is a partition of size  $n$ , i.e., with  $\sum \lambda_i = n$ , let  $\sigma^\lambda$  be the associated irreducible representation of  $S_n$ .

To a partition  $\alpha$  of size  $n$  we may associate a (conjugacy class of) subgroup of  $S_n$ . Namely, we partition the set  $\{1, 2, 3, \dots, n\}$  into subsets of size  $\alpha_j$ , and consider the subgroup which preserves each of the chosen subsets. We denote this subgroup  $S_\alpha$ . Form the induced representation  $\text{ind}_{S_\alpha}^{S_n} 1$ . The decomposition of this representation into irreducible constituents is described by the Kostka coefficients:

$$(4.1.2.11) \quad \text{ind}_{S_\alpha}^{S_n} \simeq \sum K_{\lambda\alpha} \sigma^\lambda$$

REMARKS. There is a natural partial order on partitions obtained from thinking of them as elements of the positive Weyl chamber for  $U_n$ . We say



$\lambda \geq \mu$  provided  $\sum_{j=1}^l \lambda_j \geq \sum_{j=1}^l \mu_j$  for all  $l$ ,  $1 \leq l \leq n$ . The representation  $\sigma^\lambda$  is characterized by the fact that  $\sigma^\lambda$  occurs in  $\text{ind}_{S_\lambda}^{S_n} 1$  (with multiplicity 1), but is not contained in  $\text{ind}_{S_\mu}^{S_n} 1$  unless  $\lambda \geq \mu$ . This follows from equation (4.1.2.11), but is usually proved in establishing the existence of the  $\sigma^\lambda$ .

Finally, we indicate a further appearance of the  $K_{\lambda\alpha}$  as decomposition numbers, this time for  $\text{GL}_n$  over finite fields. Let  $F_q$  be the finite field with  $q$  elements. Let  $\text{GL}_n(F_q) = G$  be the group of invertible matrices with coefficients in  $F_q$ . Let  $B$  be the Borel subgroup of upper triangular matrices in  $G$ . We want to consider the induced representation

$$\tau_B = \text{ind}_B^G 1,$$

i.e., the natural action of  $G$  on  $L^2(G/B)$ .

Let  $P \supseteq B$  be a parabolic subgroup of  $G$  containing  $B$ . Then  $P$  is a group of block upper triangular matrices for some choice of block sizes  $\beta_1, \beta_2, \dots, \beta_k$  with  $\sum \beta_j = n$ . We can consider the representations

$$\tau_P = \text{ind}_P^G 1$$

which will be subrepresentations of  $\tau_B$ . For a given parabolic  $P$ , with block sizes  $\beta_1, \beta_2, \dots, \beta_k$  consider another parabolic  $P'$ , whose block sizes are  $\beta'_1, \beta'_2, \dots, \beta'_k$ , where  $\beta'_j = \beta_{s(j)}$  are a permutation of the block sizes of  $P$ . We call two parabolics related in this way *associate parabolics*. Although distinct associate parabolics are not conjugate in  $G$ , nevertheless we have  $\tau_P \simeq \tau_{P'}$ . Hence, the representation  $\tau_P$  depends only on the partition of  $n$  defined by the block sizes  $\beta_j$ . Thus, without loss of generality, we may assume the  $\beta_j$  are arranged in order of size:  $\beta_j \geq \beta_{j+1}$ . Supplementing the  $\beta_j$  with  $n - k$  zeros, we can attach to  $P$  a partition  $\beta = (\beta_1, \beta_2, \dots, \beta_k, 0, 0, \dots, 0)$ , and we have a well-defined representation  $\tau_\beta = \tau_P$  associated to  $\beta$ .

From the Bruhat decomposition for  $\text{GL}_n$  (cf. §1.2), we can conclude that the intertwining number (i.e.,  $\dim \text{Hom}_{\text{GL}_n}$ ; cf. §A.1.4) of  $\tau_\beta$  and  $\tau_\alpha$  is equal to the intertwining number of the representations  $\text{ind}_{S_\beta}^{S_n} 1$  and  $\text{ind}_{S_\alpha}^{S_n} 1$  of  $S_n$  (cf. [HoMy1, Lus1]). From this and the fact that the representations  $\text{ind}_{S_\beta}^{S_n} 1$  form a basis for the representation ring of  $S_n$  (cf. the Remark following equations (4.1.2.11)), we can conclude that the constituents of  $\tau_B$  can be labeled by partitions in such a way that

$$(4.1.2.12) \quad \tau_\alpha \simeq \sum K_{\lambda\alpha} \tau_\lambda^0,$$

where  $\tau_\lambda^0$  is the irreducible representation labeled by the partition  $\lambda$ . We see that the  $\tau_\lambda^0$  can be characterized in a fashion similar to the irreducible representations  $\sigma^\lambda$  of  $S_n$ :  $\tau_\lambda^0$  does not occur in  $\tau_\alpha$  unless  $\alpha \leq \lambda$ , and it occurs in  $\tau_\lambda$  with multiplicity 1.

For a fuller account of the above facts, we refer to [HoLe, Iwah1, HoMy1, Lus1]. The argument just sketched applies only to  $GL_n$ : the Bruhat Decomposition is of course available for general reductive groups; however, the representation ring is spanned by the representations  $\text{ind}_{W_p}^W 1$ , for Weyl groups  $W_p$  of parabolic subgroups, only for Weyl groups of type  $A_n$ .

4.1.3: CURRENT EVENTS. (i) *Symmetric functions*. The characters of  $U_n$  are not the only symmetric functions which arise naturally in representation theory of Lie groups. Other examples are functions on  $GL_n(\mathbf{R})$  which are invariant under both right and left translation by  $O_n$ . Such functions which are polynomials and eigenfunctions for the center of the universal enveloping algebra for  $GL_n(\mathbf{R})$  are called “zonal spherical polynomials” for  $GL_n(\mathbf{R})$ . They are matrix coefficients for  $O_n$ -invariant vectors in the finite dimensional irreducible representations of  $GL_n(\mathbf{R})$  (which are in natural correspondence with the representations of  $U_n$ ). They are used particularly in probability [Jame, Meht] and also elsewhere [Dyso1, GrRi1, 2]. There is an analogous family of polynomials, also parametrized by partitions, associated to  $GL_n(\mathbf{H})$ , the group of nonsingular  $n \times n$  matrices with entries in  $\mathbf{H}$ , the quaternions.

In addition to these, other symmetric functions arise naturally from  $p$ -adic groups [Macd1]. Recently I. G. Macdonald has defined a two-parameter family of bases of the symmetric functions, which include all the above functions for appropriate values of the parameters [Macd4]. These are currently under active study (cf. [Cher2, Dunk, Heck1, 2, HeOp, Opda1–3], etc.).

These developments grow out of earlier work of Macdonald [Macd3], which established a class of identities involving the Dedekind  $\eta$ -function:

$$(4.1.3.1) \quad \eta = q^{1/24} \prod_{l=1}^{\infty} (1 - q^l).$$

Except for the factor  $q^{1/24}$ , the function  $\eta$  is the reciprocal of the generating function for the partition function [Andr1]; thus it is an object of deep combinatorial interest. Macdonald established one identity for each “affine root system” [Bour, Morr, Hump, Hill]. The identity for the system  $\widetilde{A}_1$  is the Jacobi triple product identity (cf. [Andr2], etc.); for the system  $\widetilde{BC}_1$  it is the “quintuple product identity,” which is usually attributed to Watson, but which can be found in Fricke-Klein. (The identities for the classical root systems were found by Dyson [Dyso1, 2], who, however, did not see the connection with root systems.) Macdonald saw his identities as analogs for affine root systems of the Weyl denominator identity (cf. [Jaco1, §VIII.3; Hump, §24.3], etc.) for finite root systems. (For root systems of type  $A_n$ , this is the formula for the Vandermonde determinant, cf. formula (3.5.4.20).) Somewhat after his work, a full analog of the Weyl character formula was established for appropriate representations of “affine Lie algebras” and even more general Kac-Moody algebras [Kac4, GaLe]. In a more recent reconsider-

ation of his identities, Macdonald drew attention to certain finite truncations of the infinite products involved and formulated some conjectures regarding their evaluation. Macdonald's conjectures were also extended by Morris [Morrr]. In a flurry of activity by numerous authors, these have recently been established for most affine root systems [GaGo, Gust, Habs, Stem, Zeil1, 2].

We also mention a closely related, but rather different line of work pursued by Gustafson and Milne [GuMi1–3], who formulate a notion of “well-poised hypergeometric series.” These ideas are inspired by the efforts of Biedenharn, with Louck and others [BiLo2], to understand tensor products of representations in an explicit way.

(ii) *Kazhdan-Lusztig polynomials.* The rather different incarnations of the Kostka coefficients, involving both infinite-dimensional representations of Lie groups, and representations of finite reductive groups, provides an easily grasped example of the intimate connections that exist between these superficially different topics. Probably the most striking such example is provided by the Kazhdan-Lusztig polynomials [KaLu1, Shi]. These polynomials are defined in quite a technical way, in connection with what might appear to be a minor issue in the structure theory of “Hecke algebras for  $G/B$ ”; but they turn out to have some extraordinary connections with phenomena in topology, algebraic geometry, and representation theory of Lie groups [Shi, Lu, BoBM].

We review briefly the construction/definition of the Kazhdan-Lusztig polynomials. Let  $\Sigma$  be a finite root system,  $W$  the associated Weyl group, and  $R$  a generating set of fundamental reflections. Define the Hecke algebra  $\mathcal{H}_W$  associated to  $W$  to be the algebra over the field  $\mathbb{C}(q^{1/2})$  of rational functions in an indeterminate  $q^{1/2}$  generated by elements  $T_s$ ,  $s \in R$ , subject to relations

$$(4.1.3.2a) \quad T_s^2 = (q - 1)T_s + q,$$

$$(4.1.3.2b) \quad T_s^{\beta_{rs}}(T_r T_s)^{\alpha_{rs}} = T_r^{\beta_{rs}}(T_s T_r)^{\alpha_{rs}}, \quad r \neq s; r, s \in R,$$

where  $m_{rs} = 2\alpha_{rs} + \beta_{rs}$ ,  $0 \leq \beta_{rs} \leq 1$ , is the order of the element  $rs$  in  $W$ .

Some readers may wonder why we use  $\mathbb{C}(q^{1/2})$  as coefficients, when only  $q$ , not  $q^{1/2}$ , is involved in the defining relations (4.1.3.2). These readers should ask Lusztig.

If we specialize  $q$  to be the power of a prime, then  $\mathcal{H}_W$  has a direct interpretation as the algebra of intertwining operators for the induced representation  $\text{ind}_B^G 1$ , where  $G$  is the Chevalley group, over the finite field  $F_q$ , associated to the root system  $R$ , and  $B \subseteq G$  is a Borel subgroup [Spri, Crtr]. If  $q = 1$ , then relations (4.1.3.2) just reduce to the defining relations for  $W$ , so we get the group algebra of  $W$ . Thus  $\mathcal{H}_W$  may be thought of as a one-parameter family of algebras containing  $\mathbb{C}(W)$  and all the intertwining algebras for  $\text{ind}_{B(F_q)}^{G(F_q)} 1$  as  $F_q$  varies over all finite fields; in other words, we may regard the intertwining algebras as “deformations” of  $\mathbb{C}(W)$ . Using

this observation it is fairly easy to show [Iwah2] that all the intertwining algebras are isomorphic to  $\mathbf{C}(W)$ . This fact accounts, at least in a philosophical way, for phenomena such as the persistence of the Kostka coefficients across characteristics.

Consider an element  $w$  in  $W$ . Factor

$$(4.1.3.3) \quad w = s_1 s_2 \cdots s_k, \quad s_j \in R.$$

The minimum number  $k$  of factors in equation (4.1.3.3) is called the *length* of  $w$  and written  $l(w)$ . Given a factorization (4.1.3.3) of  $w$ , with  $k = l(w)$ , define

$$(4.1.3.4) \quad T_w = T_{s_1} T_{s_2} \cdots T_{s_k}.$$

The relation (4.1.3.2)(b) guarantees that  $T_w$  is well defined, i.e., independent of the minimal length factorization (4.1.3.3) of  $w$ . Then the relations (4.1.3.2)(a) and (b) together imply that the  $T_w$ ,  $w \in W$ , define a basis of  $\mathcal{H}_W$ .

We recall [Hilr, Dixm1] that there is defined on  $W$  a partial order, the *Bruhat order*. We say  $u \leq w$  if there is some minimal expression (4.1.3.3) for  $w$  from which we may obtain  $u$  by simply deleting some of the  $s_j$ . In geometric terms, if  $G$  is the (say, complex) semisimple group attached to the root system  $\Sigma$  and  $B \subseteq G$  is the Borel subgroup for which  $R$  is the set of fundamental generators of  $W$ , then  $u \leq w$  if and only if  $BuB$  is contained in the (Zariski) closure of  $BwB$ .

Following Kazhdan and Lusztig [KaLu1] we define an involution  $a \rightarrow \bar{a}$  of the algebra  $\mathcal{H}_W$  as follows:

$$(4.1.3.5) \quad \bar{q} = q^{-1} \quad (\text{i.e., } (q^{1/2})^- = q^{-1/2}), \quad \bar{T}_w = (T_w)^{-1}$$

It is easy to check that definitions (4.1.3.5) preserve the defining relations (4.1.3.2) of  $\mathcal{H}_W$ , so that  $a \rightarrow \bar{a}$  extends uniquely to an automorphism of  $\mathcal{H}_W$  (as an algebra over  $\mathbf{C}$ ).

**THEOREM 4.1.3.6 [KaLu1].** *For each pair  $(y, w)$  of elements of  $W$ ,  $y \leq w$ , there is a unique polynomial  $P_{y,w}(q)$  of degree at most  $\frac{1}{2}(l(w) - l(y) - 1)$ , such that  $P_{ww} = 1$  for each  $w \in W$ , and the element*

$$C_w = \sum_{y \leq w} (-1)^{l(w)+l(y)} q^{l(w)/2 - l(y)} \bar{P}_{y,w} T_w$$

satisfies  $C_w = \bar{C}_w$ .

This is proved by induction on  $w$ . For example,

$$C_s = q^{-1/2} T_s - q^{1/2}, \quad s \in R.$$

The  $P_{y,w}$  are the *Kazhdan-Lusztig polynomials*. They have the following

rather amazing connections with geometry and representation theory:

(4.1.3.7)(a) The  $P_{y,w}$  are the Poincaré polynomials for the local intersection cohomology (see [GoMP1, 2, Kirw3] for a description of this) at  $y$  in the Schubert variety  $((BwB)/B)^-$  in the flag variety  $G/B$ . Here  $G, B$  are, as above, the complex Chevalley group and Borel subgroup associated to  $\Sigma$  and  $R$ . See [Crtr, Spri].

(b) The values of  $P_{y,w}$  at  $q = 1$  describe how the Verma modules (cf. §3.5.3) of fixed infinitesimal characters break up into irreducible highest weight modules. This transfers to a description of the composition series of principal series for complex Lie groups, at least when the infinitesimal character is the infinitesimal character of a finite-dimensional representation. This was conjectured in [KaLu], and proved by Bernstein-Beilinson [BeBe] and Brylinski-Kashiwara [BrKa]. The description of composition series was extended to all semisimple Lie groups by Vogan [Voga7].

The Kazhdan-Lusztig polynomials can be used to express other quantities of interest in representation theory, notably the structure of primitive ideals in the universal enveloping algebra of a semisimple Lie algebra [BeBe, BrKa, Shi]. They can also be used to express the Kostka coefficients [Lus3; Shi, §2.7]. Unfortunately, the  $P_{y,w}$  are themselves rather difficult to compute. Nevertheless, the defining conditions of Theorem 4.1.3.6 define the  $P_{y,w}$  recursively, so they can in principle be computed mechanically. This “in principle” caveat applies to many Lie-theoretic quantities one would like to know. It is a challenge to devise more effective means of computation in particular situations of interest.

**4.2. Automorphic forms.** The theory of automorphic forms is a major customer and source of problems for representation theory, to the extent that it is sometimes difficult to draw a boundary between the two fields. This state of affairs is the culmination of a long gradual approach, going back at least to the theory of binary quadratic forms and the demonstration by Jacobi of the functional equations for his  $\theta$ -functions by means of the Poisson summation formula [Lang1, Rade]. In recent years, the intimacy between the two subjects was strongly fostered by a constellation of conjectures figured by R. P. Langlands [Lgld3, 6], which, supplemented, refined, and amended by various followers, are generally known under the rubric “Langlands program.” These conjectures envision a vast web of “reciprocity laws” (the term comes from quadratic reciprocity through Artin reciprocity) linking “geometric objects” (naively, algebraic varieties defined over  $\mathbf{Q}$  or another number field; more sophisticatedly schemes over  $\mathbf{Z}$ ; or now “motives,” a rather more elusive notion) with “automorphic representations,” the representation-theoretic refinement of the notion of automorphic form, by means of various classes of  $L$ -functions. Expositions or developments at several levels of the Langlands program have been published in recent years [Bor15, Gelb, ArCl, Roga1, CIMi, Lgld9].

Here we would like to explain how the study of automorphic forms motivates broadening the purview of representation theory to include not only representations of Lie groups, discussed in §3, but also of Lie group analogs—algebraic groups defined over  $p$ -adic (nonarchimedean local) fields.

With some oversimplification, the main problem of the theory of automorphic forms may be said to be the spectral decomposition of  $L^2(G/\Gamma)$ , where  $G$  is a Lie group and  $\Gamma \subseteq G$  is a discrete subgroup. That is, we let  $G$  act on  $L^2(G/\Gamma)$  by left translations, and we want to decompose  $L^2(G/\Gamma)$  into (a direct integral of) irreducible representations of  $G$ . This formulation is in itself the product of a very substantial conceptual development. It may not appear clearly in traditional presentations [Lang, Rade, Scho] of automorphic forms. Originally interest was attached to certain functions on  $G/\Gamma$  with special properties. However, it was gradually realized that many of the key properties of these functions (especially Euler product expansions and criteria for their existence) were naturally expressed in terms of the representation generated by the function and that the function could be retrieved from the representation as a special vector in the space of the representation. This yoga is implicit, for example, in Langlands' (partial) definition of an  $L$ -function for every automorphic form [Lgld3]. Usually one is interested in the case where  $\Gamma$  is a lattice in  $G$ , in fact an arithmetic subgroup (cf. Endnote 4 of §1). We will keep in mind as basic examples the group  $SL_n(\mathbf{Z}) \subseteq SL_n(\mathbf{R})$ , the group of determinant one  $n \times n$  matrices with integer entries inside the group of determinant one  $n \times n$  real matrices, and  $Sp_{2n}(\mathbf{Z}) \subseteq Sp_{2n}(\mathbf{R})$ , the group of symplectic  $2n \times 2n$  matrices with integer entries inside the group of all real  $2n \times 2n$  symplectic matrices.

A key feature of arithmetic groups is that they come in families, defined by congruence conditions. Thus, if  $\Gamma_1 = SL_n(\mathbf{Z})$  or  $Sp_{2n}(\mathbf{Z})$  or other arithmetic group, and  $m$  is any positive integer, we define  $\Gamma_m$ , the  $m$ th *principal congruence subgroup* to be the subgroup of elements  $\gamma$  of  $\Gamma_1$  which are congruent to 1 (the identity matrix) modulo  $m$ , in the sense that the entries of  $\gamma - 1$  are all divisible by  $m$ :

$$(4.2.1) \quad \Gamma_m = \{\gamma \in \Gamma_1 : (\gamma - 1) \equiv 0 \pmod{m}\}.$$

More generally, a *congruence subgroup*  $\Gamma \subseteq \Gamma_1$  is any subgroup which contains  $\Gamma_m$  for some  $m$ . (Clearly, any congruence subgroup of  $\Gamma_0$  will have finite index in  $\Gamma_1$ .) The *congruence subgroup problem* asks when the converse is true: when is any subgroup of finite index in  $\Gamma_1$  a congruence subgroup? It is often, in fact usually, the case, or almost the case [Bak, BaMS, Ragh2, PrRa]. But it, like rigidity (cf. §1.5.2), fails for  $SL_2(\mathbf{R})$ . This failure makes  $SL_2(\mathbf{R})$  useful for a variety of problems, including the realization of Galois groups (cf. [FeFo, Frie, Matz, Thom], etc.).

The theory of automorphic forms asks not simply about  $L^2(G/\Gamma_0)$  for a fixed arithmetic subgroup  $\Gamma_0$ , but wishes to describe  $L^2(G/\Gamma)$  for all congruence subgroups of  $\Gamma$ . Thus in the classical theory of modular forms, one

speaks of the *level* of a modular form; a rough translation to the context of our discussion would take “level” to mean the smallest  $n$  such that a given function (or the representation it generates) lives on  $G/\Gamma_n$ .

If  $\Gamma'' \subset \Gamma' \subseteq \Gamma_1$  are two arithmetic subgroups, then there is an obvious inclusion

$$L^2(G/\Gamma') \hookrightarrow L^2(G/\Gamma''),$$

so we may consider the union or inductive limit

$$(4.2.2) \quad \mathcal{L}(G, \Gamma_1) = \bigcup_{\Gamma \subseteq \Gamma_1} L^2(G/\Gamma)$$

over all arithmetic subgroups  $\Gamma \subseteq \Gamma_1$ . The theory of automorphic forms aspires to describe  $\mathcal{L}(G, \Gamma_1)$ , including its level structure, namely where a given representation sits in the hierarchy of  $L^2(G/\Gamma)$  inside  $\mathcal{L}(G, \Gamma_1)$ .

It was Hecke [Heck] who first observed (in a context where the issues were considerably more obscure) that  $\mathcal{L}(G, \Gamma_1)$  allows substantially more symmetry than just the action of  $G$  by left translations. Implicit in the definition of  $\Gamma_1$  is a group  $G_{\mathbf{Q}}$ , the rational points of  $G$ . We have  $\Gamma_1 \subseteq G_{\mathbf{Q}} \subseteq G$ , and  $G_{\mathbf{Q}}$  is dense in  $G$ . Thus, if  $\Gamma_1 = \mathrm{SL}_n(\mathbf{Z})$  and  $G = \mathrm{SL}_n(\mathbf{R})$ , then  $G_{\mathbf{Q}} = \mathrm{SL}_n(\mathbf{Q})$ ; and similarly for the example of the symplectic group. The group  $G_{\mathbf{Q}}$  has the property that for any  $g$  in  $G_{\mathbf{Q}}$ , and any congruence subgroups  $\Gamma', \Gamma''$  in  $\Gamma_1$ , the intersection  $(g\Gamma'g^{-1}) \cap \Gamma''$  is again a congruence subgroup. (The reader should find this easy to verify for  $\Gamma_1 = \mathrm{SL}_n(\mathbf{Z})$ ,  $G_{\mathbf{Q}} = \mathrm{SL}_n(\mathbf{Q})$ .) We should note that  $\mathcal{L}(G, \Gamma_1)$  in fact depends only on  $G_{\mathbf{Q}}$ , not on a particular arithmetic subgroup  $\Gamma_1$ . Because of this, we will write  $\mathcal{L}(G, G_{\mathbf{Q}})$ .

Consider  $f \in L^2(G/\Gamma)$ . For  $g \in G_{\mathbf{Q}}$ , consider the right translate of  $f$  by  $g$ :

$$R_g(f)(x) = f(xg), \quad x \in G.$$

If  $\gamma \in \Gamma \cap g\Gamma g^{-1}$ , then

$$R_g(f)(x\gamma) = f(x\gamma g) = f(xg(g^{-1}\gamma g)) = f(xg) = R_g(f)(x).$$

Hence  $R_g(f)$  belongs to  $L^2(G/(\Gamma \cap g\Gamma g^{-1}))$ . We have shown the following:

LEMMA 4.2.3. *Right translation by  $g \in G_{\mathbf{Q}}$  preserves  $\mathcal{L}(G, G_{\mathbf{Q}})$ .*

It is obvious that  $R_g$  commutes with the action of  $G$  on  $\mathcal{L}(G, G_{\mathbf{Q}})$  by left translations. Thus  $\mathcal{L}(G, G_{\mathbf{Q}})$  is actually a  $(G \times G_{\mathbf{Q}})$ -module. We may go further. We observe that the action of  $G_{\mathbf{Q}}$  is of a special sort: any given  $f$  in  $\mathcal{L}(G, G_{\mathbf{Q}})$  is stabilized by some congruence subgroup  $\Gamma \subseteq \Gamma_1$ . This makes reasonable, if it does not directly suggest, the following construction which extends the action of  $G_{\mathbf{Q}}$  to a larger group, obtained by a certain process of completion. Define a topology on  $G_{\mathbf{Q}}$  by considering the congruence subgroups  $\Gamma \subseteq \Gamma_1$  to be open subgroups of  $G_{\mathbf{Q}}$ . It is easy to check that

this makes  $G_Q$  into a Hausdorff topological group [HeRo, Loom, MoZi2]. A topological group has a natural sense of uniform structure, hence of Cauchy sequence: a sequence  $\{\gamma_j\}_{j=1}^\infty$  is Cauchy, provided  $\gamma_j \gamma_k^{-1}$  converges to the identity. This is the same as saying  $\gamma_j \gamma_k^{-1} \in \Gamma_m$  for any  $m$  and all sufficiently large  $j, k$ . Denote by  $G_{A_f}$  the completion [Kell] of  $G_Q$  with respect to this uniform structure. Then, because of the nature of the action of  $G_Q$  on  $\mathcal{L}(G, G_Q)$ , as noted above, it is easy to check that the action of  $G_Q$  on  $\mathcal{L}(G, G_Q)$  extends continuously to an action of  $G_{A_f}$ . Thus  $\mathcal{L}(G, G_Q)$  is a  $(G \times G_{A_f})$ -module.

To make this statement more concrete, let us examine the structure of  $G_{A_f}$ . First, look at the topology of  $G_{A_f}$ . By definition, the closure  $\bar{\Gamma}$  in  $G_{A_f}$  of a congruence subgroup  $\Gamma$  will be open in  $G_{A_f}$ . Since the congruence subgroups of  $\Gamma_1$  all have finite index in  $\Gamma_1$ , it is easy to see that any sequence  $\{\gamma_g\}$  in  $\Gamma$  will have a convergent subsequence in  $G_{A_f}$ . In other words,  $\bar{\Gamma}$  will be open and compact in  $G_{A_f}$ . It is elementary to check that an open subgroup of a topological group is also closed. Thus  $G_{A_f}$  has a system  $\{\bar{\Gamma}\}$  of compact, open, and (hence) closed subgroups which form a neighborhood base for the identity in  $G_{A_f}$ . In other words,  $G_{A_f}$  is a totally disconnected, locally compact group.

Look also at the algebraic structure of  $G_{A_f}$ . For each  $m$ , the completion  $\bar{\Gamma}_m$  of  $\Gamma_m$  will be a normal, open subgroup of  $\bar{\Gamma}_1$ , and we will have

$$\bar{\Gamma}_1 / \bar{\Gamma}_m \simeq \Gamma_1 / \Gamma_m.$$

A review of the definition of  $\bar{\Gamma}$  reveals that it may be regarded as an inverse limit [HeRo, KeNa, Lang3] of the  $\Gamma_1 / \Gamma_m$ :

$$(4.2.4) \quad \bar{\Gamma} = \varprojlim \Gamma_1 / \Gamma_m.$$

To get a feel for the structure of  $\bar{\Gamma}$ , consider the case of

$$N = \left\{ \begin{bmatrix} 1 & x \\ 0 & 1 \end{bmatrix} : x \in \mathbf{R} \right\}$$

and its subgroup  $\Gamma_1$ , of integral matrices, which is isomorphic to  $\mathbf{Z}$ . We see then that  $\Gamma_1 / \Gamma_m \simeq \mathbf{Z} / m\mathbf{Z}$ , so that

$$\bar{\Gamma}_1 \simeq \varprojlim \mathbf{Z} / m\mathbf{Z}.$$

The Chinese Remainder Theorem [Jaco2, Lang3] tells us that if  $m = \prod_p p^{j_p}$  is the prime factorization of  $m$ , then

$$\mathbf{Z} / m\mathbf{Z} \simeq \prod_p \mathbf{Z} / (p^{j_p} \mathbf{Z}).$$

It follows in the inverse limit that  $\bar{\Gamma}_1 \simeq \prod_p \bar{\mathbf{Z}}_p$ , where

$$(4.2.5) \quad \bar{\mathbf{Z}}_p = \varprojlim \mathbf{Z} / (p^n \mathbf{Z}).$$



The group  $\overline{\mathbf{Z}}_p$ , which, in fact, inherits from  $\mathbf{Z}$  the structure of ring as well as additive group, is known as the *p-adic integers*. It is an integral domain: if  $x \in \mathbf{Z} - p^k \mathbf{Z}$ , and  $y \in \mathbf{Z} - p^l \mathbf{Z}$ , then  $xy \notin p^{k+l} \mathbf{Z}$ . It has a unique prime ideal  $p\overline{\mathbf{Z}}_p$ . Its field of quotients is called  $\mathbf{Q}_p$ , the *p-adic numbers*. It is a non-discrete, locally compact field—for short, a *local field*. The embedding  $\mathbf{Z} \rightarrow \overline{\mathbf{Z}}_p$  extends to an embedding  $\mathbf{Q} \rightarrow \mathbf{Q}_p$  with dense image. Thus  $\mathbf{Q}_p$  may be regarded as a *completion* of  $\mathbf{Q}$ : a local field in which  $\mathbf{Q}$  embeds densely. A classical result [Weil2] asserts that the fields  $\mathbf{Q}_p$ , the *p-adic numbers* for a prime  $p$ , and  $\mathbf{R}$ , the reals, constitute all possible completions of  $\mathbf{Q}$ . More generally, any local field of characteristic zero is either  $\mathbf{R}$ ,  $\mathbf{C}$ , or a finite extension of  $\mathbf{Q}_p$  for some  $p$  [Weil2].

The product

$$(4.2.6) \quad \mathbf{Z}_{A_f} = \prod_p \overline{\mathbf{Z}}_p$$

over all primes  $p$  of the *p-adic integers* is, like the individual factors  $\overline{\mathbf{Z}}_p$ , a compact ring. It is not, of course, a domain, because the product of two elements from different factors will be zero. However, it contains  $\mathbf{Z}$  (embedded diagonally) as a dense subring, and if we invert elements of  $\mathbf{Z}$  we obtain a ring  $A_f$ , the ring of *finite adeles*. Since a given  $m \in \mathbf{Z}$  is invertible in  $\overline{\mathbf{Z}}_p$  if and only if  $p$  does not divide  $m$ , we can see that  $A_f$  may be described as follows. For every finite set  $S = \{p_1, p_2, \dots, p_s\}$  of primes, let  $A_S$  be the ring obtained from  $\mathbf{Z}_{A_f}$  by inverting the primes in  $S$ . It is easy to see that

$$(4.2.7) \quad A_S \simeq \left( \prod_{p \in S} \mathbf{Q}_p \right) \times \prod_{p \notin S} \overline{\mathbf{Z}}_p.$$

One has further that

$$(4.2.8) \quad A_f = \bigcup_S A_S.$$

Another description, easily checked to be equivalent to (4.2.8), is that  $A_f$  is the *restricted direct product* of the  $\mathbf{Q}_p$  with respect to the  $\overline{\mathbf{Z}}_p$ : the set of sequences  $(x_2, x_3, x_5, \dots)$ , where  $x_p \in \mathbf{Q}_p$ , and, for almost all (in the sense: all but a finite number)  $p$ , we have  $x_p \in \overline{\mathbf{Z}}_p$ . The topology on  $A_f$  is such that each  $A_S$  is an open subring. In particular, the ring  $\mathbf{Z}_{A_f}$  is open in  $A_f$ , and  $A_f$  is a locally compact ring.

The ring  $A_f$  of finite adeles can be used to describe the results of our completion construction for a general group  $G_{\mathbf{Q}}$ , as we anticipated by using the notation  $G_{A_f}$  for this completion. Thus for  $G = \mathrm{SL}_n(\mathbf{R})$ ,  $G_{\mathbf{Q}} = \mathrm{SL}_n(\mathbf{Q})$ , and  $\Gamma_1 = \mathrm{SL}_n(\mathbf{Z})$ , one has  $\overline{\Gamma}_1 = \prod_p \mathrm{SL}_n(\overline{\mathbf{Z}}_p)$  and  $G_{A_f} = \mathrm{SL}_n(A_f)$ , which can be described either as the group of matrices of determinant 1 with coefficients in  $A_f$ , or as the restricted direct product of the groups  $\mathrm{SL}_n(\mathbf{Q}_p)$  with

respect to the open compact subgroups  $SL_n(\mathbf{Z}_p)$ . With some caveats, this is the general pattern: if  $G_{\mathbf{Q}}$  is described as a group of matrices with rational entries satisfying some equations, then  $G = G_{\mathbf{R}}$  is the group of matrices satisfying the same equations, but with real numbers as entries; the  $G_{\mathbf{A}_f}$  are the matrices satisfying these equations, but with entries in  $\mathbf{A}_f$ ; and, up to finite index,  $\bar{\Gamma}_f$  is the product of the  $G(\bar{\mathbf{Z}}_p)$ .

Let us summarize the consequences of this discussion for the structure of  $\mathcal{L}(G, G_{\mathbf{Q}})$ . We see we have found that  $\mathcal{L}(G, G_{\mathbf{Q}})$  is a module not just for the Lie group  $G = G_{\mathbf{R}}$  but for a product group  $G_{\mathbf{R}} \times G_{\mathbf{A}_f}$ , where the factor  $G_{\mathbf{A}_f}$  is itself a (restricted) product of groups  $G_{\mathbf{Q}_p}$ , for all primes  $p$ . The factors  $G_{\mathbf{Q}_p}$  of  $G_{\mathbf{A}_f}$  look "just like"  $G_{\mathbf{R}}$ , in the sense that they are a group of matrices satisfying the same equations as the equations defining  $G_{\mathbf{Q}_p}$ , only the entries of the matrices are in  $\mathbf{Q}_p$  rather than  $\mathbf{R}$ .

The previous paragraph suggests that  $G_{\mathbf{R}}$  and the  $G_{\mathbf{Q}_p}$  are on an essentially equal footing as far as  $\mathcal{L}(G, G_{\mathbf{Q}})$  is concerned. This is so. We give a slight reformulation of the previous paragraph which emphasizes this viewpoint. Set

$$(4.2.9) \quad \mathbf{A} = \mathbf{R} \times \mathbf{A}_f.$$

This is called the ring of *adeles*. Since  $\mathbf{R}$  is a completion of  $\mathbf{Q}$ , just as are the  $\mathbf{Q}_p$ , we have the diagonal embedding  $\mathbf{Q} \rightarrow \mathbf{A}$ . Slight extension of the discussion of  $\mathbf{A}_f$  [Lang4, Weil2] shows that

$$(4.2.10) \quad \begin{aligned} & \text{(i) } \mathbf{Q} \cap \left( \mathbf{R} \times \prod_p \bar{\mathbf{Z}}_p \right) = \mathbf{Z}; \text{ hence} \\ & \text{(ii) } \mathbf{Q} \text{ is discrete in } \mathbf{A}, \text{ and} \\ & \text{(iii) } \mathbf{A}/\mathbf{Q} \text{ is compact.} \end{aligned}$$

We can also form  $G_{\mathbf{A}} = G_{\mathbf{R}} \times G_{\mathbf{A}_f}$ . Under often satisfied assumptions on  $G$  [Pras, Plat], analogs of facts (4.2.10) hold for  $G$  also.

$$(4.2.11) \quad \begin{aligned} & \text{(i) } G_{\mathbf{Q}} \subseteq G_{\mathbf{A}} \text{ is a discrete subgroup.} \\ & \text{(ii) The quotient space } G_{\mathbf{Q}}/G_{\mathbf{A}} \text{ has finite volume.} \\ & \text{(iii) If } \Gamma \subseteq G_{\mathbf{Q}} \text{ is an arithmetic group, and } \bar{\Gamma} \text{ is the} \\ & \quad \text{completion of } \Gamma \text{ in } G_{\mathbf{A}_f}, \text{ then } \bar{\Gamma} \text{ is compact, and} \\ & \quad \text{open in } G_{\mathbf{A}_f}. \text{ Further } G_{\mathbf{R}} \times \bar{\Gamma} \text{ is open in } G_{\mathbf{A}}, \text{ and} \\ & \quad G_{\mathbf{Q}} \cap (G_{\mathbf{R}} \times \bar{\Gamma}) = \Gamma. \\ & \text{(iv) } G_{\mathbf{R}} \cdot G_{\mathbf{Q}} \text{ is dense in } G_{\mathbf{A}}. \end{aligned}$$

Property (iv) is known as *strong density* [Pras, Plat]. It follows from points (iii) and (iv), that the inclusion  $G_{\mathbf{R}} \hookrightarrow G_{\mathbf{A}}$  gives rise to an identification

$$(4.2.12a) \quad G_{\mathbf{R}}/\Gamma \simeq \bar{\Gamma} \backslash G_{\mathbf{A}}/G_{\mathbf{Q}}.$$

Observe that since  $G_{\mathbf{R}}$  commutes with  $\bar{\Gamma}$ , it will act on  $\bar{\Gamma} \backslash G_{\mathbf{A}}/G_{\mathbf{Q}}$  on the left, and (4.2.12a) is a  $G_{\mathbf{R}}$ -equivariant identification. Of course (4.2.12a) leads to an identification of  $L^2$  spaces:

$$(4.2.12b) \quad L^2(G_{\mathbf{R}}/\Gamma) \simeq L^2(\bar{\Gamma} \backslash G_{\mathbf{A}}/G_{\mathbf{Q}}).$$

All the spaces  $L^2(\bar{\Gamma} \backslash G_{\mathbf{A}}/G_{\mathbf{Q}})$  may be considered as subspaces of  $L^2(G_{\mathbf{A}}/G_{\mathbf{Q}})$ . Doing this, we see that the inclusions  $L^2(G_{\mathbf{R}}/\Gamma') \subseteq L^2(G_{\mathbf{R}}/\Gamma'')$  when  $\Gamma'' \subseteq \Gamma'$  are simply inclusions of subspaces of  $L^2(G_{\mathbf{A}}/G_{\mathbf{Q}})$ . Thus the space  $\mathcal{L}(G, G_{\mathbf{Q}})$  of equation (4.2.2) is also seen as a subspace of  $L^2(G_{\mathbf{A}}/G_{\mathbf{Q}})$ —precisely the subspace of all vectors which are fixed by some open subgroup of  $G_{\mathbf{A}_f}$ . Clearly  $\mathcal{L}(G, G_{\mathbf{Q}})$  is dense in  $L^2(G_{\mathbf{A}}/G_{\mathbf{Q}})$ ; or from the point of view of (4.2.2),  $L^2(G_{\mathbf{A}}/G_{\mathbf{Q}})$  is the result of taking the Hilbert space completion of  $\mathcal{L}(G, G_{\mathbf{Q}})$ —in some sense it is the result of following construction (4.2.2) to its natural end. For groups not satisfying strong density (property (4.2.11)(iv)), the connection between  $L^2(G/\Gamma)$  and  $L^2(G_{\mathbf{A}}/G_{\mathbf{Q}})$  is more complicated than (4.2.12b), but the adelic viewpoint is still illuminating.

In summary, we find that, if we are interested in describing the spaces  $L^2(G_{\mathbf{R}}/\Gamma)$  for all congruence subgroups  $\Gamma$  of the arithmetic subgroup  $\Gamma_1$  and the relations between these spaces, then essentially we are interested in  $L^2(G_{\mathbf{A}}/G_{\mathbf{Q}})$ . This space supports an action not simply of  $G_{\mathbf{R}}$  but of the adèle group  $G_{\mathbf{A}}$ , in which each  $p$ -adic completion  $G_{\mathbf{Q}_p}$  of  $G_{\mathbf{Q}}$  participates on an equal footing with  $G_{\mathbf{R}}$ . The unitary dual of  $G_{\mathbf{A}}$  is identifiable to the (restricted) *product* of the unitary duals of the local factors  $G_{\mathbf{Q}_p}$ , via a tensor product construction [Flat]. Thus an irreducible  $G_{\mathbf{A}}$ -subspace of  $L^2(G_{\mathbf{A}}/G_{\mathbf{Q}})$ —which is called an *automorphic representation*—carries vastly more information than simply the  $G_{\mathbf{R}}$ -isomorphism class that it defines: it also determines a point in  $\hat{G}_{\mathbf{Q}_p}$  for all primes  $p$ . Langlands [Lgld3] has shown how to use the parameters from all the  $\hat{G}_{\mathbf{Q}_p}$  to define (for almost all primes) local Euler factors, to be multiplied together to form an  $L$ -function (the Langlands automorphic  $L$ -function) to be attached to the automorphic representation. One can then hope that this automorphic  $L$ -function is equal to an  $L$ -function attached to some geometric object, which equality would constitute a reciprocity law of some sort. A large number of workers have been engaged in this project, and have made notable progress [GeSh, JaLa, ArCl, Roga, CIMi, GePs], including the establishment of some nonobvious cases of the Artin conjecture on holomorphicity of  $L$ -functions [Lgld5, Tunn]. However, in some sense, the work has only begun.

REMARK. At this point, we should perhaps deal with an issue that may have been bothering the reader, namely what if we do indeed wish to be very classical and deal with  $L^2(G/\Gamma)$  for fixed  $\Gamma$ . Individual elements of  $G_{\mathbf{A}_f}$  do

not preserve individual spaces  $L^2(G/\Gamma)$ , only the system  $\mathcal{L}(G, G_Q)$  formed from all  $L^2(G/\Gamma)$ . However, when we make the identification  $L^2(G/\Gamma) \simeq L^2(\bar{\Gamma} \backslash G_A/G_Q)$ , we see that the convolution algebra  $C_c(G_{A_f}/\bar{\Gamma})$ , of compactly supported functions on  $G_{A_f}$  which are both left and right invariant under  $\bar{\Gamma}$ , will preserve  $L^2(\bar{\Gamma} \backslash G_A/G_Q)$ . Interpreted in terms of  $L^2(G/\Gamma)$ , elements of  $C_c(G_{A_f}/\bar{\Gamma})$  are the original Hecke operators [Heck, Lang5, Shim1]. For this reason,  $C_c(G_{A_f}/\bar{\Gamma})$ , is sometimes called a *Hecke algebra*. If  $\bar{\Gamma} = \prod_p \bar{\Gamma}_p$  is a product of local factors, then also  $C_c(G_{A_f}/\bar{\Gamma})$  can be decomposed as a (restricted) tensor product

$$(4.2.13) \quad C_c(G_{A_f}/\bar{\Gamma}) \simeq \bigotimes_p C(G_{Q_p}/\bar{\Gamma}_p)$$

of “local” Hecke algebras  $C(G_{Q_p}/\bar{\Gamma}_p)$ . These local Hecke algebras (also known as spherical function algebras) are very interesting. They control the representation theory of  $G_{Q_p}$  [HoMy1, 2, BDKM, BuKu]. Also for a particular choice of  $\bar{\Gamma}_p$  (the Iwahori subgroup [Iwah2, Roga2]), they are related to quantum groups, knot theory, etc. [Cher2, GoHJ] (cf. the articles by Jones and Witten in this volume).

The adelic formulation of the theory of automorphic forms raises the question of understanding representations of  $G_{Q_p}$  as well as representations of  $G_{\mathbf{R}}$ . The issue of representations of  $p$ -adic groups has been studied since the early 1960s. Considerable progress has been made, but substantial mysteries remain: the situation is not nearly so complete as in the real case. The construction of principal series by means of induction from parabolic subgroups works as for  $F$ , and there is a parallel of “Langlands classification” (Theorem 3.6.4.5) [Silb, BeZe]. However, the essence of Theorem 3.6.4.5 is that it describes general admissible representations in terms of discrete series, and the discrete series for  $p$ -adic groups remain obscure. It is roughly true that, as for real groups, discrete series are attached to characters of compact Cartan subgroups. However, in the  $p$ -adic case there are typically many compact Cartan subgroups instead of at most one, as for real groups, and understanding precisely how they interact has been difficult. The group  $GL_2$ , and closely related groups, have been under fairly good control since the 1960s, [Sall, Shal, Silb, GGPS], although the case of residual characteristic 2 gave considerable trouble [Kutz, Tunn1]. Some other cases, where the geometry of the compact tori is relatively simple or well behaved, have been treated more or less completely [CoHo, HoMy1, 2, Moy1, 2, KoZi]. In recent significant progress, constructions of all discrete series for  $GL_n(Q_p)$  have been devised [BuKu, Corw1, 2, Henn2]. In [BuKu], as in [HoMy1, 2], the main point is to find many copies of the Iwahori Hecke algebras ([Iwah2,]

**Roga2**) in  $C_c(G_{Q_p})$ . The representations of Iwahori Hecke algebras have been described in [KaLu2].

In fact, Langlands [Lgd2, 3] proposed a parametrization of representations of  $G_{Q_p}$  (or  $G_R$ ) in terms of the Galois theory of  $Q_p$ . His ideas have been useful even for real (i.e., Lie) groups—they provide the rationale for the “Langlands classification” (Theorem 3.6.4.5), and the duality phenomenon that they suggest provides a deep organizational principle for representations of different real forms of the same complex group [Voga8, AdVo]. His proposal, which has gone through several refinements [Arth2, 3, Bor15, Lus25], and may well go through more, can be thought of as a nonabelian generalization of the local reciprocity law of local classfield theory [Lang4, Weil2, Lgd2].

To explain the gist of this idea, which applies more or less uniformly to  $G_R$  and  $G_{Q_p}$ , let us adopt a more neutral notation, and denote by  $F$  either  $R$  or  $Q_p$ . (In fact, the discussion below will apply equally well for  $F$  a local field of characteristic zero.) By  $G_F$  we will mean (the  $F$ -rational points of) an (affine) algebraic group defined over  $F$  (cf. §1.5.2, especially Endnote 4, also [Borel, Jant, Spri2], etc.); the reader may think of  $SL_n(F)$ ,  $Sp_{2n}(F)$ , etc. Also, let us remark that, just as for real reductive groups (cf. §3.6.5 and [Knap2, Wall2]) there is for  $p$ -adic reductive groups a notion of *admissible representation* [JaLa, Cart2], with analogous properties: the admissible representations contain the irreducible unitary representations [Bern 4, Cart 2], and constitute a sort of analytic continuation of them. Denote the set of irreducible admissible representations of  $G_F$  by  $\text{Adm}(G_F)$ .

The core of Langlands’ idea is that irreducible representations of  $G_F$  should be parametrized “naturally” by representations of  $\text{Gal}(F)$ , the (absolute) Galois group of  $F$  (which with its standard (inverse limit) topology is thought of as a compact, totally disconnected group [Lang3, p. 351]). To be useful, this idea needs considerable clarification and qualification.

The most important qualification regards the nature of the representations of  $\text{Gal}(F)$  used to parametrize  $\text{Adm}(G_F)$ . Since the sets  $\text{Adm}(G_F)$  vary considerably for varying  $G_F$ , it is not reasonable to parametrize all  $\text{Adm}(G_F)$  by the same representations of  $\text{Gal}(F)$ —one needs a way of segregating the representations of  $\text{Gal}(F)$  which should be associated to  $\text{Adm}(G_F)$ . Langlands’ proposal for doing this is to construct a complex Lie group  ${}^L G^0$ , the *L-group* (more correctly, its identity component), whose structure reflects that of  $G_F$ . The construction of  ${}^L G^0$  exploits a kind of duality in the family of semisimple groups [Lgd3, 4, Bor15]. The existence of this duality emerges from the abstract specification of semisimple groups in terms of Weyl groups and root systems, refining Theorem 2.12.2 [Spri2, Crtr]. In this duality, symplectic groups are matched with odd-dimensional orthogonal groups, while all other members of the Killing-Cartan classification (cf. §2.10) are matched with themselves on the Lie algebra level; however, centers also change with

duality:  $SL_n$  is dual to  $PGL_n$ . However,  $GL_n$  is dual to itself. Langlands suggests that, rather than look at representations in the usual sense of  $Gal(F)$ , which are homomorphisms to  $GL_m(\mathbb{C})$  for some  $m$ , we should consider homomorphisms to  ${}^L G^0$  to parametrize  $Adm(G_F)$ .

The second key qualification of the basic idea is that the Galois group is too confined to correctly reflect  $Adm(G_F)$ , because as one sees from parabolic induction, there are continuous families in  $Adm(G_F)$ , whereas  $Gal(F)$ , being compact, will have only a discrete set of representations. Here one takes a hint from abelian class field theory [Weil2, Lang4], and replaces  $Gal(F)$  by  $W(F)$ , the *Weil group* of  $F$  [Tate, ArTa]. When  $F$  is an extension of  $\mathbb{Q}_p$ , this is a group which fits in a diagram

$$(4.2.14a) \quad \begin{array}{ccc} W(F) & \longrightarrow & F^\times \\ \downarrow & & \downarrow \\ Gal(F) & \longrightarrow & Gal(F)^{ab} \end{array}$$

where  $Gal(F)^{ab}$  is the maximal abelian quotient of  $Gal(F)$ . The vertical maps are injections, and the right-hand one is the reciprocity map of abelian classfield theory (cf. [Weil2, Lang4], etc.). Thus  $W(F)$  is a loosened-up version of  $Gal(F)$ : it is a dense subgroup, equipped with a stronger topology so that it is again locally compact. For  $\mathbb{R}$  and  $\mathbb{C}$ , one has, by definition

$$(4.2.14b) \quad \begin{aligned} W(\mathbb{C}) &= \mathbb{C}^\times, \\ 1 &\rightarrow \mathbb{C}^\times \rightarrow W(\mathbb{R}) \rightarrow \mathbb{Z}/2\mathbb{Z} \rightarrow 1, \end{aligned}$$

where the nontrivial element of  $W(\mathbb{R})/\mathbb{C}^\times$  acts on  $\mathbb{C}^\times$  by complex conjugation, and has square equal to  $-1$  in  $\mathbb{C}^\times$ . For  $\mathbb{R}$  and  $\mathbb{C}$ , it is especially important to replace the Galois group, which is only of order 2 for  $\mathbb{R}$  and trivial for  $\mathbb{C}$ , by the much larger Weil group.

The main part of Langlands' proposal, then, is to parametrize  $Adm(G_F)$  by homomorphisms from  $W(F)$  to  ${}^L G^0$ . Further modifications are still necessary:  ${}^L G^0$  must be extended by  $W(F)$ , and the homomorphisms restricted in various ways, in order to better reflect the structure of  $G_F$ , as opposed to  $G_{\tilde{F}}$ , where  $\tilde{F}$  is the algebraic closure of  $F$ . Thus we want to be able to distinguish between  $GL_m(\mathbb{R})$  and the unitary groups  $U_{p,q}$ ,  $p+q=m$ . Also some extra data is needed, provided by using the Weil-Deligne group [Tate] instead of the Weil group, or adding a nilpotent element in the Lie algebra of  ${}^L G^0$  to the parametrizing data, or other means [Bor15]. The end of this process is a set  $\Phi(G_F)$  of "admissible homomorphisms" from  $W(F)$  to  ${}^L G^0$ , and these should parametrize  $Adm(G_F)$  up to finite ambiguity. That

is, there should be a map

$$(4.2.15) \quad \begin{array}{c} \text{Adm}(G_F) \\ \downarrow \\ \Phi(G_F) \end{array}$$

with finite fibers. The fibers are called *L-packets*, and two representations in the same fiber are called *L-indistinguishable*, to suggest there is no way to separate them using the theory of *L-functions*. This map should satisfy various desiderata [Bor15].

Although the definition of the map (4.2.15) is quite involved for general  $G$ , in special cases it has immediate intuitive impact. For example, for  $G_F = \text{GL}_n(\mathbf{Q}_p)$ , the map (4.2.15) amounts to a canonical bijection between the sets

- (i) irreducible  $n$ -dimensional complex representations of  $W(\mathbf{Q}_p)$ ,
- (ii) irreducible supercuspidal representations of  $\text{GL}_n(\mathbf{Q}_p)$ .

Supercuspidal representations are a remarkable phenomenon of  $p$ -adic representation theory: they are discrete series representations whose coefficients are not merely square-integrable, but have compact support (!). For a group like  $\text{GL}_n$ , with a noncompact center, terms like “square-integrable” or “completely supported” must be understood “modulo the center” [Cart2].

A parametrization of type (4.2.15) is known to exist for  $\text{GL}_l$  if  $l$  is prime, over all  $p$ -adic fields [Kutz, KuMo, Tunn1, Henn1], and for  $\text{GL}_n(F)$  if  $n$  is less than  $p$ , when  $F$  is an extension of  $\mathbf{Q}_p$  [Moy]. It is also known for all real groups [Lgld4, Bor15]. This is the true “Langlands classification.” Theorem 3.6.4.5 is a bowdlerized version, expressed solely in terms of the structure of  $G_{\mathbf{R}}$ , with reference to  $\Phi(G_{\mathbf{R}})$  expunged.

*L-indistinguishability* is connected with a concrete phenomenon in the theory of automorphic forms, a phenomenon which leads to tremendous complications: the failure of “strong multiplicity one.” We have noted that a representation  $\Pi$  of the adèle group  $G_{\mathbf{A}}$  is essentially constructed as a tensor product of representations  $\Pi_p$  of the local factors  $G_{\mathbf{Q}_p}$  of  $G_{\mathbf{A}}$ . In particular, representations  $\Pi$  and  $\Pi'$  of  $G_{\mathbf{A}}$  will be equivalent if and only if  $\Pi_p$  is equivalent to  $\Pi'_p$  for all  $p$  (including the case of  $G_{\mathbf{R}}$  as “the infinite prime” [Weil2, Lang4]). Let us call  $\Pi$  and  $\Pi'$  *nearly equivalent* if  $\Pi_p$  and  $\Pi'_p$  are equivalent for all but a finite number of  $p$ . Strong multiplicity one for  $G$  says that given two automorphic representations  $\Pi$  and  $\Pi'$  of  $G_{\mathbf{A}}$ , i.e., constituents of  $L^2(G_{\mathbf{A}}/G_{\mathbf{Q}})$ , then if  $\Pi$  and  $\Pi'$  are nearly equivalent, they *are* equal. In other words, if we are given representations  $\Pi_p$  of the local factors  $G_{\mathbf{Q}_p}$  of  $G_{\mathbf{A}}$ , for all but a finite number of  $p$ , and told to make an automorphic representation out of them, then there is at most one way to choose the remaining  $\Pi_p$  in order to do this and at most one way to put

the resulting representation of  $G_A$  in  $L^2(G_A/G_Q)$ . Strong multiplicity one is known to hold for  $GL_n$ , and inside large classes of automorphic forms for other groups [Piat]. In fact, for  $GL_n$ , knowing an automorphic representation at a sufficiently large *finite* number of places is enough to determine it completely [More]. However, for many groups, strong multiplicity one fails. Piatetski-Shapiro has been especially vigorous in providing examples of this failure, and of other peculiar phenomena of automorphic forms [CoPS1, 2, HoPS1, 2].

Thus, for a group (such as the symplectic groups) for which strong multiplicity one fails, we can find two distinct but nearly equivalent automorphic representations  $\Pi$  and  $\Pi'$ . Let  $\Pi_p$  and  $\Pi'_p$  be local components of  $\Pi$  and  $\Pi'$ . Then it should be the case that  $\Pi_p$  and  $\Pi'_p$  are  $L$ -indistinguishable, i.e., are in the same fiber of the parametrization map (4.2.14). This is the concrete meaning of  $L$ -indistinguishability. Langlands and his school have spent much effort in recent years grappling with problems arising from  $L$ -indistinguishability [LaLa, Lgld8, LaSh, Shel]. Current interest is focussed on Hasse-Weil zeta-functions of Shimura varieties (cf. [Lgld6, ClMi, Roga1, Miln], etc.). A major tool in this endeavor is the trace formula [Selb2, Arth4, 5, Labe, Roga1].

Even in its conjectural state, the parametrization (4.2.15) is of considerable interest, because of the astounding amount of structure it suggests in the representation theory of groups defined over a given local field. It points toward the existence of an intricate interlocking system of correspondences between representations of different groups. Let  $G_F$  and  $H_F$  be two groups over the local field  $F$ . Suppose there is a homomorphism

$$\alpha : {}^L G \rightarrow {}^L H$$

between their respective  $L$ -groups. If  $\varphi : W(F) \rightarrow {}^L G$  is a homomorphism from the Weil group to  ${}^L G$ , then  $\alpha \circ \varphi$  will be a homomorphism from  $W(F)$  to  $H$ . Thus, modulo technicalities, we would expect composition with  $\alpha$  to define a mapping

$$\alpha_* : \Phi(G_F) \rightarrow \Phi(H_F).$$

The obvious question is whether the mapping  $\alpha_*$  somehow lifts to define a commutative square

$$(4.2.16) \quad \begin{array}{ccc} \mathrm{Adm}(G_F) & \xrightarrow{\tilde{\alpha}_*} & \mathrm{Adm}(H_F) \\ \downarrow & & \downarrow \\ \Phi(G_F) & \xrightarrow{\alpha_*} & \Phi(H_F) \end{array}$$

where the vertical maps are as in (4.2.15). Langlands' *Principle of Functoriality* posits the existence of such squares [Bor15]. In particular, it posits the



existence of maps  $\tilde{\alpha}_*$  between irreducible admissible representations of  $G_F$  and  $H_F$ . We may look for such maps  $\tilde{\alpha}_*$  whether or not we know that the vertical maps exists.

Many maps of the form  $\tilde{\alpha}_*$  are known to exist. The maps of parabolic or cohomological induction (cf. §§3.6.1, 3.6.4) can be regarded as such maps (corresponding to maps  $\alpha$  given by embeddings of Levi components of parabolic subgroups). Similar remarks apply to the correspondences defined by homomorphisms of Hecke algebras ( $p$ -adic “Harish-Chandra homomorphisms”) constructed in [HoMy1, 2, BuKu, Wald3]. Viewing such correspondences as being of the type (4.2.16) in no way simplifies establishing that they exist, but it makes them seem plausible and suggests they are part of a larger pattern.

Other examples are the “base change” mappings between a group over a given field  $F$  and the same group but with coefficients in an extension field  $F'$  of  $F$ . The most frequently studied case is when  $F'$  is a cyclic extension of prime degree over  $F$  [ArCl, Lgld5, GeLa]. This correspondence has been studied globally (i.e., for automorphic forms) as well as locally (for admissible representations) and has resulted, among other things, in the establishment of some new cases of the Artin conjecture [Lgld5, Tunn2]. Establishment and use of correspondences suggested by the Principle of Functoriality is becoming standard operating procedure in the theory of automorphic forms [Lgld5, BiRa, Roga1]. Here also the trace formula is heavily used.

The theory of  $\theta$ -series, which is the automorphic aspect of the oscillator representation [Howe5] also produces, in a god-given way, correspondences between automorphic forms, made up of local correspondences between admissible representations of certain pairs of groups  $(G, G')$  (precisely “reductive dual pairs”—mutually centralizing subgroups of the symplectic group  $\mathrm{Sp}_{2n}$ ) (cf. [Howe5, Gelb2, MoVW, Moeg, Prze, Mand, Wald, Shim2, Shin2, Niwa], etc.). In some cases, these “ $\theta$ -correspondences” can be shown to be of the form predicted by the Principle of Functoriality. However, for groups over  $\mathbf{R}$  (i.e., Lie groups), where the Langlands classification (4.2.15) is known, and which therefore currently offer the strongest test, some  $\theta$ -correspondences are incompatible with the  $L$ -packet parametrization. That is, some  $\theta$ -correspondences do not preserve  $L$ -packets [Adam1].

On the other hand, Arthur [Arth2, 3], influenced by earlier examples of  $\theta$ -series which violated the “generalized Ramanujan conjecture” [Sata, HoPS, Kuro], and by problems stemming from  $L$ -indistinguishability, was led to propose that certain parts of  $\mathrm{Adm}(G_F)$  should be coagulated into lumps larger than  $L$ -packets. These larger lumps are called  $\psi$ -packets. It seems possible that  $\theta$ -correspondences will be compatible with  $\psi$ -packets [Adam1]. Further, it appears that Arthur’s proposals are directly relevant to the determination of the unitary dual for semisimple Lie groups [Arth2, 3, BaVo] (and presumably  $p$ -adic groups also). This complex interplay of rich examples, difficult technical issues, and bold ideas makes the theory of automorphic

forms an exciting research area, one whose challenges will occupy generations to come.

**4.3. Physics, geometry, and differential equations.** Physics, geometry, and differential equations are intimately related to each other, and the interaction of Lie theory with all of them has been extensive. This is hardly surprising, since all three of geometry, differential equations, and Lie theory are the children of physics, and group theory is now understood to provide a large part of the underpinnings of both geometry and analysis. On a philosophical level, one may also observe that a great deal of physics is concerned with conservation laws and invariance principles, and group theory is a natural language for expressing such ideas. The application of Lie theory to physics is the subject of a long series of large conferences [ITGT1–17, Loeb] and numerous texts [BiLo1, 2, BeTu, Corn, Hame, Jone, Lich, Wolb], etc. We have given perhaps the most basic example in quantum mechanics, the harmonic oscillator, in §3.1. The energy states of the hydrogen atom, i.e., the quantum Coulomb problem, provide another elegant and subtle example with “accidental degeneracies” accounted for by extra symmetries [Abar, Engl, Fron, Shan]. Shortly after Einstein introduced special relativity [Eins], Minkowski [Mink] explained that the difference between classical (Newtonian or Gallilean) physics and relativity could be explained as a change in the symmetry group of space time, from the isometry group of the degenerate form  $x_1^2 + x_2^2 + x_3^2$  in four variables to that of the nondegenerate form  $x_1^2 + x_2^2 + x_3^2 - x_4^2$ . This is analogous to the passage from Euclidean to non-Euclidean (Lobachevskian) geometry. Several early cosmological models of the expanding universe are based on homogeneous spaces for various semisimple Lie groups [Eins2, Sitt]. A speculative alternative to “big bang” models of the universe has been proposed by I. Segal [Sega4], based on  $S^3 \times \mathbf{R}$  as a homogeneous space for  $SU(2, 2)^\sim$ , where  $^\sim$  indicates the universal cover. R. Penrose’s formulation of general relativity in terms of “twistors” [PeWa, Penr, Well] is based on Lie-theoretic constructions. Modern bookkeeping schemes for elementary particles, beginning with isotopic spin, and continuing with the 8-fold way [Gü Ra, Loeb, SaWe], and beyond, are based on the combinatorics of finite-dimensional representations of Lie groups ( $SU_2$ ,  $SU_3$ , etc.). The Yang-Mills equations, thought to be the governing equations of “quantum chromodynamics,” the interactions of quarks, are variational equations on the space of “gauge fields,” connections of a principal fiber bundle, whose fiber is an appropriate Lie group [AtBo, FrUh, Taub]. Recently the speculative “string theory” and its cousin conformal field theory [BePZ, Gawe, MoSe, Segl, Witt], have contributed to the explosive growth in the study of infinite-dimensional Kac-Moody Lie algebras, especially the “affine” Lie algebras (cf. [Dola, FrKa, FrLM, GrSW, Kac1], etc.). Currently the study of “quantum Lie algebras” or “quantum groups,” which grew out of certain formal identities arising in exactly solvable statistical mechanical models [Andr, AnBF, Baxt, JiMi1, 2], is proceeding furiously, and

is leading to new insights even in matters of longstanding interest in finite-dimensional representation theory of semisimple Lie groups [**Kash**, **Lusz4**, **Cher1**, **2**, **Murp**].

Some of these developments have led to dramatic new advances in geometry, for example, Donaldson's analysis of 4-manifolds as boundaries of solutions of families of solutions to the Yang-Mills equations [**FrUh**]. Donaldson constructed new invariants of 4-manifolds; there have also been many recent constructions of invariants for 3-manifolds, including [**Cran1**, **2**, **TuVi**], which construct 3-manifold invariants in terms of the quantum  $6-j$  coefficients,  $q$ -analogs of numbers which arise in the explicit description of tensor products of representations of  $SL_2$  [**BiLo1**, **2**].

The more traditional applications of Lie theory to geometry are also extensive, and many are fundamental. One of the first that can be considered explicitly Lie-theoretic (although it precedes the period of coverage of this essay) is the Erlanger Programm of Felix Klein [**Klei2**], in which the relation between Euclidean geometry and the various alternatives (the hyperbolic geometry of Lobachevsky and Bolyai, the elliptic geometry of Riemann, the projective geometry, etc.) which had arisen in the nineteenth century, are systematically related to one another in terms of their associated symmetry groups; and in which, furthermore, the word "geometry" is proposed to mean the understanding of the invariants of a group  $G$  acting on several copies of a homogeneous space. (Thus for the Euclidean group of isometries of the plane, the only invariant of two points is the distance between them; the invariants of three points are side-angle-side, or angle-side-angle, etc.) The variety of possible geometries was the question which led to Killing's [**Kill**] (see also [**Cole**]) classification of the simple Lie algebras, and which received a more or less definitive formulation in E. Cartan's definition and classification of symmetric spaces [**Crtn4**], (see also [**Helg2**, **Loos**]). Similarly, invariant theory, which was more or less explicitly Lie-theoretic in nature even before Lie theory existed, and which led to tremendous computational efforts in the nineteenth century [**Sylv1**, **2**], after being cut off at the root by Hilbert [**Hilb1**, **2**], found new life by being grafted onto the representation theory of semisimple groups [**Litt**, **Weyl2**], and made a basis for the theory of moduli of algebraic varieties [**FoMu**].

These evocations and lists could go on indefinitely. Let us mention just a very few particular examples. Efforts to understand the topology, especially the cohomology ring, of Lie groups [**Crtn5**, **6**], inspired deRham's Theorem [**Rham**], and led to the notions of  $H$ -space [**Hopf3**, **Brow**, **Thms**] and Hopf algebra [**Hopf2**, **HoSa**, **Swee**]. This latter notion has been crucial in the recent formulation of the idea of quantum group [**Drin**, **Jimb**]. In Riemannian geometry, a major theme has been the study of spaces of positive (sectional) curvature. Most of the known examples are homogeneous spaces or perturbations of them [**Barg2**, **Wall6**]. In even dimensions, there can be only a finite number of homotopy types of positively-pinched manifolds [**ChEb**], and there

are only a finite number of homogeneous examples in each even dimension. However, Wallach [Wall6, AlWa] classified all possible homogeneous, positively curved manifolds, and found an infinite family of seven-dimensional examples. Another elegant example in a similar spirit is an exotic 7-sphere with nonnegative curvature, constructed as a quotient of  $\mathrm{Sp}_2$  by an action of  $\mathrm{Sp}_1 \simeq \mathrm{SU}_2$  built from right and left translations [GrMe]. Finally, we mention the question [KacM] whether the spectrum of the Laplace-Beltrami operator on functions on a compact Riemannian manifold determines the manifold up to isometry. The answer is “no,” a stronger and stronger no as more and more counterexamples have been constructed [Miln, Iked, Vign, Suna, Bera, BrGo]. All these examples are homogeneous spaces, or almost homogeneous spaces, and rely on the structure imposed by group theory to get control of the spectrum of the Laplacian.

Differential equations, especially nonlinear differential equations, were the context for some of Lie’s original investigations [Hawk, LiEn], and certain infinite-dimensional Lie algebras were studied by Cartan [Crtn5, GuSt2] in connection with variational problems. (Recently, Cartan’s methods were used by Olver [Olve2] to establish an interesting result in the invariant theory of binary forms and Cartan’s algebras, which have finite-dimensional quotients modulo  $p$ , figure in the classification of simple Lie algebras over fields of positive characteristic [StWi].) For much of the twentieth century, however, such investigations were pursued much less vigorously than ones sparked by the internal development of the subject, but recently, Lie groups as symmetry groups of solutions to differential equations have received renewed attention [Ibra, Olve1, Ovs1].

A relatively recent area, control theory borrows techniques from geometry and differential equations both, and in particular uses Lie theory [Haze1, 2, Broc1, 2, Crou, HaMa, Herm1, 2]. Being concerned primarily with the future, control theory has stimulated the theory of Lie semigroups [HiHL].

Linear partial differential equations have been attacked by methods of harmonic analysis, whose essentially group-theoretic basis was recognized in this century [PeWe, Weil3, Weyl1]. The development of pseudo-differential and Fourier integral operators [Hörm] has led to “phase-space analysis” in which the distinction between multiplication and differentiation is blurred. This phase space analysis has a direct interpretation in terms of the Heisenberg group, and many of the basic estimates [Beal, Hörm, Unte] of the theory can be carried out efficiently using group theory, of the Heisenberg group itself, and of the oscillator representation [Foll, Howe3]. The recent theory of wavelets [Daub, Gros, Meye] to some extent breaks with Lie theory, but still draws inspiration from symmetry principles, especially scaling. Nilpotent Lie groups, especially the Heisenberg group, have been used to establish facts about particular systems of equations, especially hypoellipticity [Foll2, FoSt, RoSt]. The Lewy counterexample [Lewy] to local solvability of systems of differential equations has been interpreted as a system of left-invariant op-

erators defining a positive complex polarization (cf. §3.3) on the Heisenberg group [GrSt]. Conversely, there has been substantial study of the analytic properties of elements of  $\mathcal{U}(g)$  acting on  $C^\infty(G)$ , especially for  $G$  nilpotent (cf. [Corw3, CoHr, HeNo, Rock], etc.).

Further, many of the classical equations of physics have many symmetries, and are otherwise tightly connected to group theory (cf. [Howe6, ITGT, Hame, Jone], etc.). Dirac is said to have been guided to his equation for the dynamics of the electron by a desire to have it be Lorentz invariant. The Laplace operator

$$\Delta = \sum_{j=1}^n \frac{\partial^2}{\partial x_j^2}$$

is invariant under the full group of isometries (translations and rotations) of  $\mathbf{R}^n$ , and is characterized as the generator of the algebra of all differential operators with full Euclidean invariance. In particular, the space of harmonic functions ( $\Delta f = 0$ ) or indeed any eigenspace of  $\Delta$ , carries a representation of the Euclidean group. The matrix coefficients of these representations are Bessel functions [Mill1, Vile]. This example may be generalized tremendously. Helgason [Helg1] proposes a general program of studying a system consisting of a group  $G$  acting on a space  $X$ , together with some differential operators  $\{D_i\}$  which commute with  $G$  and with each other: the goal is to analyze the representations of  $G$  defined by the joint eigenspaces of the  $D_i$ . Such a study becomes a two way street: one can learn about the eigenspaces if one has control over the representations, or vice versa. Many of the geometric realization theorems, for discrete series and other interesting representations (cf. [JaVe, Mant, RaSW, PaRo, Tora, Schm1–3], etc.), could be considered examples in an extended version of this program. The most direct examples arise when  $G$  is the full group of isometries of a symmetric space  $X$ : in that case the full algebra of  $G$ -invariant differential operators on  $X$  is commutative, and the eigenspaces yield representations infinitesimally equivalent to the spherical principal series (cf. §3.6.1) [Helg1, GaVa]. In the past decade, there has been a large amount of work generalizing this to the case of semisimple symmetric spaces (cf. [Ban, Bien, FIJe1, 2, OlOr, Oshi1–3, Schl], etc.). Matrix coefficients associated to eigenspace representations will yield many classical families of special functions [GaVa, Helg1, Mill1, Vile].

The subject of “dynamical systems” is the modern home for the qualitative theory of differential equations. It combines geometrical, analytical, and dynamical considerations to study a variety of problems, many of which were originally motivated by physics (though in some cases investigations have strayed rather far from their origins). A (discrete time) *dynamical system* in the modern sense is simply a pair  $(X, T)$ , where  $X$  is a set (“space,” or “state-space,” or “phase space”) and  $T: X \rightarrow X$  is a mapping. We will assume  $T$  is invertible, i.e., 1-to-1 and onto, though this is not necessary for

all purposes. Usually  $X$  will be equipped with structure, e.g., as a measure space, a topological space, a smooth manifold, a metric space, and  $T$  preserves that structure. Thus one speaks of measurable dynamics, topological dynamics, smooth dynamics, rigid motions, or other specific kinds of dynamical system, in order to specify context. In all cases, the focus is on motion, which is described by  $T$ . Thus one is implicitly studying the cyclic group with generator  $T$ . If one wishes to think of continuous motion, one considers a flow or one-parameter group (cf. §1.2.1) of transformations  $T_r$ ,  $r \in \mathbf{R}$ , rather than a single  $T$ . There is a standard construction (suspension, or “flow under a function,” cf. [Pete]) to convert a discrete time dynamical system to a flow.

Even at this early stage, before any special structures have been imposed, it is clear that group theory provides a wealth of examples of dynamical systems: any triple  $(G, H, g)$ , where  $G$  is a group,  $H \subseteq G$  is a subgroup, and  $g$  is an element of  $G$ , defines a dynamical system  $(g, G/H)$ , with  $g$  acting on  $G/H$  by left translation. If  $G$  is a Lie group, we can make flows by considering one-parameter subgroups rather than single elements. We will see below that some of these examples are extremely interesting.

Many questions asked under the rubric of dynamical systems concern the long-term behavior of these systems. *Ergodic theory* studies the *average* long-term behavior of systems  $(X, T)$ , where  $X$  is equipped with a probability measure  $\mu$ , which  $T$  preserves. Ergodic theory was motivated originally by a question in statistical mechanics. To explain this, we recall some basic notions and facts of the subject. Suppose we can write  $X = X_1 \cup X_2$ , a disjoint union, with both  $X_1$  and  $X_2$  of positive measure, and both invariant under  $T$ . Then writing  $T_j = T|_{X_j}$ , we may consider that  $(T, X)$  is composed of two separate, at least in the measure-theoretic sense, subdynamical systems  $(X_j, T_j)$ ,  $j = 1, 2$ . Interest obviously focuses on systems which are not decomposable in this fashion: these indecomposable systems are called *ergodic*, and a main problem of ergodic theory has been: given a system  $(X, T)$  decide whether it is ergodic.

A basic result which brings out the significance of ergodicity is Birkhoff's Ergodic Theorem [Halm, Pete]. Let  $(X, T)$  be a measurable dynamical system, and denote by  $\mu$  the given  $T$ -invariant measure on  $X$ . Given  $p \in X$ , and a (measurable) function  $f$  on  $X$ , define the *time averages* along the orbit of  $p$  by

$$(4.3.1) \quad A_n(f)(p) = \frac{1}{n} \sum_{j=0}^{n-1} f(T^j(p)), \quad n \geq 1.$$

**THEOREM 4.3.2.** *For  $f \in L^1(X, \mu)$ , there is a function  $A_\infty(f)$  in  $L^1(X, \mu)$ , invariant by  $T$ , such that*

$$\lim_{n \rightarrow \infty} A_n(f)(p) = A_\infty(f)(p)$$

for almost all  $p$  (with respect to  $\mu$ ). In particular, if  $T$  acts ergodically on  $X$ , then

$$(4.3.3) \quad A_n(f)(p) \rightarrow \int_X f(x) d\mu(x)$$

for almost all  $p$  in  $X$ .

An easy argument convinces one that  $T$  is ergodic if and only if the only  $T$ -invariant functions in  $L^1(X, \mu)$  are the constant functions. Thus the second statement of Theorem 4.3.2 is an easy corollary of the first. It is the source of the interest of ergodicity in statistical mechanics. The time averages  $A_n(f)(p)$  of various dynamical quantities  $f$  appear in statistical mechanical calculations; but the calculation of  $A_n(f)(p)$  in terms of its definition, involving as it does detailed step-by-step knowledge of the dynamics over long periods of time, is formidably difficult. If  $A_n(f)(p)$  can be estimated by the single, predeterminable number  $\int_X f(x) d\mu(x)$ , known as the *phase average*, the life of the statistical mechanics theorist is immeasurably simplified. For further discussion of these matters, see [Mack1].

From the Kronecker line [ArAv, p. 132; Casl] (a dense one-parameter subgroup of the  $n$ -dimensional torus  $T^n$ ), through the extensive study, especially by Hedlund [Hed11, 2], and Hopf [Hopf2], in the 1920s and 1930s, of the geodesic flow on Riemann surfaces (endowed with the metric of constant curvature) to the current studies by Dani-Margulis [Dani1, 2, Marg, DaMa1, 2] and, more or less definitively, by Ratner [Ratn4], of horocycle flows on  $G/\Gamma$ ,  $\Gamma$  a lattice in semisimple  $G$ , Lie theory has provided important examples of ergodic actions. These examples frequently bear on number theory, especially questions of Diophantine approximation. The relation of the Kronecker line to Diophantine approximation is classical. We will explain shortly a result of Margulis (generalized dramatically by Ratner [Ratn4]) which implies a number-theoretic result which had been an open conjecture for sixty years.

In [FoGe], Fomin and Gelfand showed how to use representation theory to establish ergodicity of dynamical systems constructed from homogeneous spaces, or more generally, obtained from selecting one transformation from a larger group of symmetries. This is based on the fact that ergodicity and related properties can be detected using spectral theory. Precisely, consider a measurable dynamical system  $(X, T)$ , with invariant measure  $\mu$ . Let  $L^2(X, \mu)$  be the  $L^2$ -space of  $\mu$ , and let  $L^2(X, \mu)^o$  be the subspace orthogonal to the constant functions. We can use  $T$  to define an endomorphism  $U_T$  of  $L^2(X, \mu)$ , in the usual way:

$$(4.3.4) \quad U_T(f)(x) = f(T^{-1}(x)), \quad f \in L^2(X, \mu), x \in X.$$

Ergodicity of  $T$  can easily be formulated in terms  $U_T$  [ArAv, Halm]: for  $T$  to be ergodic,  $U_T$  should have only the constant functions as fixed vectors, or should have no fixed vectors in  $L^2(X, \mu)^o$ , or should not have 1 in its point spectrum on  $L^2(X, \mu)^o$ . Since fixed vectors can be detected by means

of matrix coefficients ( $v$  is a fixed vector for  $U_T$  if and only if the matrix coefficients (cf. §A.1.11)

$$\varphi_{v,v}(n) = \int_X U_T^n(v)(x) \overline{v(x)} dx, \quad n \in \mathbf{Z},$$

are constant as a function of  $n$ ), we can also formulate the condition of ergodicity in terms of matrix coefficients.

The matrix coefficient formulation of ergodicity suggests several variant conditions stronger than ergodicity. One of the most useful of these is *strong mixing* [ArAv, Halm, Pete] which is the requirement that the matrix coefficients  $\varphi_{v,v}(n)$  should decay to zero as  $n \rightarrow \infty$  if  $v \in L^2(X, \mu)^0$ . Strong mixing is a technically pleasant property because it makes sense for any group  $G$  and any unitary representation  $\rho$  of  $G$  on a Hilbert space  $\mathcal{H}$ : for  $u, v$  in  $\mathcal{H}$ , we should require that the matrix coefficient  $(\rho(g)u, v)$  decay to zero as  $g$  goes to  $\infty$  in  $G$ . In other words, the set of  $g$  for which  $|(\rho(g)u, v)| \geq \varepsilon$  should be compact. A further useful property of strong mixing is that it is clearly inherited by closed subgroups. By interpreting the geodesic flow on a Riemann surface in terms of  $\mathrm{SL}_2(\mathbf{R})$ , and then essentially showing that any nontrivial irreducible unitary representation of  $\mathrm{SL}_2(\mathbf{R})$  has the strong mixing property (actually they formulated their result differently), Gelfand and Fomin gave a new proof of the results of Hedlund and Hopf. The Gelfand-Fomin argument went through several stages of generalization until today it can be used to show that nearly any measurable dynamical system coming from a homogeneous space will be ergodic, unless it fails to be for obvious reasons [HoMo, Moor2, Zimm1].

The influence between Lie theory and ergodic theory has been mutual. A particularly striking example of this was Margulis's use of ergodic theory in the proof of his Superrigidity Theorem [Zimm1], which was then reinterpreted as being a result in ergodic theory by Zimmer. A very recent example of this mutual interaction is the Margulis proof of the Oppenheim Conjecture [Marg1], followed quickly by Ratner's [Ratn4] broad generalization of the key ergodic-theoretic result underlying his proof.

As techniques for establishing ergodicity of dynamical systems became established, workers in ergodic theory refined their investigations. One kind of finer question frequently raised is loosely referred to as "unique ergodicity." It is a kind of inverse to the standard ergodic problem. Suppose  $(X, T)$  is a dynamical system and  $X$  is a locally compact Hausdorff space. Then  $X$  will support many probability measures, and one can ask for a description of all  $T$ -invariant probability measures on  $X$ . These will form a closed convex set in the unit ball of the space of all measures, and the ergodic ones are essentially the extreme points. Hence ergodicity figures in this question too. There is also clearly a connection between knowing all possible ergodic measures and knowing the closures of orbits.

Ratner [Ratn1-4] has more or less definitively answered these questions



for the action of unipotent groups on homogeneous spaces. For purposes of explaining this, we will call an element  $g$  of a Lie group  $G$  *unipotent in  $G$*  if all eigenvalues of  $\text{Ad } g$  acting on the Lie algebra of  $G$  are equal to 1 (sometimes this is called *Ad-unipotence*). We will call a subgroup  $H \subseteq G$  *unipotently generated* if the subset of elements of  $H$  which are unipotent in  $G$  generate  $H$  as group.

**THEOREM 4.3.5.** (a) *Let  $H$  be a unipotently generated subgroup of the Lie group  $G$ , and let  $\Gamma \subseteq G$  be a lattice. For  $x \in \Gamma \backslash G$ , consider the closure  $\text{Cl}(xH)$  of the  $H$  orbit of  $x$  in  $\Gamma \backslash G$ . Then there is a subgroup  $\tilde{H}_x \subseteq G$ , containing  $H$ , such that*

- (i)  $\text{Cl}(xH) = x\tilde{H}_x$ ,
- (ii)  $x^{-1}\Gamma x$  is a lattice in  $\tilde{H}_x$ .

(b) *Let  $\mu$  be an  $H$ -invariant probability measure on  $\Gamma \backslash G$ . Then there is a subgroup  $\tilde{H}_\mu \subseteq G$ , containing  $H$ , and a point  $x \in \Gamma \backslash G$  such that*

- (i)  $x^{-1}\Gamma x \cap \tilde{H}_\mu$  is a lattice in  $\tilde{H}_\mu$ ,
- (ii)  $\mu$  is the invariant probability measure on  $((x^{-1}\Gamma x) \cap \tilde{H}_\mu) \backslash \tilde{H}_\mu \simeq x\tilde{H}_\mu$ .

In particular, these results hold if  $H$  is a one-parameter subgroup of unipotent elements. It follows from this result and the compactness criterion [BoHC, MoTa] for arithmetic lattices, that if  $H$  is a one-parameter unipotent subgroup of  $G = \text{SL}_2(\mathbf{R})$ , and if  $\Gamma$  is an arithmetic lattice with  $\Gamma \backslash G$  compact, then

- (i) every  $H$ -orbit is dense,
- (ii) the only  $H$ -invariant measure is the standard measure on  $\Gamma \backslash G$ .

Such extraordinary rigidity of behavior is in striking contrast to the situation when  $H$  consists of noncompact semisimple elements, e.g., elements from the split Cartan subgroup  $A$  of the Iwasawa decomposition (cf. §A.2.3.3). For such  $H$ , Dani [Dani3] has shown the resulting transformations on  $\Gamma \backslash G$  are isomorphic to Bernoulli shifts—and therefore determined up to isomorphism by their entropy. The rigidity of unipotent flows comes from the long term coherence of their trajectories, which diverge from one another slowly. In current parlance, unipotent flows are not chaotic [Deva]; whereas flows defined by semisimple elements have rapidly diverging trajectories and are chaotic (many of them are Anosov flows [AbMa, ArAv]). The coherence of trajectories, which Ratner terms “Property  $H$ ,” is illustrated by the following calculation comparing the “ $LU$ ” with the “ $UL$ ” decompositions (cf. §1.1) for  $\text{SL}_2(\mathbf{R})$ :

$$(4.3.6a) \quad \begin{bmatrix} a & 0 \\ z & a^{-1} \end{bmatrix} \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix} \begin{bmatrix} b & 0 \\ 0 & b^{-1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ w & 1 \end{bmatrix},$$

where

$$(4.3.6b) \quad s = \frac{a^2 t}{1 + azt}, \quad b = \frac{a}{azt + 1}, \quad w = \frac{az}{azt + 1}.$$

We interpret these formulas as follows. We imagine we flow by right multiplications of the group  $N^+ = \{ \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \}$  ( $t$  connoting time). From the  $N^+$ -trajectory through some point  $x_0$ , we move to a point  $x_1$  by the small motion  $\begin{bmatrix} a & 0 \\ z & a^{-1} \end{bmatrix}$  in a direction transverse to the flow lines, and then flow along the  $N^+$ -trajectory through  $x_1$ . This is the procedure indicated by the left side of equation (4.3.6a). The right-hand side tells us how to track the flow on the nearby  $N^+$  trajectory by moving along the original trajectory (through  $x_0$ ), at not quite a steady pace, specified by  $s$ , then moving transversely to the other trajectory, first by  $\begin{bmatrix} b & 0 \\ 0 & b^{-1} \end{bmatrix}$  (which preserves the set of all  $N^+$ -trajectories), then by  $\begin{bmatrix} 1 & 0 \\ w & 1 \end{bmatrix}$ . The point to note is that, as long as we do not flow too far (precisely, as long as  $1 + azt$  remains near 1, or at least, away from zero) we do not have to move very far to get from our original trajectory to the adjacent one. Further, the nearer the perturbed trajectory, the longer we can track it easily: if we divide  $z$  by 2, then the time until  $b - 1$  or  $w$  exceed some prespecified limit at least doubles. If we were dealing with the rapidly diverging trajectories defined by semisimple elements, halving the original perturbation would only delay the time to divergence by a fixed increment, rather than doubling it.

Now let us describe a consequence of Theorem 4.3.5, the proof of a conjecture of Oppenheim [Opp], which was proved by Margulis [Marg] (see also [DaMa2]) on the basis of the relevant special case of Theorem 4.3.5. Consider a quadratic form

$$(4.3.7) \quad Q(x) = \sum_{j,k=1}^n b_{jk} x_j x_k$$

on  $\mathbf{R}^n$ . If  $Q$  is rational in the sense that its coefficients  $b_{jk}$  are rational numbers, then the values of  $Q$  on  $\mathbf{Z}^n$  live in  $\frac{1}{r}\mathbf{Z}$  for an appropriate denominator  $r$ ; in particular the values form a discrete set. This remains true if  $Q$  is projectively rational in the sense that the ratios  $b_{jk}/b_{lm}$  are in  $\mathbf{Q}$ . It is also true if  $Q$  is definite. The conjecture of Oppenheim said essentially that these were the only conditions under which  $Q$  will have discrete values.

**THEOREM 4.3.8.** *Suppose  $n \geq 3$ , and that the form  $Q$  of formula (4.3.7) is indefinite and not projectively rational. Then the set of values  $\{Q(z) : z \in \mathbf{Z}^n\}$  is dense in  $\mathbf{R}$ .*

To prove this using Theorem 4.3.5, observe first that it suffices to treat the case  $n = 3$ . Let  $G = \mathrm{SL}_3(\mathbf{R})$ , and let  $H = \mathrm{SO}(Q)$  be the determinant one isometry group of  $Q$ . Consider the transforms  $h(\mathbf{Z}^3)$  of the lattice  $\mathbf{Z}^3 \subseteq \mathbf{R}^3$  by elements of  $H$ . This will be a collection of lattices inside the set of all lattices  $L \subseteq \mathbf{R}^3$  such that the volume of  $\mathbf{R}^n/L$  (measured with the push-down of Lebesgue measure) is 1, which set is identifiable with  $\mathrm{SL}_3(\mathbf{Z}) \backslash \mathrm{SL}_3(\mathbf{R})$ . If the set  $\{h(\mathbf{Z}^3), h \in H\}$  is dense in the space of lattices,

then in particular the set of points  $\{h(z) : h \in H, z \in \mathbf{Z}^3\}$  is dense in  $\mathbf{R}^3$ . On the other hand, since  $H$  preserves  $Q$ , the values of  $Q$  on  $h(\mathbf{Z}^3)$  are the same as the values on  $\mathbf{Z}^3$ . Hence  $Q(\mathbf{Z}^3)$  must be dense in  $\mathbf{R}$ .

So suppose  $\{h(\mathbf{Z}^3)\}$  is not dense in the space of lattices. Since the form  $Q$  is indefinite, the group  $\mathrm{SO}(Q)$ , or at least its identity component, which has index 2 in the full group, is generated by unipotent elements. Hence Theorem 4.3.5 gives us a group  $\tilde{H}$ ,  $H \subseteq \tilde{H} \subseteq G$ , such that the closure of  $H(\mathbf{Z}^3)$  is  $\tilde{H}(\mathbf{Z}^3)$ . But there are no groups between  $H$  and  $G$  (check the adjoint action of  $H$  on  $\mathrm{Lie} G / \mathrm{Lie} H$ : it is irreducible). Hence either  $\tilde{H} = G$  or  $\tilde{H} = H$ . The first possibility amounts to the denseness of  $H(\mathbf{Z}^3)$ , which has already been rejected. Hence  $H(\mathbf{Z}^3)$  is closed, and  $H \cap \mathrm{SL}_3(\mathbf{Z})$  is a lattice in  $H$ . The Borel Density Theorem [Zimm1] then implies that  $H \cap \mathrm{SL}_3(\mathbf{Z})$  is Zariski-dense in  $H$ , and from this Theorem 4.3.8 follows directly; for if we conjugate everything by an automorphism of  $\mathbf{C}$  over  $\mathbf{Q}$ , the group  $H \cap \mathrm{SL}_3(\mathbf{Z})$  will remain fixed, since it consists of matrices with rational (indeed, integer) entries. Hence its Zariski closure  $H$  remains fixed. Hence  $Q$ , which is determined up to multiples by  $H$ , remains fixed up to multiples. Hence the ratios  $b_{jk}/b_{lm}$  remain fixed, hence are in  $\mathbf{Q}$ .

Our second topic in dynamical systems will be essentially completely opposite from ergodic theory—completely integrable systems. Stimulated by discoveries of several examples of such systems, this topic was extremely active in the decade around 1980. Probably greatest interest attached to several infinite-dimensional systems associated to the “inverse scattering problem” [BeCo, BeDT, TrPo] for second-order ordinary differential operators, in particular the Korteweg-DeVries equation [Adle2, GGKM, BeDT]. However, some finite-dimensional systems were found which had strong analogies with the infinite-dimensional ones, and these, especially the Toda lattice [GoWa, Kost6, Syme], also received considerable attention. Both the infinite-dimensional and the finite-dimensional systems were found to have close connections with Lie groups. We will discuss the (nonperiodic) Toda lattice because it is quite accessible and because it highlights the geometry of Lie groups, especially the relations between the various standard decompositions (Cartan, Iwasawa, Bruhat, cf. §A.2.3). Our account is largely based on [GoWa] and [DLNT].

We should recall some features of Hamiltonian dynamics. We refer to §3.2 or [AbMa, GuSt1] for background about symplectic manifolds. Let  $M$  be a symplectic manifold, let  $q$  be a function on  $M$ , and let  $\alpha^{-1}(dq)$  be the Hamiltonian vector field associated to  $f$  (cf. formula (3.2.1.6)). Let  $q_2$  be another function on  $M$ . From our discussion in §3.2, we know that the conditions

$$(4.3.9a) \quad \alpha^{-1}(dq)(q_2) = 0$$

and

$$(4.3.9b) \quad [\alpha^{-1}(dq), \alpha^{-1}(dq_2)] = 0$$

are equivalent. However, the geometric interpretations of these two equations are quite different. Equation (4.3.9a) means that  $q_2$  will be invariant under the flow generated by  $\alpha^{-1}(dq)$ . In other words, the flow defined by  $\alpha^{-1}(dq)$  takes place in the level sets of  $q_2$ . Thus an equation of type (4.3.9a) helps us locate the trajectories of the flow of  $\alpha^{-1}(dq)$ . To suggest these roles,  $q_2$  is sometimes called a *conserved quantity*, or an *integral* for the flow of  $\alpha^{-1}(dq)$ . On the other hand, equation (4.3.9b) means that the flows associated to  $\alpha^{-1}(dq)$  and  $\alpha^{-1}(dq_2)$  will commute with each other, hence each permutes the trajectories of the other, or each defines a one-parameter group of symmetries of the other. Thus, if we have two functions satisfying equations (4.3.9), both their flows are contained in their simultaneous level sets, and preserve each other. The fact that for Hamiltonian flows a conserved quantity plays a dual role of symmetry, and vice versa, is a particularly enriching feature of Hamiltonian mechanics.

Suppose that, given  $q = q_1$  as above, we can find several functions  $q_2, q_3, \dots, q_m$  such that any pair  $q_j, q_k, 1 \leq j, k \leq m$ , satisfies the mutually equivalent conditions (4.3.9). Then all the flows generated by the  $\alpha^{-1}(dq_j)$ , which (under the obvious necessary condition of functional independence, viz., that the vector fields (equivalently the differentials  $dq_j$ ) should be linearly independent) will fill out submanifolds of dimension  $m$ , will all simultaneously preserve the level sets of all the  $q_j$ , which (under the same assumption of functional independence) will have codimension  $m$ . Thus we must have  $m \leq \dim M - m$ , or  $\dim M \geq 2m$ . In the extreme case, when  $\dim M = 2m$ , the systems of trajectories of the  $\alpha^{-1}(dq_j)$  will completely fill, at least locally, the joint level surfaces of the  $q_j$ , so that the whole situation is determined: the possible motions generated by the  $\alpha^{-1}(dq_j)$  completely fill up the surfaces defined by constancy of the  $q_j$ , and vice versa. Such a system is called *completely integrable*.

In two dimensions, i.e., if  $\dim M = 2$ , all Hamiltonian systems are completely integrable, but in dimensions greater than two, complete integrability is very special, and the discovery of completely integrable systems is an interesting event. In the 1960s, the physicist Toda [Toda1] defined a dynamical system interpretable as a system of  $n$  points moving on the real line subject to attractive forces depending exponentially on the distance between the particles. Precisely, it is the Hamiltonian system in  $\mathbf{R}^{2m}$ , with its standard symplectic form, defined by the function (in symplectic coordinates  $x_j, y_j, 1 \leq j \leq n$ )

$$(4.3.10) \quad \frac{1}{2} \sum_{j=1}^n y_j^2 + \sum_{j=1}^{n-1} \exp(x_j - x_{j+1}).$$

It is easy to check that the subspace of  $\mathbf{R}^{2n}$  defined by

$$\sum_{j=1}^n y_j = 0 = \sum_{j=1}^n x_j$$

is invariant for the flow defined by the Hamiltonian (4.3.10). If on this subspace we make the change of variables

(4.3.11)

$$a_j = -\frac{y_j}{4}, \quad b_k = \frac{1}{2} \exp \frac{(x_k - x_{k+1})}{2}, \quad 1 \leq j \leq n, \quad 1 \leq k \leq n-1,$$

then we find the equations

$$(4.3.12) \quad \frac{da_j}{dt} = (b_{j-1}^2 - b_j^2), \quad \frac{db_k}{dt} = 2b_k(a_k - a_{k+1}).$$

We may interpret the equations (4.3.12) in terms of matrices in two different ways.

Flaschka [Flas] and Moser [Mose] interpreted equations (4.3.12) as follows. Let  $M^+$  be the tridiagonal symmetric matrix

$$(4.3.13) \quad M^+ = \frac{1}{2} \begin{bmatrix} 2a_1 & b_1 & 0 & \cdots & 0 \\ b_1 & 2a_2 & b_2 & & 0 \\ 0 & & \ddots & & 0 \\ \vdots & & & & b_{n-1} \\ 0 & \cdots & & b_{n-1} & 2a_n \end{bmatrix}$$

and let  $M^-$  be the skew-symmetric matrix

$$(4.3.14) \quad M^- = \frac{1}{2} \begin{bmatrix} 0 & -b_1 & 0 & \cdots & 0 \\ b_1 & 0 & -b_2 & & \vdots \\ & b_2 & 0 & & 0 \\ \vdots & & & \ddots & -b_{n-1} \\ 0 & \cdots & & b_{n-1} & 0 \end{bmatrix}.$$

Then the system of equations (4.3.12) can be checked to be equivalent to the matrix equation

$$(4.3.15) \quad \frac{dM^+}{dt} = 2[M^-, M^+].$$

Adler [Adle1] and Kostant [Kost6] gave another interpretation of the system (4.3.12). Let  $\mathfrak{b}^0$  be the Lie algebra of upper triangular matrices of trace zero and  $\mathfrak{n}^+$  the subalgebra of strictly upper triangular matrices. By means of the bilinear form on  $n \times n$  matrices

$$(4.3.16) \quad X, Y \rightarrow \text{tr}(XY), \quad X, Y \in M_n(\mathbf{R}),$$

where  $\text{tr}$  denotes the trace of a matrix, we can identify the space  $M_n(\mathbf{R})$  of real  $n \times n$  matrices with its dual. In this identification, the traceless matrices

$M_r(\mathbf{R})^0 \simeq \mathfrak{sl}_n(\mathbf{R})$  are self-dual, and the dual of  $\mathfrak{b}^{0+}$  is identified to  $\mathfrak{sl}_n(\mathbf{R})$  modulo its annihilator, which is  $\mathfrak{n}^+$ . Thus

$$(4.3.17) \quad (\mathfrak{b}^{0+})^* \simeq \mathfrak{sl}_n(\mathbf{R})/\mathfrak{n}^+ \simeq \mathfrak{b}^{0-},$$

where  $\mathfrak{b}^{0-}$  is the lower triangular matrices of trace zero. In  $\mathfrak{b}^{0-}$ , consider the set of matrices

$$(4.3.18) \quad M = \begin{bmatrix} a_1 & 0 & 0 & \cdots & 0 \\ b_1 & a_2 & 0 & & \\ 0 & b_2 & a_3 & & \\ & & & \ddots & \\ 0 & \cdots & 0 & b_{n+1} & a_n \end{bmatrix}, \quad b_j \neq 0, \sum_{j=1}^n a_j = 0.$$

It is easy to check that under the coadjoint action  $\text{Ad}^* B^{0+}$  (where  $B^{0+} = \exp \mathfrak{b}^{0+}$ ) on  $\mathfrak{b}^{0-} \simeq (\mathfrak{b}^{0+})^*$ , the matrices (4.3.18) form a single orbit, which we will refer to as the *Toda orbit*. One can also verify that the invariant symplectic form (cf. §3.2) on the Toda orbit is, up to multiples,

$$(4.3.19) \quad \omega = \sum_{1 \leq k \leq j \leq n-1} \frac{db_j \wedge da_k}{b_j} = \sum_{j=1}^{n-1} \frac{db_j \wedge dh_j}{b_j},$$

where  $h_j = \sum_{k=1}^j a_k$ . If on this orbit one takes the Hamiltonian function

$$(4.3.20) \quad \frac{1}{2} \sum_{j=1}^{n-1} b_j^2 + \sum_{k=1}^n a_k^2 = \frac{1}{4} \text{tr}(M + M^t)^2,$$

then the Hamiltonian flow defined by (4.3.20) with respect to the symplectic form (4.3.19) is again the system (4.3.12). In equation (4.3.20),  $M^t$  denotes the transpose of the matrix  $M$ .

Thus we have two Lie-theoretic interpretations of the Toda lattice: one as a Hamiltonian flow on a coadjoint orbit for  $B^{0+}$ , with respect to a quadratic Hamiltonian related to the embedding  $\mathfrak{b}^{0+} \subseteq \mathfrak{sl}_n(\mathbf{R})$ , the other as a flow on the tridiagonal symmetric matrices, defined by commutator with a skew-symmetric matrix. The two interpretations are related by the equations

$$(4.3.21) \quad M^+ = \frac{1}{2}(M + M^t), \quad M^- = \frac{1}{2}(M - M^t)$$

with  $M$  as in equation (4.3.18) and  $M^+$  and  $M^-$  as in (4.3.13) and (4.3.14).

The form (4.3.15) of the Toda lattice equations, involving the commutator of a matrix with a skew-symmetric matrix depending on it, is known as *Lax form*. Its significance is that, since commutator is the infinitesimal version of conjugation, it immediately implies that all matrices in the trajectory through  $M^+$  will be conjugate to  $M^+$  (and by an orthogonal matrix). Hence the flow will be *isospectral*; the eigenvalues of  $M^+$  will be constant on each trajectory of the flow. If these eigenvalues (or, more properly, symmetric functions of them) are pulled back to the Toda orbit (4.3.18) via the map (4.3.21), they

define  $n - 1$  integrals of the Toda flow. Since the Toda orbit has dimension  $2(n - 1)$ , we see that if these integrals of the Toda flow Poisson commute with each other (i.e., if their associated Hamiltonian vector fields commute), then the Toda flow will be completely integrable, and the orbits of all the flows together will fill out the tridiagonal matrices of given eigenvalues. This is indeed the case, as we shall show.

Involved in this story are several actions related to one another by the geometry of  $G = \mathrm{SL}_n(\mathbf{R})$ . As above, let  $B^{0+}$  be the connected group whose Lie algebra is  $\mathfrak{b}^{0+}$ . Let  $K = \mathrm{SO}_n(\mathbf{R})$ —this is a maximal compact subgroup of  $\mathrm{SL}_n(\mathbf{R})$ . We have the Iwasawa, or for  $\mathrm{SL}_n(\mathbf{R})$ , the Gram-Schmidt decomposition (cf. §A.2.3)

$$(4.3.22) \quad \mathrm{SL}_n(\mathbf{R}) = B^{0+}K, \quad K \cap B^{0+} = 1.$$

Thus we have an identification

$$(4.3.23) \quad B^{0+} \backslash \mathrm{SL}_n(\mathbf{R}) \simeq K$$

so that the action of  $K$  on itself by right translation extends to an action of all of  $\mathrm{SL}_n(\mathbf{R})$ . Understanding the Toda lattice involves looking at this action from several points of view.

The first point of view is to express the  $\mathrm{SL}_n(\mathbf{R})$  action pointwise in terms of the  $K$ -action. Let

$$(4.3.24) \quad g = b(g)k(g), \quad g \in \mathrm{SL}_n(\mathbf{R}),$$

be the decomposition of an element  $g$  in  $\mathrm{SL}_n(\mathbf{R})$  according to (4.3.22): thus  $b(g) \in B^{0+}$  and  $k(g) \in K$ . We clearly have

$$(4.3.25) \quad b(b_1 g k_1) = b_1 b(g), \quad k(b_1 g k_1) = k(g)k_1$$

for  $g \in \mathrm{SL}_n(\mathbf{R})$ ,  $b_1 \in B^{0+}$ ,  $k_1 \in K$ . We also observe that  $g \rightarrow k(g)$  is the mapping which implements the identification (4.3.23). We may compute

$$(4.3.26) \quad k(k_1 g) = k((k_1 g k_1^{-1})k_1) = k(k_1 g k_1^{-1})k_1.$$

We can also express this infinitesimally. If  $g = \exp tX$ , for  $X \in \mathfrak{sl}_n(\mathbf{R})$ , then we can find  $\kappa(X, k_1)$  such that

$$\frac{d}{dt}(k_1^{-1}k(k_1 \exp tX))|_{t=0} = \kappa(X, k_1).$$

To express  $\kappa$ , we need the decomposition

$$(4.3.27) \quad \mathfrak{sl}_n(\mathbf{R}) \simeq \mathfrak{so}_n \oplus \mathfrak{b}^{0+} \simeq \mathfrak{k} \oplus \mathfrak{b}^{0+},$$

which is the infinitesimal version of decomposition (4.3.22). Let

$$(4.3.28) \quad p_k: \mathfrak{sl}_n(\mathbf{R}) \rightarrow \mathfrak{k}, \quad p_b: \mathfrak{sl}_n(\mathbf{R}) \rightarrow \mathfrak{b}^{0+}$$

be the projection maps associated with equation (4.3.27). In terms of these maps, by plugging  $g = \exp tX$  into formula (4.3.26) and differentiating, we can compute

$$(4.3.29) \quad \kappa(X, k_1) = \mathrm{Ad} k_1^{-1}(p_k(\mathrm{Ad} k_1(X))).$$

Now consider the Cartan decomposition (cf. §A.2.3.1)

$$(4.3.30) \quad \mathfrak{sl}_n(\mathbf{R}) \simeq \mathfrak{k} \oplus \mathfrak{p},$$

$$M = M^+ + M^- = \frac{1}{2}(M + M^t) + \frac{1}{2}(M - M^t)$$

of a matrix into its symmetric and skew-symmetric parts. The summands  $\mathfrak{k}$  and  $\mathfrak{p}$  are invariant under  $\text{Ad } K$ . In particular, we have an action of  $K$  on  $\mathfrak{p}$ , the space of symmetric matrices. The basic theorem on diagonalizing symmetric matrices tells us that

$$(4.3.31) \quad \text{Ad } K(\mathfrak{a}^0) = \mathfrak{p}.$$

Furthermore, if  $X$  in  $\mathfrak{a}^0$  is regular, in the sense that all the eigenvalues of  $X$  are distinct (so that no roots of  $\mathfrak{a}^0$  in  $\mathfrak{sl}_n(\mathbf{R})$ , which are differences of eigenvalues, vanish on  $X$ ), then the stabilizer of  $X$  in  $K$  is discrete, so that the map  $k \rightarrow \text{Ad } k(X)$  is almost injective, i.e., it defines a covering by  $K$  of the  $\text{Ad } K$  orbit of  $X$ .

The decomposition (4.3.30) is orthogonal with respect to the inner product (4.3.16), hence that inner product restricts nondegenerately to  $\mathfrak{p}$ . (In fact, it is positive definite.) Given a function  $f$  on  $\mathfrak{p}$ , we can form its gradient  $\nabla f$  in the standard way [CoSe; Lang2, p.186], by the formula

$$(4.3.32) \quad \text{tr}(\nabla f(y)z) = df(y)(z) = \partial_z(f)(y), \quad y, z \in \mathfrak{p},$$

where  $df(y) \in \mathfrak{p}^*$  is the usual differential of  $f$  at  $y$ , and  $\partial_z f$  indicates the directional derivative in the direction  $z$ . Using  $\nabla f$  and the projection  $p_k$  (cf. (4.3.28)) to  $\mathfrak{k}$  along  $\mathfrak{b}^{0+}$ , we can define a Lax form dynamical system on  $\mathfrak{p}$  by the differential equation

$$(4.3.33) \quad \frac{dy}{dt} = [p_k(\nabla f(y)), y].$$

This will give a flow which preserves each  $\text{Ad } K$ -orbit in  $\mathfrak{p}$ , since the tangent vector at each point  $y$  is by definition in  $\text{ad } \mathfrak{k}(y)$ , which is the tangent space to  $\text{Ad } K(y)$  at  $y$ . Thus for each regular  $y$ , we get by transport of structure a flow on  $K$ . This is explicitly defined as follows. Fix a regular  $y_0$ , and represent the general  $y$  in  $\text{Ad } K(y_0)$  in the form  $y = \text{Ad } k_1(y_0)$ . Then equation (4.3.33) reads

$$\begin{aligned} \frac{d}{dt}(\text{Ad } k_1(y_0)) &= [p_k \nabla f(\text{Ad } k_1(y_0)), \text{Ad } k_1(y_0)] \\ &= \text{Ad } k_1[\text{Ad } k_1^{-1}(\nabla f(\text{Ad } k_1(y_0))), y_0] \end{aligned}$$

or

$$(4.3.34) \quad (\text{Ad } k_1)^{-1} \frac{d}{dt}(\text{Ad } k_1(y_0)) = \text{ad}(\text{Ad } k_1^{-1}(p_k(\nabla f(\text{Ad } k_1(y_0))))(y_0).$$

Now suppose that the function  $f$  is invariant under  $\text{Ad } K$ . From the chain rule

$$d(f \circ \text{Ad } k) = ((df) \circ \text{Ad } k) \text{Ad } k, \quad k \in K,$$



we deduce from the invariance of  $f$  ( $f = f \circ \text{Ad } k$ ) and the definition (4.3.32) of  $\nabla f$ , that

$$(4.3.35) \quad \nabla f(\text{Ad } k(y)) = \text{Ad } k(\nabla f(y)), \quad k \in K, y \in \mathfrak{p}.$$

If we plug this into equation (4.3.34), and suppress the  $y_0$ , writing  $\nabla f(y_0) = X_0$ , we obtain the following equation for  $k_1$ :

$$(4.3.36) \quad k_1^{-1} \frac{d}{dt} k_1 = \text{Ad } k_1^{-1}(p_k(\text{Ad } k_1(X_0))).$$

If we compare this with equation (4.3.29) for the right action of  $\text{SL}_n(\mathbf{R})$  on  $B^{0+} \backslash \text{SL}_n(\mathbf{R})$  we obtain the following result.

**PROPOSITION 4.3.37.** *Let  $f$  be an  $\text{Ad } K$ -invariant function on  $\mathfrak{p}$ . Then the flow on  $\mathfrak{p}$  defined by the Lax-type equations (4.3.33) is equivalent on each  $\text{Ad } K$ -orbit  $\text{Ad } K(y_0)$  to a flow induced by right translations by a one parameter group of  $\text{SL}_n(\mathbf{R})$  acting on  $B^{0+} \backslash \text{SL}_n(\mathbf{R}) \simeq K$ . Precisely, under the covering  $K \rightarrow \text{Ad } K(y_0)$  by the map  $k_1 \rightarrow \text{Ad } k_1(y_0)$ , the flow (4.3.33) on  $\text{Ad } K(y_0)$  is identified to right translation on  $B^{0+} \backslash \text{SL}_n(\mathbf{R})$  by  $\exp(t \nabla f(y_0))$ .*

To tighten the connection made by Proposition 4.3.37, we need another observation about  $\nabla f$  for  $\text{Ad } K$ -invariant  $f$ . For regular  $y_0 \in \mathfrak{p}$ , let  $\mathfrak{c}(y_0)$  be the centralizer of  $y_0$  in  $\mathfrak{p}$ . Thus if  $y_0 = \text{Ad } k_0(a_0)$ , for  $a_0 \in \mathfrak{a}^0$ , we have  $\mathfrak{c}(y_0) = \text{Ad } k_0(\mathfrak{a}^0)$ . We have

$$(4.3.38) \quad \nabla f(y_0) \in \mathfrak{c}(y_0).$$

Indeed, we know  $\nabla f(y_0)$  is orthogonal to the level set of  $f$  through  $y_0$ . Since  $f$  is  $\text{Ad } K$ -invariant, the level set includes  $\text{Ad } K(y_0)$ , and the tangent space at  $y_0$  to  $\text{Ad } K(y_0)$  is  $\text{ad } k(y_0)$ . Since we have assumed  $y_0$  to be regular, this equals  $\mathfrak{c}(y_0)^\perp$ , the orthogonal complement to  $\mathfrak{c}(y_0)$  in  $\mathfrak{p}$ , as can easily be seen in the case  $y_0 \in \mathfrak{a}^0$ . Since all elements of  $\mathfrak{c}(y_0)$  commute with one another, combining Proposition 4.3.37 with observation (4.3.38) yields the following conclusion.

**COROLLARY 4.3.39.** *All the flows on  $\mathfrak{p}$ , associated to the Lax-type equations (4.3.33) for  $f$  which are  $\text{Ad } K$ -invariant, commute with each other.*

To complete the picture, it remains to relate the Lax-type equations to Hamiltonian systems on coadjoint orbits. Here again the relation (4.3.38) is a key ingredient. It implies that, again taking  $f$  to be  $\text{Ad } K$ -invariant,

$$(4.3.40) \quad [\nabla f(y), y] = 0, \quad y \in \mathfrak{p}.$$

Take  $x, y$  in  $\mathfrak{p}$ , and decompose  $x$  as in equation (4.3.27). Then if  $x, y$  commute, we have

$$(4.3.41) \quad [p_k(x), y] = -[p_b(x), y], \quad x, y \in \mathfrak{p}; [x, y] = 0.$$

Let  $f$  be an  $\text{Ad } K$ -invariant function on  $\mathfrak{p}$ . Extend  $f$  to a function on all  $\mathfrak{g}$  by letting  $f$  be constant on cosets of  $\mathfrak{k}$ . Then  $\nabla f(y)$ ,  $y \in \mathfrak{p}$ , is the same as

when  $f$  was considered as a function on  $\mathfrak{p}$ . Thus from (4.3.40) and (4.3.41) we deduce

$$(4.3.42) \quad [p_k(\nabla f(y)), y] = -[p_b(\nabla f(y)), y].$$

There is a natural projection

$$(4.3.43) \quad \alpha : \mathfrak{g} \rightarrow (\mathfrak{b}^{0+})^*$$

induced by the bilinear form (4.3.16). The restriction of  $\alpha$  to  $\mathfrak{p}$  is a linear isomorphism of  $\mathfrak{p}$  onto  $(\mathfrak{b}^{0+})^*$ . Since  $\alpha$  is  $\text{Ad } B^{0+}$ -equivariant, we have

$$(4.3.44) \quad \alpha(\text{ad } p_b(\nabla f(y))(y)) = \text{ad } p_b(\nabla f(y))\alpha(y).$$

The function  $f$  on  $\mathfrak{p}$  pushes down via  $\alpha$  to define a function  $H_f = f \circ \alpha^{-1}$  on  $\mathfrak{b}^{0+}$ . Checking through the various identifications and using equations (4.3.42) and (4.3.44) we can verify [GoWa, Syme]

**PROPOSITION 4.3.45.** *Let  $f$  be an  $\text{Ad } K$ -invariant function on  $\mathfrak{p}$ , and let  $H_f = f \circ \alpha^{-1}$  be the push-forward of  $f$  to  $(\mathfrak{b}^{0+})^*$  via  $\alpha$ . Then the Lax-type system (4.3.33), pushed forward to  $(\mathfrak{b}^{0+})^*$  by  $\alpha$ , becomes the Hamiltonian system defined by  $-H_f$  on each  $B^{0+}$  coadjoint orbit in  $\mathfrak{b}^{0+}$ .*

If we now combine Proposition 4.3.45 with Corollary 4.3.39, we find the  $n-1$  Hamiltonians corresponding to  $\text{trace}(\alpha^{-1}(M))^k$ , for  $k = 2, 3, \dots, n$ , form a Poisson commutative family on any  $\text{Ad}^* B^{0+}$ -orbit in  $(\mathfrak{b}^{0+})^*$ . In particular, since the Toda orbit (4.3.18) has dimension  $2(n-1)$  they define on it (assuming they are functionally independent, which is easy to check) a completely integrable family. Thus the Toda lattice is completely integrable.

**REMARKS.** (a) The argument above shows that the  $H_f = f \circ \alpha^{-1}$  form a Poisson commutative family on any  $\text{Ad}^* B^{0+}$ -orbit in  $(\mathfrak{b}^{0+})^*$ . However, only for special orbits which have dimension  $2(n-1)$  or less could we hope to conclude complete integrability from this fact. In [GoWa] other orbits where this scheme gives complete integrability are described. On the other hand, in [DLNT] more Poisson commuting functions are found on  $(\mathfrak{b}^{0+})^*$ , enough to provide complete integrability for the Toda flow on a generic  $\text{Ad}^* B^{0+}$ -orbit in  $(\mathfrak{b}^{0+})^*$ .

(b) Systems analogous to the Toda lattice can be constructed on any split real semisimple group [Kost6, Syme1].

(c) A number of other completely integrable systems on coadjoint orbits have been found. For example, by considering a nested chain  $\mathfrak{u}_1 \subseteq \mathfrak{u}_2 \subseteq \mathfrak{u}_3 \subseteq \dots \subseteq \mathfrak{u}_n$  of unitary Lie algebras  $\text{Thimm}([\text{Thim}]$  (see also [GuSt5]) was able to construct completely integrable systems on coadjoint orbits in  $\mathfrak{u}_n^*$ . Included among these are the geodesic flows on Grassmannians (cf. §1.4). Similar considerations apply to orthogonal Lie algebras.

(d) Using the Bruhat decomposition (cf. §A.2.3.3), one can give a very clear description of the dynamics of right translation on  $B^{0+} \backslash \text{SL}_n(\mathbf{R})$  by

$\exp tX$ ,  $X \in \mathfrak{p}$ . It behaves very simply: it is essentially a gradient flow, with nondegenerate, isolated fixed points, parametrized by the Weyl group. In fact, the Bruhat decomposition arises as the Morse decomposition [Miln2] associated to this flow. Similar facts can be shown to hold for the holomorphic action of  $\mathbb{C}^\times$  on projective varieties [Carr]. (Of course  $B^{0+} \backslash \mathrm{SL}_n(\mathbb{R})$  is not a complex variety, but if we had been working with  $\mathrm{SL}_n(\mathbb{C})$  rather than  $\mathrm{SL}_n(\mathbb{R})$ , we would have been dealing with the complex flag manifold, which is a complex variety.) The geometric analysis allows one to describe quite precisely the asymptotics of Toda trajectories.

(e) The Toda lattice can be solved quite explicitly [Kost6, Syme1, GoWa]. It turns out to be closely related to the famous “QR algorithm” for diagonalizing matrices. It is essentially a continuous-time version of this algorithm [Syme2, DeNT, GoWa].

(f) It should be noted that Proposition 4.3.45 is a generalization of the results of the explicit calculations (4.3.10) through (4.3.21).

**Appendix 1: Basic concepts of representation theory.** An account of representation theory must begin with some basic definitions and some remarks about certain technical issues. These latter can be somewhat off-putting, but if openly acknowledged, their negative effects can be minimized. The reader is advised to skim this section, and refer to it as necessary. For greater detail on this basic material, we refer to [FeDo, Gaal, Kiri, Lang1], etc.

A.1.1. Let  $G$  be a Lie group. A *representation*  $\rho$  of  $G$  on a vector space  $V$  is a homomorphism of  $G$  into the group of invertible linear transformations of  $V$ :

$$(A.1.1.1) \quad \rho : G \rightarrow \mathrm{GL}(V).$$

To be complete, in referring to a representation, we should specify both  $\rho$  and  $V$ , but often we will only specify  $\rho$ , letting  $V$  be understood implicitly; or we may just specify  $V$  and let  $\rho$  be implicit, in which case we call  $V$  a *G-module*.

A.1.2. Very often  $V$  will be infinite dimensional, and then usually it is equipped with a topology. Although the case of greatest general interest is when  $V$  is a Hilbert space, sometimes it is a Banach space, and it is not really possible to avoid considering situations when  $V$  is only locally convex. Whatever the topology of  $V$ , one wants to put a continuity condition on  $\rho$ . The correct one is that  $\rho$  should be *strongly continuous*:

$$(A.1.2.1) \quad \text{The map } g \rightarrow \rho(g)v, \text{ from } G \text{ to } V, \text{ should be continuous for all } v \in V.$$

**REMARK.** If  $V$  is a Banach space, one might be tempted to think the map  $g \rightarrow \rho(g)$  should be continuous with respect to the norm topology on the operators on  $V$ . But this condition is far too restrictive, and hardly ever holds when  $V$  is infinite dimensional.

$\exp tX$ ,  $X \in \mathfrak{p}$ . It behaves very simply: it is essentially a gradient flow, with nondegenerate, isolated fixed points, parametrized by the Weyl group. In fact, the Bruhat decomposition arises as the Morse decomposition [Miln2] associated to this flow. Similar facts can be shown to hold for the holomorphic action of  $\mathbb{C}^\times$  on projective varieties [Carr]. (Of course  $B^{0+} \backslash \mathrm{SL}_n(\mathbb{R})$  is not a complex variety, but if we had been working with  $\mathrm{SL}_n(\mathbb{C})$  rather than  $\mathrm{SL}_n(\mathbb{R})$ , we would have been dealing with the complex flag manifold, which is a complex variety.) The geometric analysis allows one to describe quite precisely the asymptotics of Toda trajectories.

(e) The Toda lattice can be solved quite explicitly [Kost6, Syme1, GoWa]. It turns out to be closely related to the famous “QR algorithm” for diagonalizing matrices. It is essentially a continuous-time version of this algorithm [Syme2, DeNT, GoWa].

(f) It should be noted that Proposition 4.3.45 is a generalization of the results of the explicit calculations (4.3.10) through (4.3.21).

**Appendix 1: Basic concepts of representation theory.** An account of representation theory must begin with some basic definitions and some remarks about certain technical issues. These latter can be somewhat off-putting, but if openly acknowledged, their negative effects can be minimized. The reader is advised to skim this section, and refer to it as necessary. For greater detail on this basic material, we refer to [FeDo, Gaal, Kiri, Lang1], etc.

A.1.1. Let  $G$  be a Lie group. A *representation*  $\rho$  of  $G$  on a vector space  $V$  is a homomorphism of  $G$  into the group of invertible linear transformations of  $V$ :

$$(A.1.1.1) \quad \rho : G \rightarrow \mathrm{GL}(V).$$

To be complete, in referring to a representation, we should specify both  $\rho$  and  $V$ , but often we will only specify  $\rho$ , letting  $V$  be understood implicitly; or we may just specify  $V$  and let  $\rho$  be implicit, in which case we call  $V$  a *G-module*.

A.1.2. Very often  $V$  will be infinite dimensional, and then usually it is equipped with a topology. Although the case of greatest general interest is when  $V$  is a Hilbert space, sometimes it is a Banach space, and it is not really possible to avoid considering situations when  $V$  is only locally convex. Whatever the topology of  $V$ , one wants to put a continuity condition on  $\rho$ . The correct one is that  $\rho$  should be *strongly continuous*:

$$(A.1.2.1) \quad \text{The map } g \rightarrow \rho(g)v, \text{ from } G \text{ to } V, \text{ should be continuous for all } v \in V.$$

**REMARK.** If  $V$  is a Banach space, one might be tempted to think the map  $g \rightarrow \rho(g)$  should be continuous with respect to the norm topology on the operators on  $V$ . But this condition is far too restrictive, and hardly ever holds when  $V$  is infinite dimensional.

From now on, all representations under discussion will be understood to be strongly continuous, unless the contrary is specifically stated.

A.1.3. A very important special class of representations are those for which  $V$  is a Hilbert space, and  $\rho(G)$  consists of unitary operators. These are called *unitary representations*.

A.1.4. Let  $\rho$  be a representation of  $G$  on the space  $V$ , and  $\sigma$  another representation on the space  $U$ . Then a (continuous) linear map  $T: V \rightarrow U$  is called an *intertwining operator* (or if we want to be modern, morphism of  $G$ -modules or  $G$ -morphism) if

$$(A.1.4.1) \quad \sigma(g)T = T\rho(g), \quad g \in G.$$

That is, if the diagram

$$\begin{array}{ccc} V & \xrightarrow{\rho(g)} & V \\ T \downarrow & & \downarrow T \\ U & \xrightarrow{\sigma(g)} & U \end{array}$$

commutes.

A.1.5. With notation as in §A.1.4, if  $T$  is a topological isomorphism between  $V$  and  $U$ , we call  $T$  an *equivalence*. If such a  $T$  exists, we say  $\rho$  and  $\sigma$  are equivalent. Very often, we are interested in representations only up to equivalence, and often when we say "representation" we actually mean equivalence class of representations. Context should usually make this clear.

A.1.5.1. If  $\rho, \sigma$  are unitary representations, then we feel best if an equivalence  $T$  is also unitary, in which case we speak of *unitary equivalence*. An easy argument using the polar decomposition (cf. [Gaal, Lang2, RiNa], etc.) of an operator guarantees that if  $\rho, \sigma$  are unitary and equivalent, then they are unitarily equivalent.

A.1.5.2. The notion of equivalence is a vexed one, owing to the extraordinary variety of linear topological vector spaces one has available. For example, consider the unit circle  $T$ . There are billions and billions of function spaces on  $T$ , perhaps most prominently the spaces  $L^p(T)$ , on which  $T$  acts via its action on itself by translations. Clearly these spaces bear a strong family resemblance to each other. However, the representations of  $T$  on them are not equivalent in the sense just defined, and generations of abelian harmonic analysts have derived pleasure from sorting out the differences [Katz]. They still are.

However, at the more primitive level which characterizes much of non-abelian harmonic analysis, one often would like to ignore differences such as those between  $L^p$  and  $L^q$ . For this purpose, various looser notions of equivalence have been formulated. None so far are terribly satisfactory for a wide range of applications, although there has been some progress made

in particular areas, [Fell1, 2, Warn]. Currently this is an unsettled, highly technical topic, which we will mostly ignore. An exception is the notion of infinitesimal equivalence of Harish-Chandra modules, which we will explain in §A.1.20 (see also §3.6.5).

A.1.6. Let  $\rho$  be a representation of  $G$  on  $V$ . Let  $V_1 \subseteq V$  be a closed subspace. We say  $V_1$  is  $G$ -invariant if  $\rho(g)V_1 \subseteq V_1$  for all  $g \in G$ . If  $V_1$  is  $G$ -invariant, then the map

$$\rho_1 : G \rightarrow \mathrm{GL}(V_1), \quad \rho_1(g) = \rho(g)|_{V_1},$$

defined by restriction to  $V_1$  is a representation of  $G$  on  $V_1$ . We call  $\rho_1$  arising in this way a *subrepresentation* of  $\rho$ .

A.1.7. Notations as in §A.1.6. If the only  $G$ -invariant subspaces of  $V$  are  $\{0\}$  and  $V$ , then we say  $V$  is *irreducible*.

A.1.7.1. Irreducibility, like equivalence, is not a completely satisfactory notion. A better one for many purposes is *topological complete irreducibility* (t.c.i.). We say  $\rho$  is t.c.i. if for any finite-dimensional subspace  $Y \subseteq V$ , any linear map  $T : Y \rightarrow V$ , and any open neighborhood  $U$  of  $T$  in  $\mathrm{Hom}(Y, V)$ , we can find a linear combination  $A = \sum a_i \rho(g_i)$  of elements of  $\rho(G)$  such that  $A|_Y$  is in  $U$ . That is, we can approximate an arbitrary operator of  $\mathrm{End}(V)$  arbitrarily closely on finite-dimensional subspaces of  $V$  by linear combinations from  $\rho(G)$ .

A.1.7.2. The condition t.c.i. implies irreducibility trivially. For unitary representations, irreducibility implies t.c.i. [Warn].

A.1.7.3. The set of unitary equivalence classes of irreducible unitary representations of a group  $G$  is called the *unitary dual* of  $G$ , and denoted  $\widehat{G}$ .

A.1.7.4. It is easy to check that if  $\rho$  is a t.c.i. representation of  $G$  on a space  $V$ , then the only operators on  $V$  which commute with  $\rho(G)$  are scalar multiples of the identity. In particular, if  $ZG$  is the center of  $G$ , then  $\rho(z)$ ,  $z \in ZG$ , must be a scalar multiple of the identity operator:

$$\rho(z) = \psi(z)I$$

for an appropriate complex number  $\psi(z) \in \mathbf{C}^x$ . It is immediate that  $\psi : ZG \rightarrow \mathbf{C}^x$  must be a group homomorphism, often called a *quasicharacter*. We call  $\psi$  the *central character* of  $\rho$ .

A.1.8. Notations as in §A.1.6. Let  $V_2$  be another  $G$ -invariant subspace of  $V$ . Let  $\rho_2$  be the subrepresentation of  $V$  defined by  $V_2$ . Suppose that  $V$  is the direct sum of  $V_1$  and  $V_2$ , i.e., the natural map

$$\begin{aligned} \alpha : V_1 \oplus V_2 &\rightarrow V, \\ \alpha(v_1, v_2) &= v_1 + v_2, \quad v_i \in V_i, \end{aligned}$$

is a linear isomorphism. Then we say  $\rho$  is the *direct sum* of the representations  $\rho_1$  and  $\rho_2$ .

A.1.8.1. If  $\rho$  is unitary, so that  $V$  is a Hilbert space, let  $V_1^\perp$  be the orthogonal complement of  $V_1$ . Then  $V_1^\perp$  is also  $G$ -invariant. Hence, if a unitary representation is reducible, it decomposes as a direct sum.

A.1.9. Let  $f$  be a complex-valued function on  $G$ . For  $g \in G$ , define the *left translation* of  $f$  by  $g$ ,  $L_g(f)$ , and the *right translation* of  $f$  by  $g$ ,  $R_g(f)$ , by the formulas

$$(A.1.9.1) \quad L_g(f)(h) = f(g^{-1}h), \quad R_g(f)(h) = f(hg), \quad g, h \in G.$$

(Actually, there is no need to require  $f$  to be complex-valued for these formulas—it could be vector-valued, or even set-valued.) Let  $C^G$  be the vector space of all complex-valued functions on  $G$  (no topology!). Then  $L: g \rightarrow L_g$  and  $R: g \rightarrow R_g$  are homomorphisms of  $G$  into  $GL(C^G)$ . They are called the *left-regular* and *right-regular* representations of  $G$ .

Suppose  $Y \subseteq C^G$  is some space of functions on  $G$ , invariant by left translations by  $G$ , and equipped with a topology such that the restriction of  $L$  to  $Y$  is strongly continuous (cf. (A.1.2.1)). As a condition of nondegeneracy, to ensure that  $Y$  consists of “most” functions on  $G$ , we will require that  $Y$  contain  $C_c^\infty(G)$ , the space of smooth functions of compact support. We will refer to the restriction of  $L$  to  $Y$  as the  *$Y$ -left-regular representation*, or the *left-regular representation on  $Y$* . Similar definitions apply to the right-regular representation. Frequently considered examples are  $C_c^\infty(G)$  itself;  $C_c(G)$ , the continuous functions of compact support;  $C_0(G)$ , the continuous functions vanishing at  $\infty$ ; and  $L^p(G)$ ,  $1 \leq p < \infty$  (these  $L^p$  spaces are understood to be with respect to the left-invariant Haar measure on  $G$  [HeRo, Loom, Nach]. Note that the action of  $G$  on  $L^\infty(G)$ , and if  $G$  is non-compact, even on  $C(G)$ , the bounded continuous functions, is *not* strongly continuous.

A.1.10. Given a representation  $\rho$  of  $G$  on a space  $V$ , we can define an action  $\rho^*$  of  $G$  on the dual space  $V^*$  of  $V$  by the formula

$$(A.1.10.1) \quad \rho^*(g)(\lambda)(v) = \lambda(\rho(g^{-1})v), \quad \lambda \in V^*, v \in V, g \in G.$$

The action  $\rho^*$  may not be strongly continuous. For example, if  $\rho$  is the left-regular representation on  $L^1(G) = V$ , then  $V^* = L^\infty(G)$ , and the resulting  $\rho^*$ , which is just the left-regular representation on  $L^\infty(G)$ , is not strongly continuous. However, if  $V$  is, for example, a reflexive Fréchet space, then  $\rho^*$  will be strongly continuous [Moor3, Warn]. Whenever it is, we call  $\rho^*$  the *contragredient* representation to  $\rho$ .

A.1.11. Let  $\rho$  be a representation of  $G$  on  $V$ . Select  $v \in V$  and  $\lambda \in V^*$ .

The function on  $G$  defined by

$$(A.1.11.1) \quad \varphi_{u,\lambda}(g) = \lambda(\rho(g)(u))$$

is called the *matrix coefficient* of  $\rho$  defined by  $\lambda$  and  $u$ . It is a continuous function on  $G$ . If  $\rho$  is unitary, or more generally if  $\rho$  is a bounded representation (i.e.,  $\|\rho(g)\| \leq M$  for all  $g \in G$  and some number  $M$ ) on a Banach space, then  $\varphi_{u,\lambda}$  is a bounded function on  $G$ .

A trivial formal calculation shows that

$$(A.1.11.2) \quad L_g \varphi_{u,\lambda} = \varphi_{u,\rho^*(g)(\lambda)}, \quad R_g \varphi_{u,\lambda} = \varphi_{\rho(g)(u),\lambda}.$$

Thus the maps

$$\Phi_\lambda : u \rightarrow \varphi_{u,\lambda}, \quad \Phi_u^* : \lambda \rightarrow \varphi_{u,\lambda}, \quad u \in V, \lambda \in V^*,$$

are intertwining maps; between  $\rho$  and the right-regular representation, and between  $\rho^*$  and the left-regular representation, respectively. (We will be vague here about exactly which space of functions we are using for our regular representations.) Thus matrix coefficients serve as a bridge between the abstract world of representations and the concrete world of functions, specifically on  $G$ . In doing this, they play a pivotal role in representation theory. They also provide an intimate link with classical mathematics, as most of the special functions of the nineteenth century physics are matrix coefficients of appropriate representations of appropriate groups [Mill1, Vile].

A.1.12. If  $\rho$  is a representation of  $G$  on  $V$ , then the linear span of  $\rho(G)$  is an algebra, as is its closure. It is often useful to be able to produce fairly general elements in this algebra. This is done by constructing the “integrated form” of  $\rho$ .

Let  $dg$  be the left-invariant Haar measure on  $G$  [HeRo, Loom, Nach]. For  $f_1, f_2$  in  $C_c^\infty(G)$ , we define the convolution

$$(A.1.12.1) \quad f_1 * f_2(h) = \int_G f_1(g) f_2(g^{-1}h) dg = \int_G f_1(g) L_g(f_2)(h) dg.$$

This product turns  $C_c^\infty(G)$  into an associative algebra. There is also an involutive, conjugate-linear, antiautomorphism (in brief: an involution) on  $C_c^\infty(G)$  defined (in the case when  $G$  is unimodular [HeRo, Loom, Nach]) by the formula

$$(A.1.12.2) \quad f^*(g) = \overline{f(g^{-1})}.$$

Here the overbar indicates complex conjugation. Note that

$$(A.1.12.3) \quad \check{f}(g) = f(g^{-1})$$

is a complex linear antiautomorphism of  $G$ .

Given a representation  $\rho$  of  $G$  on a space  $V$ , and a function  $f$  in  $C_c^\infty(G)$ , define  $\rho(f)$  by the recipe

$$(A.1.12.4) \quad \rho(f)(v) = \int_G f(g) \rho(g)(v) dg, \quad v \in V.$$



For this integral to be defined, the space  $V$  must have some mild completeness properties [Moor3]. It is more than enough that  $V$  be Fréchet. One checks by a formal computation that this definition of  $\rho(f)$  defines an algebra homomorphism from  $C_c^\infty(G)$ , with convolution as product, to the algebra  $\text{End}(V)$  of continuous linear transformations on  $V$ .

Unless the contrary is stated, we will understand that  $\rho$  has an integrated form defined by (A.1.12.4).

If  $\rho$  is unitary, then formula (A.1.12.4) defines a  $*$ -homomorphism of  $C_c^\infty(G)$  into  $\text{End}(V)$ :

$$(A.1.12.5) \quad \rho(f^*) = \rho(f)^*,$$

where the  $*$  on the right-hand side of the equation indicates the adjoint of an operator on a Hilbert space (and the  $*$  on the left is as in formula (A.1.12.2)).

The algebra  $C_c^\infty(G)$  can be completed in various norms to produce a Banach  $*$ -algebra. The two most commonly considered ones are

$$(A.1.12.6) \quad \text{The } L^1\text{-norm: } \|f\|_1 = \int_G |f(g)| dg,$$

$$\text{The } C^*\text{-norm: } \|f\|_* = \sup\{\|\rho(f)\| : \rho \text{ unitary}\}.$$

An easy estimate shows that if  $\rho$  is a representation of  $G$  by isometries on a Banach space, then  $\rho$  is norm decreasing with respect to  $\|\cdot\|_1$ :

$$(A.1.12.7) \quad \|\rho(f)\| \leq \|f\|_1 \quad \text{if } \rho \text{ is isometric in a Banach space.}$$

Here  $\|\cdot\|$  indicates the norm of an operator on the space of the representation. Estimate (A.1.12.7) shows in particular that the supremum involved in defining  $\|\cdot\|_*$  exists, such that

$$(A.1.12.8) \quad \|f\|_* \leq \|f\|_1.$$

The completion of  $C_c^\infty(G)$  with respect to  $\|\cdot\|_*$  is a  $C^*$ -algebra [FeDo, Gaal].

The algebra  $C_c^\infty(G)$  does not have an identity element. However, it does have an *approximate identity* or *Dirac sequence* (cf. [HeRo, Lang1, 2], etc.). This is a sequence of functions  $\{f_n\}$  such that

$$(A.1.12.9) \quad \lim_{n \rightarrow \infty} f_n * \varphi = \varphi, \quad \varphi \in C_c^\infty(G).$$

One constructs a Dirac sequence by choosing functions  $f_n$  whose support is concentrated in smaller and smaller neighborhoods of the identity. Such a sequence is of course far from unique. When convenient, one can assume that  $f_n$  is nonnegative, and that  $\int_G f_n(g) dg = 1$  for all  $n$ . Given an approximate identity  $\{f_n\}$ , and a representation  $\rho$  of  $G$ , one has

$$(A.1.12.10) \quad \lim_{n \rightarrow \infty} \rho(f_n)(v) = v.$$

In particular, the span of vectors of the form  $\rho(f)v$ , for  $f \in C_c^\infty(G)$  and  $v \in V$ , is dense in  $V$ .

A.1.13. The construction of A.1.12, with  $C_c^\infty(G)$  replaced by  $C_c(G)$ , works for a general locally compact group. For a Lie group  $G$ , one also has the Lie algebra  $\text{Lie}(G)$ , and its universal enveloping algebra, and it is very much of the essence to represent them also. The procedure for doing this is somewhat more involved than in §A.1.12, since the operators of  $\text{Lie}(G)$  will typically be unbounded if  $V$  is infinite dimensional.

Consider  $X \in \text{Lie}(G)$  and the associated one-parameter group  $\{\exp tX\} \subseteq G$ . Then  $\rho(\exp tX)$  is a one-parameter group in  $\text{GL}(V)$ , and  $\rho(X)$  should be the infinitesimal generator of this one-parameter group. The procedure for finding  $\rho(X)$  is clear from §2 (see formula (2.5.1)). We set

$$(A.1.13.1) \quad \rho(X)(v) = \lim_{t \rightarrow 0} \frac{\rho(\exp tX)v - v}{t}$$

for any  $v$  for which this exists. (To be consistent with formula (2.5.1), we should denote  $\rho(X)$  by  $d\rho(X)$ ; but we now drop the  $d$ .) The formula

$$(A.1.13.2) \quad \rho(\exp sX)v - v = \int_0^s \rho(\exp tX)(\rho(X)v) dt$$

guarantees that  $\rho(X)$  has a closed graph. If  $f \in C_c^\infty(G)$ , then one checks directly from the definitions that  $\rho(X)\rho(f)v$  exists, for any  $v \in V$ ; precisely, one has the formula

$$(A.1.13.3) \quad \rho(X)\rho(f)(v) = \rho(X(f))(v),$$

where  $X(f)$  denotes the usual operation of differentiating the smooth function  $f$  with respect to the vector field  $X$ . From the final result in §A.1.12, it follows that  $\rho(X)$  is densely defined. Thus  $\rho(X)$  is a closed, densely defined operator on  $V$ .

In fact, much more is true. The span of the vectors  $\rho(f)(v)$ ,  $f \in C_c^\infty(G)$ ,  $v \in V$ , form a dense subspace on which all  $\rho(X)$ ,  $X \in \text{Lie}(G)$ , are defined; and moreover, this subspace is stable under the  $\rho(X)$ 's by formula (A.1.13.3). This suggests the following construction. Define  $V^\infty$  to be the subspace of  $V$  consisting of vectors such that, for any sequence  $X_1 X_2 \cdots X_n$  of elements in  $\text{Lie}(G)$ , the composition  $\rho(X_n)\rho(X_{n-1}) \cdots \rho(X_1)(v)$  is defined. Topologize  $V^\infty$  by requiring that a new  $\{v_\alpha\}$  converges to a limit  $v_0$  in  $V^\infty$  if and only if, for all sequences of  $X_i$ 's, as above  $\rho(X_n)\rho(X_{n-1}) \cdots \rho(X_1)(v_\alpha)$  converges to  $\rho(X_n) \cdots \rho(X_1)(v_0)$  in  $V$ . Then one shows

$$(A.1.13.4) \quad \text{Scholium: The space } V^\infty \text{ is dense in } V; \text{ it is stable under } \rho(X), X \in \text{Lie}(G); \text{ also, it is stable under } G; \text{ and the action of } G \text{ on } V^\infty \text{ is strongly continuous.}$$

We call  $V^\infty$  the space of *smooth vectors*, and we call the action of  $G$  on  $V^\infty$  the *smooth representation associated to*  $\rho$ . This is denoted sometimes by  $\rho^\infty$ , or sometimes simply again by  $\rho$ .

If  $V = V^\infty$  (as topological vector spaces) we say  $\rho$  is a *smooth representation*. It is clear that  $(V^\infty)^\infty = V^\infty$ , so that  $\rho^\infty$  is always a smooth representation.

If  $V$  is complete, then the fact that the  $\rho(X)$  have closed graphs allows one to conclude that  $V^\infty$  is complete. Since, in the definition of  $V^\infty$  it would have sufficed to choose the elements  $X_i$  from a basis of  $\text{Lie}(G)$ , we see that if  $V$  is Fréchet, in particular, if  $V$  is Banach, then  $V^\infty$  is Fréchet.

We have seen that vectors of the form  $\rho(f)v$ ,  $f \in C_c^\infty(G)$ ,  $v \in V$ , belong to  $V^\infty$ . The span of these vectors is known as the *Gårding space*. For a long time the relation between  $V^\infty$  and the Gårding space was unclear. Then Dixmier and Malliavin showed they were equal [DiMa].

REMARK. If  $\mathfrak{g}$  is a Lie algebra, the associative algebra generated by  $\mathfrak{g}$  subject to the relations  $xy - yx = [x, y]$  (where the left-hand side indicates the usual commutator in an associative algebra, and the right-hand side is the Lie bracket in  $\mathfrak{g}$ ) is called the *universal enveloping algebra* of  $\mathfrak{g}$ . We will denote it by  $\mathcal{U}(\mathfrak{g})$ . Any representation

$$\rho : \mathfrak{g} \rightarrow \text{End}(V)$$

of  $\mathfrak{g}$  on a vector space  $V$  automatically extends to a homomorphism

$$\rho : \mathcal{U}(\mathfrak{g}) \rightarrow \text{End}(V)$$

of associative algebras. Thus, in particular, in the context of this section, we have an action of  $\mathcal{U}(\text{Lie}(G))$  on  $V^\infty$ .

A.1.14. A major method for constructing representations is by *induction*. The study of induced representations for locally compact groups was pioneered by Mackey (cf. [Mack4, FeDo, Warn], etc.).

Let  $H \subseteq G$  be a closed subgroup. Let  $\sigma$  be a representation of  $H$  on a space  $U$ . Define  $C_c^\infty(H \backslash G; \sigma)$  to be the space of smooth functions from  $G$  to  $U$  whose support lies in some set of the form  $HS$ ,  $S$  compact, and which transform by  $\sigma$  under left translations by  $H$ :

$$(A.1.14.1) \quad \varphi(hg) = \sigma(h)\varphi(g), \quad h \in H, g \in G, \varphi \in C_c^\infty(H \backslash G; \sigma).$$

It is easy to check that  $C_c^\infty(H \backslash G; \sigma)$  is invariant under right translations by  $G$ . Further  $C_c^\infty(H \backslash G; \sigma)$  comes equipped with a standard topology [Warn], and it is not hard to see that with respect to this topology the action of  $G$  is strongly continuous. The representation of  $G$  so defined will be called the *representation of  $G$  induced by  $\sigma$* , or more particularly, the  *$C_c^\infty$ -induced representation induced by  $\sigma$* . It will be denoted by  $\text{ind}_H^G \sigma$ , or  $C_c^\infty - \text{ind}_H^G \sigma$  if we wish to be more specific. More generally, if  $Y$  is a space obtained by completing  $C_c^\infty(H \backslash G; \sigma)$  in some topology weaker than its natural topology, we will also call the action of  $G$  on  $Y$ , extended from  $C_c^\infty(H \backslash G; \sigma)$  by continuity, a representation induced by  $\sigma$ .

A.1.15. Continue the notations of §A.1.1.4. Consider the space  $C^\infty(G; U)$  of smooth functions on  $G$  with values in  $U$ . We can consider two actions of  $H$  on  $C^\infty(G; U)$ , the action by left translations, as defined in (A.1.9.1), and the action by transforming the values of elements of  $C^\infty(G; U)$  by  $\sigma$ .

These two actions commute with the action of  $G$  by right translations and they also commute with each other, so we can form their “tensor product,” which we will denote by  $\sigma \otimes R$ . Explicitly in formulas, these actions are given by

$$(A.1.15.1) \quad \begin{aligned} L_h f(g) &= f(h^{-1}g), \quad f \in C^\infty(G; U), \\ \sigma(h)(f)(g) &= \sigma(h)(f(g)), \\ (\sigma \otimes L)(h)(f)(g) &= \sigma(h)(f(h^{-1}g)). \end{aligned}$$

In terms of these actions, we can see that  $C_c^\infty(H \backslash G; \sigma)$  is defined by a condition of invariance under the action  $\sigma \otimes L$  of  $H$ , as well as a support condition relative to  $H \backslash G$ .

Let  $C_c^\infty(G; U)$  be the elements of  $C^\infty(G; U)$  with compact support. For  $\varphi \in C_c^\infty(G; U)$ , define a function  $p_\sigma(\varphi)$  in  $C^\infty(G; U)$  by the recipe

$$(A.1.15.2a) \quad p_\sigma(\varphi)(g) = \int_H \sigma(h)^{-1} \varphi(hg) d_r h,$$

where  $d_r h$  is a right-invariant Haar measure on  $H$ . In terms of notations (A.1.15.1), we may write

$$(A.1.15.2b) \quad p_\sigma(\varphi) = \int_H \sigma \otimes L(h)(\varphi) d_l h,$$

where  $d_l h = d_r h^{-1}$  is a left-invariant Haar measure on  $H$ . Note that we can also write

$$(A.1.15.3) \quad d_l h = \delta_H(h)^{-1} d_r h,$$

where  $\delta_H$  is the modular function of  $H$  [Gaal, Loom, Weil4]. From formulas (A.1.15.2) and (A.1.15.3), it is a simple matter to check that

$$(A.1.15.4a) \quad \sigma \otimes L(h)(p_\sigma(\varphi)) = p_\sigma(f), \quad h \in H, \quad \varphi \in C_c^\infty(G; U),$$

$$(A.1.15.4b) \quad p_\sigma(\sigma \otimes L(h)(\varphi)) = \delta_H(h) p_\sigma(\varphi).$$

Since it is clear that  $p_\sigma(\varphi)$  will have support contained in  $H \text{ supp}(\varphi)$ , we see that formula (A.1.15.4a) is just the statement that  $p_\sigma(\varphi) \in C_c^\infty(H \backslash G; \sigma)$ . We conclude that  $p_\sigma$  defines a linear map

$$(A.1.15.5) \quad p_\sigma : C_c^\infty(G; U) \rightarrow C_c^\infty(H \backslash G; \sigma).$$

Both these spaces have standard topologies, and it is easy to check that  $p_\sigma$  is a continuous map. The results of Dixmier and Malliavin [DiMa] imply that if  $\sigma$  is smooth and  $U$  is Fréchet, then the map  $p_\sigma$  is surjective. In particular, if  $\sigma$  is finite dimensional, then  $p_\sigma$  is surjective (but this is elementary).

Regarding the kernel of  $p_\sigma$ , the formula (A.1.15.4b) makes it clear that functions of the form

$$(\sigma \otimes L)(h)(\varphi) - \delta_H(h)(\varphi), \quad \varphi \in C_c^\infty(G; U),$$

are sent to zero by  $p_\sigma$ . If we let  $h$  vary in a one-parameter group and differentiate at the origin, we may conclude that

$$(A.1.15.6) \quad (\sigma(x) + L(x) - \delta_H(x))\varphi \in \ker p_\sigma$$

for all  $\varphi \in C_c^\infty(G; U)$  and  $x \in \text{Lie}(H)$ . (Here we are following the notation of (A.1.13.1) for the action of  $\text{Lie}(H)$ .)

**LEMMA A.1.15.7.** *If  $\sigma$  is finite-dimensional, then every element of  $\ker p_\sigma$  is a finite sum of functions of the form (A.1.15.6).*

This lemma is essentially a generalization of the Poincaré Lemma [Gold, Ster], for forms of top degree.

Let  $\lambda$  be a linear functional on  $C_c^\infty(H \backslash G; \sigma)$ . Then  $\lambda \circ p_\sigma$  is a linear functional on  $C_c^\infty(G; U)$ —a “ $U^*$ -valued distribution.” Using formulas (A.1.15.4) and (A.1.10.1) we can compute that

$$\begin{aligned} (\sigma \otimes L)^*(h)((\lambda \circ p_\sigma)(\varphi)) &= (\lambda \circ p_\sigma)((\sigma \otimes L)(h)^{-1}\varphi) \\ &= \lambda(p_\sigma((\sigma \otimes L)(h)^{-1}(\varphi))) \\ &= \lambda(\delta_H(h)^{-1}p_\sigma(\varphi)) = \delta_H(h)^{-1}\lambda \circ p_\sigma(\varphi), \quad h \in H, \varphi \in C_c^\infty(G; U). \end{aligned}$$

In other words, the distribution  $\mu = \lambda \circ p_\sigma$  satisfies the transformation law

$$(A.1.15.8) \quad \delta_H(h)(\sigma \otimes L)^*(h)(\mu) = \mu.$$

Suppose on the other hand that we have a functional  $\mu$  on  $C_c^\infty(G; U)$  which satisfies the transformation law (A.1.15.8). By a differentiation, we conclude  $\mu$  vanishes on functions of the form (A.1.15.6). Thus the following statement follows directly from Lemma A.1.15.7.

**COROLLARY A.1.15.9.** *Suppose  $\sigma$  is a smooth representation on a Fréchet space  $U$ . Then any distribution  $\mu$  on  $C_c^\infty(G; U)$  satisfying the transformation law (A.1.15.8) is of the form  $\mu = \lambda \circ p_\sigma$  for suitable  $\lambda \in C_c^\infty(H \backslash G; \sigma)$ .*

This corollary is the basic fact behind various Frobenius reciprocity statements (see [Gaal, Knap2, Warn], etc.).

As an example of Corollary A.1.15.9, consider right-invariant Haar measure  $d_r g$  on  $G$ . Using formulas (A.1.15.3) and (A.1.10.1), we compute that

$$(A.1.15.10) \quad L^*(g_1)(d_r g) = \delta_G(g_1)^{-1}d_r g, \quad g_1 \in G.$$

From Corollary A.1.15.9, we conclude that the measure  $d_r g$  factors to the space  $C_c^\infty(H \backslash G; \delta_H/\delta_G)$ ; in other words, there is a right  $G$ -invariant functional on this space. (Note that if  $\delta_H = \delta_{G|H}$ , this amounts to the statement that  $H \backslash G$  carries a  $G$ -invariant measure [Gaal, Weil4, FeDo], etc.)

**A.1.16.** Isometric Banach space representations, especially unitary representations, are of particular interest, so we want to know how they fare with respect to induction. For this, it is convenient to deal with a class of representations slightly more general than isometric ones. Let us call a representation

$\rho$  of  $G$  on a Banach space  $V$  *quasi-isometric* if it satisfies the following mutually equivalent conditions

- (A.1.16.1) (a)  $\rho(g)$  is a scalar multiple of an isometry of  $V$  for all  $g$  in  $G$ ,  
 (b)  $\|\sigma(g)\| \|\sigma(g)^{-1}\| = 1$ . Here  $\|\cdot\|$  denotes the operator norm in  $\text{End } V$ .  
 (c)  $\sigma(g) = \alpha(g)\sigma_1(g)$ , where  $\sigma_1$  is an isometric representation of  $G$  on  $V$ , and  $\alpha: G \rightarrow \mathbf{R}^{+\times}$  is a continuous homomorphism, i.e., a real-valued quasicharacter on  $G$ .

Note that, in situation (A.1.16.1)(c), we have

$$(A.1.16.2) \quad \alpha(g) = \|\sigma(g)\|.$$

Thus  $\|\sigma(g)\|$  is a quasicharacter; we call it the *dilation character* of  $\rho$ .

If  $\rho$  is quasi-isometric, and  $\beta: G \rightarrow \mathbf{C}^\times$  is a quasicharacter of  $G$ , then  $\beta \otimes \rho$ , defined by

$$(A.1.16.3a) \quad \beta \otimes \rho(g)(v) = \beta(g)\rho(g)(v),$$

is also quasi-isometric, with dilation character

$$(A.1.16.3b) \quad \|\beta \otimes \rho\| = |\beta| \|\rho\|,$$

where  $|\beta|$  denotes the absolute value of  $\beta$ .

Let  $H \subseteq G$  be a closed subgroup, and let  $\sigma$  be a quasi-isometric representation of  $H$  on a Banach space  $U$ . Since  $\sigma$  is quasi-isometric, we see that for any  $f \in C_c^\infty(H \backslash G; \sigma)$ , the function  $\|f\|$ , defined by

$$(A.1.16.4) \quad \|f\|(g) = \|f(g)\|,$$

where  $\|\cdot\|$  is the norm on  $U$ , will belong to  $C_c(H \backslash G; \|\sigma\|)$ . Suppose that  $\|\sigma\| = (\delta_H/\delta_G)^{1/p}$  for some real number  $p \geq 1$ . Then  $\|f\|^p \in C_c(H \backslash G; \delta_H/\delta_G)$ . According to the final remark of §A.1.15, there is a right-invariant functional on  $C_c(H \backslash G; \delta_H/\delta_G)$ , a projection of right-invariant Haar measure on  $G$ . Let us denote it by  $\varphi \rightarrow \int_{H \backslash G} \varphi \, d\dot{g}$ . Then the recipe

$$(A.1.16.5) \quad \|f\|_p = \left( \int_{H \backslash G} \|f\|^p \, d\dot{g} \right)^{1/p}$$

defines a  $G$ -invariant norm on  $C_c^\infty(H \backslash G; \sigma)$ , which may be completed to define an isometric representation of  $G$ .

The above construction involves a particular assumption on the dilation character  $\|\sigma\|$  of the representation  $\sigma$ . However, we see that for any quasi-isometric representation  $\sigma$  of  $H$ , the representation  $((\delta_H/\delta_G)^{1/p} \|\sigma\|^{-1}) \otimes \sigma$  will satisfy the assumption. Thus we have

**PROPOSITION A.1.16.6.** *If  $\sigma$  is a quasi-isometric representation of the closed subgroup  $H$  on the Banach space  $U$ , then formula (A.1.16.5) defines a  $G$ -invariant norm on  $C_c^\infty(H \backslash G; ((\delta_H/\delta_G)^{1/p} \|\sigma\|^{-1}) \otimes \sigma)$ .*

The isometric representation of  $G$  that results from completing the norm of Proposition A.1.16.6 will be denoted

$$(A.1.16.7) \quad p - \text{ind}_H^G \sigma$$

and will be called the  $p$ -normalized induced representation derived from  $\sigma$ .

Finally, suppose that  $\sigma$  is a unitary representation of  $H$ . Let  $(\cdot, \cdot)$  denote the  $H$ -invariant inner product which defines the norm on  $U$ . Then if  $f_1, f_2$  are in  $C_c^\infty(H \backslash G; (\delta_G/\delta_H)^{1/2} \otimes \sigma)$ , we see that the inner product

$$(f_1, f_2)(g) = (f_1(g), f_2(g))$$

will be a scalar-valued function in  $C_c^\infty(H \backslash G; \delta_G/\delta_H)$ . Hence the recipe

$$(A.1.16.8) \quad (f_1, f_2)^\sim = \int_{H \backslash G} (f_1, f_2) dg$$

defines a  $G$ -invariant inner product on  $C_c^\infty(H \backslash G; (\delta_G/\delta_H)^{1/2} \otimes \sigma)$ . Further, the norm defined by this inner product is clearly the norm attached to  $2 - \text{ind}_H^G \sigma$ . In summary, we have shown

**COROLLARY A.1.16.9.** *If  $\sigma$  is a unitary representation of  $H$ , then  $2 - \text{ind}_H^G \sigma$  is a unitary representation of  $G$ , with invariant inner product defined by (A.1.16.8).*

**A.1.17.** Let  $V$  be a vector space and  $V^*$  its dual. Given  $v \in V$  and  $\lambda \in V^*$ , one can form the *dyad*  $E_{v, \lambda}$ , which is an operator on  $V$ , by the formula

$$(A.1.17.1) \quad E_{v, \lambda}(x) = \lambda(x)v, \quad x \in V.$$

The bilinear map  $(v, \lambda) \rightarrow E_{v, \lambda}$  extends to an embedding

$$(A.1.17.2) \quad E: V \otimes V^* \hookrightarrow \text{End}(V)$$

of the algebraic tensor product into the algebra of linear transformations on  $V$ . The image  $E(V \otimes V^*)$  consists of all operators of finite rank on  $V$ . If  $V$  is a locally convex topological vector space and  $V^*$  is the topological dual of continuous linear functionals on  $V$ , then the image of  $E$  is the algebra of continuous finite rank operators on  $V$ .

On  $V \otimes V^*$  there is defined a canonical linear functional induced by the canonical bilinear pairing between  $V$  and  $V^*$ . When considered as a function of operators, this functional is called the *trace* and is denoted by  $\text{tr}$ . We have the formula

$$(A.1.17.3) \quad \text{tr} \left( \sum E_{v_i, \lambda_i} \right) = \sum \lambda_i(v_i).$$

Suppose  $V$  is a Banach space. We can define a norm  $||| \cdot |||_1$  on  $E(V \otimes V^*)$  by the rule

$$(A.1.17.4) \quad |||T|||_1 = \inf \left\{ \sum_i \|v_i\| \|\lambda_i\| : \sum E_{v_i, \lambda_i} = T \right\}.$$

We call this the *trace norm* on  $E(V \otimes V^*)$ . It is easy to check that the trace norm dominates the usual operator norm on  $E(V \otimes V^*)$ . Hence Cauchy sequences with respect to  $||| \cdot |||_1$  in  $E(V \otimes V^*)$  will also be Cauchy with respect to  $\| \cdot \|$ , so the Banach space completion of  $E(V \otimes V^*)$  can be regarded as a certain space of operators on  $V$ . We call it  $\mathcal{T}(V)$ , the space of *trace class* or *nuclear operators* on  $V$  (cf. [CoGr, Gaal, Lang1, 2], etc.). In fact,  $\mathcal{T}(V)$  is a two-sided ideal in  $\text{End}(V)$ , and one can easily check that

$$(A.1.17.5) \quad |||ATB|||_1 \leq |||A||| |||B||| |||T|||_1, \quad A, B \in \text{End}(V), T \in \mathcal{T}(V).$$

Here  $||| \cdot |||$  denotes the usual operator norm on  $\text{End}(V)$ .

Since the trace linear functional on  $E(V \otimes V^*)$  is clearly dominated by the trace norm, it extends by continuity to a linear function, still called the trace, on  $\mathcal{T}(V)$ .

A.1.18. Let  $\rho$  be a representation of  $G$  on a Banach space  $V$ . It may happen that, for some element  $X$  in the universal enveloping algebra  $\mathcal{U}(\text{Lie}(G))$  (see the Remark at the end of §A.1.13) the operator  $\rho(X)^{-1}$  is trace class. By this we mean

- (A.1.18.1) (i) As an operator on  $V^\infty$ ,  $\rho(X)$  is invertible.  
(ii) The inverse operator  $\rho(X)^{-1}$  extends to a continuous operator on  $V$ .  
(iii) This continuous extension is in the space  $\mathcal{T}(V)$  of trace class operators on  $V$ .

If this happens, we call  $\rho$  a *strongly trace class representation*. Many Lie groups, including all nilpotent groups and all semisimple groups (with a finite number of connected components), have all their irreducible representations strongly trace class [CoGr, Warn].

Under the conditions of the previous paragraph, consider  $f \in C_c^\infty(G)$ . We can write

$$\rho(f) = \rho(X)^{-1} \rho(X) \rho(f) = \rho(X)^{-1} \rho(L(X)(f)).$$

We conclude that  $\rho(f)$  is trace class; further,  $f \rightarrow \rho(f)$  is continuous from  $C_c^\infty(G)$  to  $\mathcal{T}(V)$ . In particular, the functional

$$(A.1.18.2) \quad \theta_\rho(f) = \text{tr } \rho(f)$$

is a distribution on  $G$ . We call  $\theta_\rho$  the *distributional character*, or simply the character, of  $\rho$ .



A.1.19. Let  $K$  be a compact group, and let  $\rho$  be a representation of  $K$  on a vector space  $V$ . A vector  $v$  in  $V$  is called  $K$ -finite if the span of  $\{\rho(k)v, k \in K\}$ , the  $K$ -transforms of  $v$ , is finite dimensional. The set of all  $K$ -finite vectors in  $V$  is a subspace of  $V$ , denoted  $V_K$ . An argument using approximate identities and the Peter-Weyl Theorem shows that  $V_K$  is dense in  $V$  (cf. [Lang1, Warn, Knap2], etc.).

Let  $\sigma \in \widehat{K}$  be an (isomorphism class of) irreducible representation(s) of  $K$ . A vector  $v$  in  $V$  is of type  $\sigma$  if the span of the  $K$ -transforms is an irreducible representation isomorphic to  $\sigma$ . The  $\sigma$ -isotypic component of  $V$  is the span of all vectors of type  $\sigma$ . It is denoted  $V_\sigma$ . The Peter-Weyl Theorem (cf. Theorem 3.5.4.23) provides a function  $e_\sigma$  such that  $\rho(e_\sigma)$  is a projection from  $V$  to  $V_\sigma$ . It follows that every vector in  $V_\sigma$  is  $K$ -finite, and any finite-dimensional,  $K$ -invariant subspace of  $V_\sigma$  is isomorphic to a direct sum of copies of  $\sigma$ . The ratio (perhaps infinite)  $(\dim V_\sigma)/\dim \sigma$  is called the *multiplicity* of  $\sigma$  in  $V$ .

We have

$$V_K = \sum_{\sigma \in \widehat{K}} V_\sigma.$$

This may be considered simply as the algebraic direct sum of the spaces  $V_\sigma$ , or, more elaborately, may be considered as a topological vector space which is the inductive limit over  $\sum_{\sigma \in F} V_\sigma$ , for finite sets  $F \subseteq \widehat{K}$ . Each  $\sum_{\sigma \in F} V_\sigma$  is given its topology as a subspace of  $V$ .

A.1.20. Let  $K$  be a compact subgroup of the Lie group  $G$ . Let  $\rho$  be a representation of  $G$  on a vector space  $V$ . Let  $V_K$  be the subspace of  $K$ -finite vectors (cf. §A.1.19). Since  $\text{Lie}(G)$  is a finite-dimensional  $K$ -module under the adjoint action, and the action of  $\text{Lie}(G)$  on  $V^\infty$  (cf. §A.1.13) can be expressed in terms of a map

$$\text{Lie}(G) \otimes V^\infty \rightarrow V^\infty, \quad x \otimes v \rightarrow \rho(x)v,$$

one can check that  $V_K^\infty$  is invariant under  $\mathcal{U}(\text{Lie}(G))$ , the universal enveloping algebra (cf. §A.1.13). Thus  $V_K^\infty$  is a module for  $K$  and for  $\mathcal{U}(\text{Lie}(G))$ .

Let  $\rho, \rho'$  be representations of  $G$  on spaces  $V, V'$ . We say  $V, V'$  are *infinitesimally equivalent* if there is a linear isomorphism

$$T: V_K^\infty \rightarrow V_K'^\infty$$

which intertwines the actions of  $K$  and of  $\mathcal{U}(\text{Lie}(G))$  on these two spaces. The situation in which this notion is most often used is when  $G$  is semisimple,  $K$  is a maximal compact subgroup, and  $V, V'$  are irreducible (t.c.i. Banach) representations of  $G$ . In these circumstances, the  $K$ -isotypic components  $V_\sigma$  and  $V'_\sigma$ ,  $\sigma \in \widehat{K}$ , are finite dimensional, by an early result of Harish-Chandra (cf. [HaCh3, Gode2, Warn, Knap2], etc.). In this case  $V_K^\infty = V_K$  constitutes an algebraic skeleton around which  $V$  is built, by means of completion with respect to some topology. Infinitesimal equivalence throws away

the fuzz introduced with the topology, and considers only the algebraic core. At the current stage of semisimple harmonic analysis, which is still primarily concerned with individual irreducible representations, this is a very useful notion of equivalence.

## Appendix 2: Structure of real semisimple Lie algebras and Lie groups.

A.2.1. *Cartan involution and invariant form.* We follow [Knap2] and [Wall2] by defining a reductive Lie group to be a closed subgroup  $G$  of  $GL_n(\mathbf{R})$ , for some  $n$ , which is left-invariant (as a set) by the “Cartan involution”

$$(A.2.1.1) \quad \theta : g \rightarrow (g^t)^{-1}.$$

Here  $g^t$  indicates the transpose of  $g \in GL_n(\mathbf{R})$ . This definition allows one to short-circuit a lot of preliminary material and get to the essential facts fairly quickly.

The “infinitesimal automorphism” of  $\mathfrak{gl}_n(\mathbf{R}) \simeq M_n(\mathbf{R})$  corresponding to  $\theta$  as defined in (A.2.1.1) is

$$(A.2.1.2) \quad \theta(x) = -x^t;$$

also known as the Cartan involution. Since it is of order two, its eigenvalues are simply  $\pm 1$ . The space of matrices fixed by  $\theta$  is exactly the space of skew-symmetric matrices; this is also the Lie algebra  $\mathfrak{o}_n$  of the orthogonal group, the isometry group of the standard inner product on  $\mathbf{R}^n$ :

$$(A.2.1.3) \quad \mathfrak{o}_n = \{x \in M_n(\mathbf{R}) : x + x^t = 0\} = \{x \in M_n(\mathbf{R}) : \theta(x) = x\}.$$

The  $-1$  eigenspace of  $\theta$  is the space  $\mathfrak{s}$  of symmetric matrices

$$(A.2.1.4) \quad \mathfrak{s} = \{y \in M_n(\mathbf{R}) : y = y^t\}.$$

We have the direct sum decomposition

$$(A.2.1.5) \quad \mathfrak{gl}_n(\mathbf{R}) \simeq \mathfrak{o}_n \oplus \mathfrak{s}.$$

The summand  $\mathfrak{o}_n$  is a Lie algebra, the Lie algebra of the compact group  $O_n \subseteq GL_n(\mathbf{R})$ , but the summand  $\mathfrak{s}$  is not at all a Lie algebra: its commutators belong to  $\mathfrak{o}_n$ . However, it is invariant under commutators from  $\mathfrak{o}_n$ . In sum, we have the relations

$$(A.2.1.6) \quad [\mathfrak{o}_n, \mathfrak{o}_n] \subseteq \mathfrak{o}_n, \quad [\mathfrak{o}_n, \mathfrak{s}] \subseteq \mathfrak{s}, \quad [\mathfrak{s}, \mathfrak{s}] \subseteq \mathfrak{o}_n.$$

Here  $[\mathfrak{o}_n, \mathfrak{o}_n]$  is interpreted as the linear span of commutators  $[x, y]$ ,  $x, y \in \mathfrak{o}_n$ , and similarly for the other expressions.

We consider on  $\mathfrak{gl}_n(\mathbf{R})$  the bilinear form

$$(A.2.1.7) \quad B(x, y) = \text{tr}(xy), \quad x, y \in M_n(\mathbf{R}),$$

where  $\text{tr}$  indicates the trace function on matrices. We observe that  $B(x, y)$  is invariant under conjugation:

$$(A.2.1.8) \quad \begin{aligned} B(\text{Ad } g(x), \text{Ad } g(y)) &= \text{tr}((gxg^{-1})(gyg^{-1})) \\ &= \text{tr}(gxyg^{-1}) = \text{tr}(xy) = B(x, y). \end{aligned}$$

The infinitesimal version of this is

$$(A.2.1.9) \quad B(\operatorname{ad} x(y), z) + B(x, \operatorname{ad} y(z)) = 0.$$

We observe that the summands  $\mathfrak{o}_n$  and  $\mathfrak{s}$  of decomposition (A.2.1.5) are orthogonal with respect to  $B$ . Furthermore, it is easy to check that  $B$  is negative definite on  $\mathfrak{o}_n$ , and positive definite on  $\mathfrak{s}$ .

Now consider a Lie algebra  $\mathfrak{g} \subseteq \mathfrak{gl}_n(\mathbf{R})$ . Assume that  $\mathfrak{g}$  is invariant under the Cartan involution  $\theta$ , given in equation (A.2.1.2). We will refer to  $\theta|_{\mathfrak{g}}$  as the Cartan involution for  $\mathfrak{g}$ .

Since  $\mathfrak{g}$  is invariant under  $\theta$ , we see that if we set

$$\mathfrak{k} = \mathfrak{o}_n \cap \mathfrak{g}, \quad \mathfrak{p} = \mathfrak{s} \cap \mathfrak{g}$$

then we will have the direct sum decomposition

$$(A.2.1.10) \quad \mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}.$$

The analogs of relations (A.2.1.6) will clearly hold for  $\mathfrak{k}$  and  $\mathfrak{p}$ . In particular,  $\mathfrak{k}$  is a Lie subalgebra of  $\mathfrak{g}$ , and  $\mathfrak{p}$  is a  $\mathfrak{k}$ -module via the adjoint action. Also,  $\mathfrak{k}$  and  $\mathfrak{p}$  are orthogonal with respect to the restriction of the inner product  $B$  of formula (A.2.1.7) to  $\mathfrak{g}$ , and  $B|_{\mathfrak{k}}$  is negative definite while  $B|_{\mathfrak{p}}$  is positive definite. It follows that  $B|_{\mathfrak{g}}$  is nondegenerate. We may also conclude that if  $K = \exp \mathfrak{k}$  is the connected subgroup of  $GL_n(\mathbf{R})$  with Lie algebra  $\mathfrak{k}$ , then  $K$  has compact closure (since it will be contained in  $O_n$ ). Furthermore,  $\mathfrak{k}$  is a maximal subalgebra of  $\mathfrak{g}$  with the property that it generates a bounded group; for if  $\exp tx$ ,  $x \in \mathfrak{gl}_n(\mathbf{R})$ , is a bounded one-parameter subgroup, the eigenvalues of  $\exp tx$  must be of absolute value 1, hence the eigenvalues of  $x$  must be pure imaginary, hence  $B(x, x) \leq 0$ . However, any subspace of  $\mathfrak{g}$  strictly containing  $\mathfrak{k}$  must have nontrivial intersection with  $\mathfrak{p}$ , on which  $B$  is positive definite.

**A.2.2. Split Cartan subalgebras and restricted roots.** Elements of  $\mathfrak{p}$ , being symmetric matrices, are diagonalizable with real eigenvalues, hence the same is true for their action on  $\mathfrak{g}$  by  $\operatorname{ad}$ . For  $x \in \mathfrak{p}$ , let  $\mathfrak{c}_{\mathfrak{g}}(x) = \ker \operatorname{ad} x|_{\mathfrak{g}}$  be the centralizer of  $x$  in  $\mathfrak{g}$ . It is easy to check that  $\mathfrak{c}_{\mathfrak{g}}(x)$  is a Lie subalgebra of  $\mathfrak{g}$ , and is stable under the Cartan involution. Thus  $\mathfrak{c}_{\mathfrak{g}}(x) = (\mathfrak{c}_{\mathfrak{g}}(x) \cap \mathfrak{k}) \oplus (\mathfrak{c}_{\mathfrak{g}}(x) \cap \mathfrak{p})$ . Clearly,  $x \in \mathfrak{c}_{\mathfrak{g}}(x) \cap \mathfrak{p}$ , and is central in  $\mathfrak{c}_{\mathfrak{g}}(x)$ . If  $\mathfrak{c}_{\mathfrak{g}}(x) \cap \mathfrak{p}$  is not central in  $\mathfrak{c}_{\mathfrak{g}}(x)$ , we can choose another element,  $x_2$ , in  $\mathfrak{c}_{\mathfrak{g}}(x)$  and look at its centralizer. Continuing in this way, we will arrive at a subalgebra

$$(A.2.2.1) \quad \mathfrak{m} \oplus \mathfrak{a}$$

such that

- (i)  $\mathfrak{m} \subseteq \mathfrak{k}$ ,  $\mathfrak{a} \subseteq \mathfrak{p}$ ,
- (ii)  $\mathfrak{m} \oplus \mathfrak{a}$  is the centralizer of  $\mathfrak{a}$  in  $\mathfrak{g}$ .

Because of the analogy between this construction and the construction of Cartan subalgebras for general Lie algebras, the abelian algebra  $\mathfrak{a}$  is called a *split Cartan subalgebra* of  $\mathfrak{g}$ .

Let  $\mathfrak{a}$  be a split Cartan subalgebra of  $\mathfrak{g}$ , as in (A.2.2.1), and consider the adjoint action of  $\mathfrak{a}$  on  $\mathfrak{g}$ . Since elements of  $\mathfrak{a}$  are individually diagonalizable over  $\mathbb{R}$ , and since they commute, we can decompose  $\mathfrak{g}$  into simultaneous eigenspaces for  $\text{ad } \mathfrak{a}$ : If  $x$  is a simultaneous eigenvector for all  $\text{ad } a$ , then

$$\text{ad } a(x) = \alpha(a)x,$$

where  $\alpha(a)$  is the appropriate eigenvalue. Clearly  $\alpha(a)$  depends linearly on  $a$ ; that is,  $\alpha$  belongs to  $\mathfrak{a}^*$ , the dual of  $\mathfrak{a}$ . Let  $\Delta$  denote the set of nonzero elements of  $\mathfrak{a}^*$  which arise as the simultaneous eigenvalue function for some  $x \in \mathfrak{g}$ . In parallel with the general situation, described in §2, the elements of  $\Delta$  are called the *restricted roots*, or simply roots, of  $\mathfrak{g}$ . We have a direct sum decomposition

$$(A.2.2.2) \quad \mathfrak{g} = \mathfrak{g}_0 \oplus \sum_{\alpha \in \Delta} \mathfrak{g}_{\alpha},$$

where

- (i)  $\mathfrak{g}_0 = \mathfrak{m} \oplus \mathfrak{a}$ ,
- (ii)  $\mathfrak{g}_{\alpha} = \{x : \text{ad } a(x) = \alpha(a)x, \text{ all } a \in \mathfrak{a}\}$ ,  $\alpha \in \Delta$ .

The  $\mathfrak{g}_{\alpha}$  are called the *root spaces* for  $\alpha$ .

The Cartan involution  $\theta$  normalizes  $\mathfrak{a}$ , so it will preserve the decomposition (A.2.2.2). In fact, it is easy to check that, if  $x \in \mathfrak{g}_{\alpha}$  and  $a \in \mathfrak{a}$ , then

$$\text{ad } a(\theta(x)) = \theta(\text{ad } \theta(a)(x)) = \theta(\alpha(-a)x) = -\alpha(a)\theta(x).$$

Thus

$$(A.2.2.3) \quad \theta(\mathfrak{g}_{\alpha}) = \mathfrak{g}_{-\alpha}.$$

In particular,

$$(A.2.2.4) \quad \mathfrak{g}_{\alpha} \oplus \mathfrak{g}_{-\alpha} = (\mathfrak{g}_{\alpha} \oplus \mathfrak{g}_{-\alpha}) \cap \mathfrak{k} \oplus (\mathfrak{g}_{\alpha} \oplus \mathfrak{g}_{-\alpha}) \cap \mathfrak{p}.$$

This decomposition allows us to show, by an argument analogous to the proof of uniqueness of Cartan subalgebras for complex Lie algebras, that the split Cartan subalgebra  $\mathfrak{a}$  of (A.2.2.1) is essentially unique.

LEMMA A.2.2.5.  $\text{Ad } K(\mathfrak{a}) = \mathfrak{p}$ .

This can be proved by the following steps.

(i) Define an element  $x \in \mathfrak{p}$  to be *regular* if its centralizer in  $\mathfrak{p}$  is of minimal dimension, equivalently if it is abelian, equivalently if it is a split Cartan subalgebra. Denote the set of regular elements in  $\mathfrak{p}$  by  $\mathfrak{p}_{\text{reg}}$ . It is open and dense in  $\mathfrak{p}$ .

(ii) An easy computation of the derivative of  $\text{Ad } K$  at  $x \in \mathfrak{a} \cap \mathfrak{p}_{\text{reg}}$ , using equation (A.2.2.4) shows that  $\text{Ad } K(\mathfrak{a} \cap \mathfrak{p}_{\text{reg}})$  is open in  $\mathfrak{p}_{\text{reg}}$ . Since the same reasoning applies to any putative other split Cartan subalgebra  $\tilde{\mathfrak{a}}$ , it follows that  $\text{Ad } K(\mathfrak{a} \cap \mathfrak{p}_{\text{reg}})$  is also closed in  $\mathfrak{p}_{\text{reg}}$ .

(iii) A computation similar to that of (ii), but at a nonregular element  $x \in \mathfrak{a}$ , shows that  $\mathfrak{p} - \mathfrak{p}_{\text{reg}}$  has codimension at least 2 in  $\mathfrak{p}$ . Hence  $\mathfrak{p}_{\text{reg}}$  is connected, whence from (ii),  $\text{Ad } K(\mathfrak{a} \cap \mathfrak{p}_{\text{reg}}) = \mathfrak{p}_{\text{reg}}$ . Taking closures gives the lemma.

Thus, in analogy with the decomposition (2.8.6) of a general complex Lie algebra with respect to a Cartan subalgebra, one finds the decomposition (A.2.2.2) of  $\mathfrak{g}$  into root spaces for  $\mathfrak{a}$  is essentially canonical (i.e., is unique up to conjugation). Further, one can find, in a similar fashion to §2.8, copies of  $\mathfrak{sl}_2$  (actually,  $\mathfrak{sl}_2(\mathbf{R})$ , the real split form) in  $\mathfrak{g}$ . Precisely, take a nonzero  $x$  in  $\mathfrak{g}_\alpha$ . A simple calculation shows that the commutator  $h_\alpha = [x, \theta(x)]$  belongs to  $\mathfrak{a}$ . (Additivity of roots for commutators, as in §2.8, shows that  $h \in \mathfrak{g}_0$ , and it is easy to check that  $\theta(h_\alpha) = -h_\alpha$ .) If  $\alpha(h_\alpha) \neq 0$ , then one can scale  $x$  to get  $\alpha(h_\alpha) = 2$ , whence  $x, \theta(x)$ , and  $h_\alpha$  form a standard basis for  $\mathfrak{sl}_2$ . Further, the possibility that  $\alpha(h_\alpha) = 0$  can be shown to contradict the fact that  $B(x, \theta(x)) < 0$ . (Note the analogy with the Cartan criterion argument, cf. §2.8.)

Thus one gets a copy of  $\mathfrak{sl}_2(\mathbf{R})$  inside  $\mathfrak{g}$  for any root  $\alpha$  of  $\mathfrak{a}$ . The corresponding copy of  $\text{SL}_2(\mathbf{R})$ , obtained by exponentiation, will contain elements which normalize  $\mathfrak{a}$ : these elements will induce a reflection in the hyperplane orthogonal to  $\alpha$ . One concludes that the restricted roots of  $\mathfrak{a}$  in  $\mathfrak{g}$  form a root system in the formal sense [Bour, Crtr, Hump, Serr1] with Weyl group generated by reflections in the hyperplanes orthogonal to the roots (the root hyperplanes). These hyperplanes divide  $\mathfrak{a}$  into a collection of convex cones, the Weyl chambers. These are the closures of connected components of  $\mathfrak{a} \cap \mathfrak{p}_{\text{reg}}$ . The Weyl chambers are permuted simply transitively by  $W$ .

**A.2.3. Decompositions of  $G$ .** Associated to the decompositions (A.2.1.10) and (A.2.2.2) are decompositions of  $G$ , the connected group with Lie algebra  $\mathfrak{g}$ . Set

$$K = G \cap O_n = \{k \in G : \theta(k) = k\}.$$

Clearly, the Lie algebra of  $K$  is the  $\mathfrak{k}$  of equation (A.2.1.10). Let  $\mathfrak{p}$  be the other summand in (A.2.1.10). Let  $\exp$  be the exponential map from  $\mathfrak{gl}_n(\mathbf{R})$  to  $\text{GL}_n(\mathbf{R})$ .

(A.2.3.1) (Cartan decomposition, I). The mapping

$$K \times \mathfrak{p} \rightarrow G, \quad (k, x) \rightarrow k \exp x$$

is a diffeomorphism. Thus each element  $g$  in  $G$  has a unique factorization  $g = k \exp x$  with  $k \in K, x \in \mathfrak{p}$ .

We use the shorthand  $G = K \exp \mathfrak{p}$  to indicate the state of affairs described in (A.2.3.1).

**REMARK.** The Cartan decomposition shows that, as topological space,  $G \sim K \times \mathbf{R}^m$ , where  $m = \dim \mathfrak{p}$ . Hence all interesting topology of  $G$ , including Betti numbers, homotopy groups, etc., is determined by  $K$ . Thus  $K$  is connected since we have assumed  $G$  is connected; also, the argument above

did not require  $G$  to be closed, so  $G$  is closed in  $\mathrm{GL}_n(\mathbf{R})$  if and only if  $K$  is.

To prove (A.2.3.1), one takes  $g$  in  $G$  and considers  $g^t g$ . This is a selfadjoint positive-definite matrix, so it has a unique selfadjoint logarithm  $y \in \mathfrak{s} \subseteq \mathfrak{gl}_n(\mathbf{R})$ :  $g^t g = \exp y$ . Suppose that  $y$  is in  $\mathfrak{p}$ . Then  $\exp(y/2) = (g^t g)^{1/2}$ , and one checks that  $k = g(g^t g)^{-1/2}$  is in  $K$ , so the desired factorization is  $g = (g \exp(-y/2)) \exp(y/2)$ .

The above argument shows that the desired factorization certainly exists in  $\mathrm{GL}_n(\mathbf{R})$  (where it is also known as the polar decomposition or principal value factorization [Stra, Lang2, Gaal]. To show that if  $g$  is in  $G$ , and  $g = k \exp x$  is its Cartan decomposition in  $\mathrm{GL}_n(\mathbf{R})$ , then  $x \in \mathfrak{p}$ , argue as follows. The map  $g \rightarrow x$  is analytic. For  $g$  near the identity, the Inverse Function Theorem implies that  $x \in \mathfrak{p}$ . Since  $G$  is connected, we always have  $x \in \mathfrak{p}$ .

REMARK. The Cartan decomposition can be extended to nonconnected subgroups satisfying appropriate conditions (such as being algebraic) [Knap2, Wall2].

If we combine decomposition (A.2.3.1) with Lemma A.2.2.5, we get another useful decomposition of  $G$ . Let  $\mathfrak{a} \subseteq \mathfrak{p}$  be a split Cartan subalgebra. Let  $A = \exp \mathfrak{a}$  be the connected abelian group with Lie algebra  $\mathfrak{a}$ . Choose a Weyl chamber (cf. §A.2.2) in  $\mathfrak{a}$ . Denote it by  $\mathfrak{a}^+$ , and set  $A^+ = \exp \mathfrak{a}^+$ .

A.2.3.2 (Cartan decomposition, II). The mapping

$$K \times A^+ \times K \rightarrow G, \quad (k_1, a, k_2) \rightarrow k_1 a k_2$$

is surjective.

We indicate this result by writing  $G = KA^+K$ , or just  $G = KAK$ .

This decomposition results from combining (A.2.3.1) with Lemma A.2.2.5. If  $g = k \exp x$ , and  $x = \mathrm{Ad} \tilde{k}(y)$  for  $y \in \mathfrak{a}^+$ , then  $g = k \tilde{k} \exp(a) \tilde{k}^{-1}$ , which is decomposition (A.2.3.2) with  $k_1 = k \tilde{k}$  and  $k_2 = \tilde{k}^{-1}$ .

Next consider decompositions of  $G$  associated to the root space decomposition (A.2.2.2). Let  $\mathfrak{a}^+$  be a Weyl chamber in  $\mathfrak{a}$ . By definition,  $\mathfrak{a}^+$  is a set where each root of  $\mathfrak{a}$  in  $\mathfrak{g}$  takes values of only one sign (i.e., either all nonnegative or all nonpositive). Let  $\Delta^+$  denote the set of roots which are positive on  $\mathfrak{a}^+$ . Set

$$(A.2.3.3) \quad \mathfrak{n}^+ = \sum_{\alpha \in \Delta^+} \mathfrak{g}_\alpha.$$

Combining equations (A.2.2.2), (A.2.2.4), and definition (A.2.3.3) gives us the equation

$$(A.2.3.4) \quad \mathfrak{g} = \mathfrak{k} \oplus \mathfrak{a} \oplus \mathfrak{n}^+.$$

This is the infinitesimal version of the *Iwasawa decomposition*.

(A.2.3.5) (Iwasawa decomposition). Let  $N^+ \subseteq G$  be the connected subgroup whose Lie algebra is  $\mathfrak{n}^+$ . Then  $N^+$  is closed in  $G$  and

$$\exp : \mathfrak{n}^+ \rightarrow N^+$$

is a diffeomorphism. Furthermore, the map

$$K \times A \times N^+ \rightarrow G, \quad (k, a, n) \rightarrow kan$$

is a surjective diffeomorphism.

This result is usually indicated by the shorthand  $G = KAN^+$ .

If  $G = \mathrm{GL}_n(\mathbf{R})$ , we may take  $\mathbf{a}$  to be the diagonal matrices and  $\mathfrak{n}^+$  to be the strictly upper triangular matrices, in which case the Iwasawa decomposition amounts to the Gram-Schmidt orthonormalization procedure [Hill, Stra]. Also the exponential map on the strictly upper triangular matrices is a polynomial map (of degree  $n - 1$ ) with polynomial inverse.

For general reductive  $G \subseteq \mathrm{GL}_n(\mathbf{R})$ , if we position  $G$  correctly, by conjugation if necessary, we can arrange that the Iwasawa decomposition for  $G$  is the same as for  $\mathrm{GL}_n(\mathbf{R})$ . Indeed we can choose an orthonormal eigenbasis  $\{b_j\}_{j=1}^n$  for  $\mathbf{R}^n$  consisting of eigenvectors for  $\mathbf{a}$ , and we can order this eigenbasis by picking an element  $x$  in  $\mathfrak{a}^+$ , and requiring that the  $x$ -eigenvalue of  $b_j$  decreases as  $j$  increases. Then the commutation relations (A.2.2.2)(ii) show, by a calculation like that of formula (3.5.1.3), that, with respect to the basis  $\{b_j\}$ , the Lie algebra  $\mathfrak{n}^+$  consists of strictly upper triangular matrices. Also,  $\mathbf{a}$  consists of diagonal matrices, and  $\mathbf{k}$  still consists of skew-symmetric matrices.

Now consider  $g \in G$ , and let  $g = kan$  be its Iwasawa decomposition as an element of  $\mathrm{GL}_n(\mathbf{R})$ . The infinitesimal decomposition (A.2.3.4), combined with the Inverse Function Theorem [Lang2], implies that for  $g$  near enough to the identity, we have  $k \in K$ ,  $a \in A$ , and  $n \in N^+$ . Since  $G$  is connected and  $k$ ,  $a$ , and  $n$  depend analytically on  $g$ , it follows that the Iwasawa decomposition for  $\mathrm{GL}_n(\mathbf{R})$  provides the Iwasawa decomposition for  $G$  also.

We also record, but do not prove, the general version of the Bruhat decomposition (cf. §1.2 for the case of  $\mathrm{GL}_n$ ). Let  $M$  be the centralizer in  $K$  of  $A$ . Set  $Q_0 = MAN^+$ . Let  $W$  be the Weyl group of  $A$ .

(A.2.3.6) (Bruhat decomposition). The group  $G$  is a disjoint union of  $(N^+, Q_0)$  double cosets with representatives from  $W$ :

$$G = \bigcup_{w \in W} N^+ w Q_0.$$

This is sometimes abbreviated  $G = N^+ W Q_0$ .

In the Bruhat decomposition, we can use, instead of  $N^+$ , the "opposite" unipotent group  $N^- = \exp \mathfrak{n}^-$ , where

$$(A.2.3.7) \quad \mathfrak{n}^- = \sum_{\alpha \in \Delta^+} \mathfrak{g}_{-\alpha}.$$

It is clear from equation (A.2.2.2) and definitions (A.2.3.3) and (A.2.3.7) that

$$(A.2.3.8) \quad \mathfrak{g} = \mathfrak{n}^- \oplus \mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}^+ = \mathfrak{n}^- \oplus \mathfrak{q}_0,$$

where  $\mathfrak{q}_0 = \mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}^+$  is the Lie algebra of  $Q_0$ . The decomposition (A.2.3.8) implies that the double coset  $N^-Q_0 = N^-MAN^+$  is open in  $G$ . This is the analog for  $G$  of the  $L-U$  decomposition of §1.1, and the Bruhat decomposition  $G = N^-WQ_0$  extends this decomposition to an arbitrary element of  $G$ .

**A.2.4. Parabolic subgroups.** The group  $Q_0 = MAN^+$  figuring in the Bruhat decomposition (A.2.3.6), or any subgroup of  $G$  conjugate to  $Q_0$ , is called a *minimal parabolic subgroup* of  $G$ . A *parabolic subgroup* of  $G$  is any subgroup which contains a minimal parabolic subgroup.

We can construct parabolic subgroups containing  $Q_0$  as follows. Let  $\mathfrak{a}$  be the split Cartan subalgebra used above, and let  $\mathfrak{a}^+ \subseteq \mathfrak{a}$  be the Weyl chamber defining  $\mathfrak{n}^+$ . Thus  $\mathfrak{a}^+$  is a convex cone, defined as the intersection of certain halfspaces cut out by the root hyperplanes in  $\mathfrak{a}$ . Let  $\mathcal{F} = \{\alpha_j\}_{j=1}^r$  be the set of roots which are nonnegative on  $\mathfrak{a}^+$ , and whose kernel intersects  $\mathfrak{a}^+$  in a set of codimension 1 in  $\mathfrak{a}$ —equivalently  $\mathfrak{a}^+ \cap \ker \alpha_j$  spans  $\ker \alpha_j$ . The  $\alpha_j$  are called the *fundamental positive roots*. Let  $\mathcal{F}_1 \subseteq \mathcal{F}$  be a subset of the fundamental positive roots. Set

$$(A.2.4.1) \quad \begin{aligned} \mathfrak{a}_1 &= \bigcap_{\alpha \in \mathcal{F}_1} \ker \alpha = \{x \in \mathfrak{a} : \alpha(x) = 0, \text{ all } \alpha \in \mathcal{F}_1\}, \\ \mathfrak{q}_1 &= \mathfrak{m} + \mathfrak{a} + \mathfrak{n}^+ + \sum_{\substack{\ker \beta \supseteq \mathfrak{a}_1}} \mathfrak{g}_\beta \end{aligned}$$

Let  $Q_1$  be the subgroup of  $G$  generated by  $Q_0$  and the exponentials of elements of  $Q_1$ . Then  $Q_1$  is a parabolic subgroup of  $G$ .

Set

$$(A.2.4.2) \quad \mathfrak{n}_1^+ = \sum_{\substack{\beta \in \Delta^+ \\ \mathfrak{a}_1 \not\subseteq \ker \beta}} \mathfrak{g}_\beta, \quad N_1^+ = \exp \mathfrak{n}_1^+.$$

Then we have the decompositions

$$(A.2.4.3) \quad \begin{aligned} \mathfrak{q}_1 &= (\mathfrak{q}_1 \cap \theta(\mathfrak{q}_1)) \oplus \mathfrak{n}_1^+, \quad Q_1 = (Q_1 \cap \theta(Q_1)) \cdot N_1^+, \\ \mathfrak{q}_1 \cap \theta(\mathfrak{q}_1) &= \mathfrak{m} \oplus \mathfrak{a} \oplus \sum_{\mathfrak{a}_1 \subseteq \ker \beta} \mathfrak{g}_\beta. \end{aligned}$$

We see that  $\mathfrak{q}_1 \cap \theta(\mathfrak{q}_1)$  is the centralizer of  $\mathfrak{a}_1$  in  $\mathfrak{g}$ , and we can likewise characterize  $Q_1 \cap \theta(Q_1)$  as the centralizer of  $A_1 = \exp \mathfrak{a}_1$  in  $G$ .

Further decompose  $\mathfrak{q}_1 \cap \theta(\mathfrak{q}_1)$  as follows. Let  $\mathfrak{m}_1$  denote the Lie algebra generated by  $\mathfrak{m}$  and the root spaces  $\mathfrak{g}_\beta$  contained in  $\mathfrak{q}_1 \cap \theta(\mathfrak{q}_1)$ . Then it can be checked that  $\mathfrak{m}_1 \cap \mathfrak{a}$  is the span of the elements  $[x_\alpha, \theta(x_\alpha)]$ ,  $x_\alpha \in \mathfrak{g}_\alpha$ ,  $\alpha \in \mathcal{F}_1$ . In particular,  $\mathfrak{a} = (\mathfrak{m}_1 \cap \mathfrak{a}) \oplus \mathfrak{a}_1$ . Thus we have

$$(A.2.4.4) \quad \mathfrak{q}_1 \cap \theta(\mathfrak{q}_1) = \mathfrak{m}_1 \oplus \mathfrak{a}_1 \quad \text{and} \quad \mathfrak{q}_1 = \mathfrak{m}_1 \oplus \mathfrak{a}_1 \oplus \mathfrak{n}_1^+.$$



This is called the *Langlands decomposition* of  $\mathfrak{q}_1$  [GaVa, Knap2, Wall2]. It is easy to see that  $\mathfrak{m}_1$  is a reductive Lie subalgebra of  $\mathfrak{g}$ . We can define a corresponding subgroup  $M_1$  of  $G$  as follows. The group  $Q_1 \cap \theta(Q_1)$  has a Cartan decomposition (cf. (A.2.3.1))

$$Q_1 \cap \theta(Q_1) = (Q_1 \cap \theta(Q_1) \cap K) \exp(\mathfrak{q}_1 \cap \mathfrak{p}).$$

Set

$$(A.2.4.5) \quad M_1 = (Q_1 \cap \theta(Q_1) \cap K) \exp(\mathfrak{m}_1 \cap \mathfrak{p}).$$

Then the Cartan decomposition plus decompositions (A.2.4.4) tell us that

$$(A.2.4.6) \quad P_1 = M_1 A_1 N_1^+$$

in the strong sense that the map from  $M_1 \times A_1 \times N_1^+$  defined by multiplication to  $P_1$  is a diffeomorphism. The factorization (A.2.4.6) is called the *Langlands decomposition* of  $P_1$ .

The procedure sketched above constructs  $2^r$ , where  $r = \#(\mathcal{F})$ , parabolic subgroups of  $G$  containing  $P_0$ . These are all possible parabolics containing  $P_0$ . To show this requires a more detailed study of root systems than we wish to give here. Instead we will finish as we started, by looking at  $GL_n$ . We will sketch how to see that possibilities for subgroups of  $GL_n$  containing the Borel subgroup of upper triangular matrices are the groups of block upper triangular matrices defined by various partial flags (cf. §1.4). Consider the basis  $\{E_{jk}\}_{j,k=1}^n$  of standard matrix units for  $\mathfrak{gl}_n$ . These satisfy the commutation relations

$$[E_{jk}, E_{lm}] = \delta_{kl} E_{jm} - \delta_{jm} E_{lk}.$$

The upper triangular matrices  $\mathfrak{b}^+$  are the span of the  $E_{jk}$  with  $j \leq k$ . Suppose we add to this another element  $x = \sum c_{lm} E_{lm}$ . Since the  $E_{lm}$ 's are eigenvectors for the  $\text{ad } E_{jj}$ , with distinct eigenvalues, we find that if  $c_{lm} \neq 0$ , then  $E_{lm}$  is in the algebra generated by  $\mathfrak{b}^+$  and  $x$ . So take  $x = E_{lm}$  for some  $l > m$ . Taking commutators with  $E_{jl}$ ,  $j \leq l$ , shows us  $E_{jm}$  belongs to the algebra generated by  $E_{lm}$  and  $\mathfrak{b}^+$ . Similarly, we must have  $E_{lk}$ ,  $k \geq m$ , in this algebra. Repeating this process, we find that all  $E_{jk}$ ,  $j \leq l$ ,  $k \geq m$ , are in the algebra. These span the whole block to the upper right of  $E_{lm}$ . Next suppose we have two elements  $E_{lm}, E_{rs}$  which generate overlapping blocks, in the sense that  $m < s \leq l < r$ . Then from the argument above, we can find  $E_{rl}$  in the algebra generated by  $\mathfrak{b}^+$  and  $E_{rs}$ . Hence  $[E_{rl}, E_{lm}] = E_{rm}$  is in our algebra, and therefore so is the smallest diagonal block containing both  $E_{lm}$  and  $E_{rs}$ . Thus we get the general parabolic containing  $\mathfrak{b}^+$  by adding *disjoint* diagonal blocks. We remark that the calculations sketched above are similar to those used in the context of general root systems.

**Acknowledgments.** In writing this paper I have benefitted from the insights and remarks of many people. I thank James Arthur, Richard Askey, Joseph

This is called the *Langlands decomposition* of  $\mathfrak{q}_1$  [GaVa, Knap2, Wall2]. It is easy to see that  $\mathfrak{m}_1$  is a reductive Lie subalgebra of  $\mathfrak{g}$ . We can define a corresponding subgroup  $M_1$  of  $G$  as follows. The group  $Q_1 \cap \theta(Q_1)$  has a Cartan decomposition (cf. (A.2.3.1))

$$Q_1 \cap \theta(Q_1) = (Q_1 \cap \theta(Q_1) \cap K) \exp(\mathfrak{q}_1 \cap \mathfrak{p}).$$

Set

$$(A.2.4.5) \quad M_1 = (Q_1 \cap \theta(Q_1) \cap K) \exp(\mathfrak{m}_1 \cap \mathfrak{p}).$$

Then the Cartan decomposition plus decompositions (A.2.4.4) tell us that

$$(A.2.4.6) \quad P_1 = M_1 A_1 N_1^+$$

in the strong sense that the map from  $M_1 \times A_1 \times N_1^+$  defined by multiplication to  $P_1$  is a diffeomorphism. The factorization (A.2.4.6) is called the *Langlands decomposition* of  $P_1$ .

The procedure sketched above constructs  $2^r$ , where  $r = \#(\mathcal{F})$ , parabolic subgroups of  $G$  containing  $P_0$ . These are all possible parabolics containing  $P_0$ . To show this requires a more detailed study of root systems than we wish to give here. Instead we will finish as we started, by looking at  $GL_n$ . We will sketch how to see that possibilities for subgroups of  $GL_n$  containing the Borel subgroup of upper triangular matrices are the groups of block upper triangular matrices defined by various partial flags (cf. §1.4). Consider the basis  $\{E_{jk}\}_{j,k=1}^n$  of standard matrix units for  $\mathfrak{gl}_n$ . These satisfy the commutation relations

$$[E_{jk}, E_{lm}] = \delta_{kl} E_{jm} - \delta_{jm} E_{lk}.$$

The upper triangular matrices  $\mathfrak{b}^+$  are the span of the  $E_{jk}$  with  $j \leq k$ . Suppose we add to this another element  $x = \sum c_{lm} E_{lm}$ . Since the  $E_{lm}$ 's are eigenvectors for the  $\text{ad } E_{jj}$ , with distinct eigenvalues, we find that if  $c_{lm} \neq 0$ , then  $E_{lm}$  is in the algebra generated by  $\mathfrak{b}^+$  and  $x$ . So take  $x = E_{lm}$  for some  $l > m$ . Taking commutators with  $E_{jl}$ ,  $j \leq l$ , shows us  $E_{jm}$  belongs to the algebra generated by  $E_{lm}$  and  $\mathfrak{b}^+$ . Similarly, we must have  $E_{lk}$ ,  $k \geq m$ , in this algebra. Repeating this process, we find that all  $E_{jk}$ ,  $j \leq l$ ,  $k \geq m$ , are in the algebra. These span the whole block to the upper right of  $E_{lm}$ . Next suppose we have two elements  $E_{lm}, E_{rs}$  which generate overlapping blocks, in the sense that  $m < s \leq l < r$ . Then from the argument above, we can find  $E_{rl}$  in the algebra generated by  $\mathfrak{b}^+$  and  $E_{rs}$ . Hence  $[E_{rl}, E_{lm}] = E_{rm}$  is in our algebra, and therefore so is the smallest diagonal block containing both  $E_{lm}$  and  $E_{rs}$ . Thus we get the general parabolic containing  $\mathfrak{b}^+$  by adding *disjoint* diagonal blocks. We remark that the calculations sketched above are similar to those used in the context of general root systems.

**Acknowledgments.** In writing this paper I have benefitted from the insights and remarks of many people. I thank James Arthur, Richard Askey, Joseph

Bernstein, Sol Friedberg, Steve Gelbart, Robert Langlands, Alex Lubotzky, Dan Mostow, Ilya Piatetski-Shapiro, Steve Rallis, George Seligman, Eli Stein, David Vogan, Nolan Wallach, and Gregg Zuckerman for helpful conversation and advice. I apologize to others I have forgotten to name. Thanks to Felix Browder and Carol Moura for their patience. Thanks to my wife Lyn for support and encouragement in the last agonies of getting this done. I have had many occasions in the past to thank Mrs. Mel DelVecchio for her superb typing and cooperative spirit. This time I would also like to thank the Lord and Phyllis Stevens for bringing Mel to the Yale Mathematics Department.

## REFERENCES

- [Abar] H. Abarbanel, *The inverse  $r$ -squared force: an introduction to its symmetries*, Studies in Mathematical Physics—Essays in Honor of Valentine Bargmann (E. Lieb et al., eds.), Princeton University Press, Princeton, 1976, pp. 3–18.
- [AbMa] R. Abraham and J. Marsden, *Foundations of mechanics*, 2nd ed., Benjamin-Cummings, Reading, MA, 1978.
- [Adam1] J. Adams,  *$L$ -functoriality for dual pairs*, Orbites Unipotentes et Représentations. II, Groupes  $p$ -Adiques et Réels, Asterisque, nos. 171–172, Soc. Math. France, Paris, 1989, pp. 85–129.
- [AdVo1] J. Adams and D. Vogan, *Harish-Chandra's method of descent*, preprint.
- [Adle1] M. Adler, *On a trace functional for pseudo-differential operators and the symplectic structure of Korteweg-DeVries type equations*, Invent. Math. **50** (1979), 219–248.
- [AdvM] M. Adler and P. van Moerbeke, *Completely integrable systems, Euclidean Lie algebras and curves*, Adv. in Math. **38** (1980), 267–317.
- [AlWa] S. Aloff and N. Wallach, *An infinite family of distinct 7-manifolds admitting positively curved Riemannian structures*, Bull. Amer. Math. Soc. (N.S.) **5** (1975), 93–97.
- [Andr1] G. Andrews, *The theory of partitions*, Encyclopedia Math. Appl., vol. 2, Addison-Wesley, Reading, MA, 1976.
- [Andr2] ———,  *$q$ -series: Their development and application in analysis, number theory, combinatorics, physics and computer algebra*, CBMS Regional Conf. Ser. in Math., vol. 66, Amer. Math. Soc., Providence, RI, 1986.
- [AnBF] G. Andrews, R. Baxter, and P. Forrester, *Eight-vertex SOS model and generalized Rogers-Ramanujan type identities*, J. Statist. Phys. **35** (1984), 193–266.
- [ArnA] V. Arnold and A. Avez, *Ergodic properties of dynamical systems*, Benjamin, New York and Amsterdam, 1968.
- [Arib] F. Aribaud, *Une nouvelle démonstration d'un théorème de R. Bott et B. Kostant*, Bull. Math. Soc. France **95** (1967), 205–242.
- [Arno1] V. Arnold, *Mathematical methods of classical mechanics*, Springer-Verlag, Berlin, 1978.
- [Arno2] ———, *Critical points of smooth functions*, Proc. Internat. Congr. Math. (Vancouver, 1974), Canadian Mathematical Congress, 1975, pp. 19–39.
- [ArCo] J. Arthur and L. Clozel, *Simple algebras, base change and the advanced theory of the trace formula*, Ann. of Math. Stud., no. 120, Princeton Univ. Press, Princeton, NJ, 1989.
- [ArTa] E. Artin and J. Tate, *Class field theory*, Benjamin, New York, 1967.
- [Arth2] J. Arthur, *Unipotent automorphic representations: Conjectures*, Orbites Unipotentes et Représentations. II, Groupes  $p$ -adiques et réels, Asterisque, nos. 171–172, Soc. Math. France, Paris, 1989, pp. 13–71.
- [Arth3] ———, *Unipotent automorphic representations: global motivation*, Automorphic Forms, Shimura Varieties and  $L$ -functions (L. Clozel and J. Milne, eds.), Perspect. Math., vol. 19, Academic Press, Boston, 1990, pp. 1–76.
- [Arth4] ———, *Harmonic analysis and an  $L^2$ -Lefschetz formula*, The Mathematical Heritage of Hermann Weyl, Proc. Sympos. Pure Math., vol. 48, Amer. Math. Soc., Providence, RI, 1988.

- [Arth5] ———, *The trace formula for non-compact quotients*, Vol. 2, Proc. Internat. Congr. Math. (Warsaw, 1983), North-Holland, Amsterdam, New York, Oxford, 1984, pp. 849–859.
- [Arti] E. Artin, *Geometric algebra*, Interscience, New York, 1957.
- [Ati1] M. Atiyah, *K-theory*, Lecture notes by D.W. Anderson, Harvard University, 1964.
- [Ati2] ———, *Convexity and commuting Hamiltonians*, Bull. London Math. Soc. **14** (1982), 1–15.
- [AtBo] M. Atiyah and R. Bott, *The Yang-Mills equations over Riemann surfaces*, Proc. Roy. Soc. London Ser. A **308** (1982), 523–615.
- [AtSc] M. Atiyah and W. Schmid, *A geometric construction of the discrete series for semisimple Lie groups*, Invent. Math. **42** (1977), 1–62.
- [AtSi] M. Atiyah and I. Singer, *The index of elliptic operators*. III, Ann. of Math. (2) **87** (1968), 546–604.
- [AuGH] L. Auslander, et al., *Flows on homogeneous spaces*, Ann. of Math. Stud., no. 53, Princeton Univ. Press, Princeton, NJ, 1963.
- [AuKo] L. Auslander and B. Kostant, *Polarization and unitary representations of solvable Lie groups*, Invent. Math. **14** (1971), 155–254.
- [AuMo] L. Auslander and C. Moore, *Unitary representations of solvable Lie groups*, Mem. Amer. Math. Soc., vol. 62, Amer. Math. Soc., Providence, RI, 1966.
- [Bak] A. Bak, *Le problème des sous-groupes de congruence et le problème métaplectique pour les groupes classiques de rang  $\geq 1$* , C. R. Acad. Sci. Paris Sér. I Math. **292** (1981), 307–310.
- [Ban] E. van den Ban, *The principal series for a reductive symmetric space. I:  $H$ -fixed distribution vectors*, Ann. Sci. École Norm. Sup. (4) **21** (1988), 354–412.
- [BaSc] E. P. van den Ban and H. Schlichtkrull, *Local boundary data of eigenfunctions on a Riemannian symmetric space*, Invent. Math. **98** (1989), 639–657.
- [Barb1] D. Barbasch, *The unitary dual for complex classical groups*, Invent. Math. **96** (1989), 103–176.
- [BaVo] D. Barbasch and D. Vogan, *Unipotent representations for complex semisimple groups*, Ann. of Math. (2) **121** (1985), 41–110.
- [Barg1] V. Bargmann, *Irreducible unitary representations of the Lorentz group*, Ann. of Math. (2) **48** (1947), 568–640.
- [Barg2] ———, *On a Hilbert space of analytic functions and an associated integral transform*, Part I, Comm. Pure Appl. Math. **14** (1961), 187–214.
- [Barg3] ——— (ed.), *Group representations in mathematics and physics* (Battelle Seattle 1969 Rencontres), Lecture Notes in Phys., vol. 6, Springer-Verlag, New York.
- [BaMS] H. Bass, J. Milnor, and J.-P. Serre, *Solution of the congruence subgroup problem for  $SL_n$  ( $n \geq 3$ ), and  $Sp_{2n}$  ( $n \geq 2$ )*, Inst. Hautes Études Sci. Publ. Math. **33** (1967), 59–139.
- [Baxt] R. Baxter, *Exactly solved models in statistical mechanics*, Academic Press, London and New York, 1982.
- [Beal] R. Beals, *A general calculus of pseudo-differential operators*, Duke Math. J. **42** (1975), 1–42.
- [BeCo] R. Beals and R. Coifman, *Scattering and inverse scattering for first order systems*, Comm. Pure. Appl. Math. **37** (1984), 39–90.
- [BeDT] R. Beals, P. Deift, and C. Tomei, *Direct and inverse scattering on the line*, Math. Surveys Monographs, vol. 28, Amer. Math. Soc., Providence, RI, 1988.
- [BeGr] Richard Beals and P. Greiner, *Calculus on Heisenberg manifolds*, Ann. of Math. Stud., no. 119, Princeton Univ. Press, Princeton, NJ, 1986.
- [BeBe] A. Beilinson and J. Bernstein, *Localisation de  $\mathfrak{g}$ -modules*, C. R. Acad. Sci. Paris (Ser. A) **292** (1981), 15–18.
- [BePZ] A. Belavin, A. Polyakov, and A. Zamolodchikov, *Infinite conformal symmetry in two-dimensional quantum field theory*, Nuclear Phys. B **241** (1984), 333–380.
- [BeTu] I. Benn and R. Tucker, *An introduction to spinors and geometry with applications in physics*, Adam Hilger, Bristol and New York, 1987.
- [Bera] P. Berard, *Variétés Riemanniennes isospectrales non-isométriques*, Sémin. Bourbaki, no. 705, (March 1989).
- [Bere] F. Berezin, *Some remarks about the associated envelope of a Lie algebra*, Functional Anal. Appl. **1** (1967), 91–102.

- [Berg1] M. Berger, *Les espaces symétriques non-compacts*, Ann. Sci. École Norm. Sup. **74** (1951), 85–177.
- [Berg2] ———, *Les variétés riemanniennes homogenes normales simplement connexes a courbure strictement positive*, Ann. Scuola Norm. Sup. Pisa **15** (1961), 179–246.
- [Bert] P. Bernat, *Sur les représentations unitaires des groupes de Lie résolubles*, Ann. Sci. École Norm. Sup. **82** (1965), 37–99.
- [Bern1] J. Bernstein, *Modules over the rings of differential operators, a study of the fundamental solutions of equations with constant coefficients*, Functional Anal. Appl. **5** (1971), 1–16.
- [Bern2] ———, *The analytic continuation of generalized functions with respect to a parameter*, Functional Anal. Appl. **6** (1972), 26–40.
- [Bern3] ———, *All reductive  $p$ -adic groups are tame*, Functional Anal. Appl. **8** (1974), 91–93.
- [BDKM] J. Bernstein, P. Deligne, D. Kazhdan, and M. Vigneras, *Représentations des groupes réductifs sur un corps local*, Hermann, Paris, 1984.
- [BGG1] J. Bernstein, I. Gelfand, and S. Gelfand, *Differential operators on the base affine space and a study of  $\mathfrak{g}$ -modules*, J. Bolyai Math. Soc., Budapest, Summer School in Math. (I. M. Gelfand, ed.), 1971, 21–64.
- [BGG2] ———, *The structure of representations generated by vectors of largest weight*, Functional Anal. Appl. **5** (1971), 1–8.
- [BGG3] ———, *On a category of  $\mathfrak{g}$ -modules*, Functional Anal. Appl. **10** (1976), 87–92.
- [BeGP] J. Bernstein, I. Gelfand, and V. Ponomarev, *Coxeter functors and Gabriel's Theorem*, Russian Math. Surveys **28** (1973), 17–32.
- [BeZe] J. Bernstein and A. Zelevinski, *Representations of the group  $GL(n, F)$* , Uspekhi Mat. Nauk. **31** (1976), 5–70; Russian Math. Surveys **31** (1976), 1–68.
- [Bers] L. Bers, *Finite dimensional Teichmüller spaces and generalizations*, Bull. Amer. Math. Soc. (N.S.) **5** (1981), 131–172.
- [BiLo1] L. Biedenharn and J. Louck, *Angular momentum in quantum physics, theory and applications*, Encyclopedia Math. and Appl., vol. 8, Addison-Wesley, Reading, MA, 1981.
- [BiLo2] L. Biedenharn and J. Louck, *The Racah-Wigner algebra in quantum theory*, Encyclopedia Math. Appl., vol. 9, Addison-Wesley, Reading, MA, 1981.
- [Bien] F. Bien,  *$\mathcal{D}$ -modules and spherical representations*, Princeton Math. Notes, no. 39, Princeton Univ. Press, Princeton, NJ, 1990.
- [BiRo] G. Birkhoff and G.-C. Rota, *Ordinary differential equations*, Gunn, Boston, 1962.
- [BIRa] D. Blasius and D. Ramakrishnan, *Maass forms and Galois representations*, Galois groups over  $\mathbb{Q}$ , Math. Sci. Res. Inst. Publ., no. 16, Springer-Verlag, Berlin and New York, 1984.
- [Blat] R. Blattner, *Quantization and representation theory*, Harmonic Analysis on Homogeneous Spaces, Proc. Sympos. Pure Math., vol. 26, Amer. Math. Soc., Providence, RI, 1973, pp. 147–166.
- [BIBR] A. Bloch, R. Bröckert, and T. Ratiu, *A new formulation of the generalized Toda lattice equations and their fixed point analysis via the momentum map*, Bull. Amer. Math. Soc. (N.S.) **23** (1990), 477–485.
- [Boch] S. Bochner, *Lectures on Fourier integrals*, Ann. of Math. Stud., no. 42, Princeton Univ. Press, Princeton, NJ, 1959.
- [Boer] H. Boerner, *Representations of groups*, North-Holland, Amsterdam, 1963.
- [Bohr] N. Bohr, *Theory of spectra and atomic constitution*, 2nd ed., Cambridge, 1924.
- [Bor1] A. Borel, *Linear algebraic groups*, Benjamin, New York, 1967.
- [Bor2] A. Borel, et al., *Algebraic  $D$ -modules*, Perspect. Math., vol. 2, Academic Press, Boston, MA, 1987.
- [Bor3] A. Borel, *Introduction aux groupes arithmétiques*, Publ. Inst. Math. Univ. Strasbourg XV, Actualités. Sci. Indust., no. 1341, Hermann, Paris, 1969.
- [Bor4] ———, *On the development of Lie group theory*, Math. Intelligencer **2** (1979–80), 67–72.
- [Bor5] ———, *Automorphic  $L$ -functions*, Automorphic Forms, Representations and  $L$ -Functions, Proc. Sympos. Pure Math., vol. 33, part 2, Amer. Math. Soc., Providence, RI, 1979, pp. 27–62.
- [Bor6] ———, *Topics in the homology theory of fibre bundles*, Lecture Notes in Math., vol. 36, Springer-Verlag, Berlin, 1967.
- [BoCa] A. Borel and W. Casselman, eds., *Automorphic forms, representations and  $L$ -functions*, Proc. Sympos. Pure Math., vol. 33, Amer. Math. Soc., Providence, RI, 1979.

- [BoHC] A. Borel and Harish-Chandra, *Arithmetic subgroups of algebraic groups*, Ann. of Math. (2) **75** (1962), 485–535.
- [BoWa] A. Borel and N. Wallach, *Continuous cohomology, discrete subgroups, and representations of reductive groups*, Ann. of Math. Stud., no. 94, Princeton Univ. Press, Princeton, NJ, 1980.
- [BoBM] W. Borho, J.-L. Brylinski, and R. MacPherson, *Nilpotent orbits, primitive ideals, and characteristic classes*, Progr. Math., vol. 78, Birkhäuser, Boston, 1989.
- [Bott] R. Bott, *Homogeneous vector bundles*, Ann. of Math. (2) **66** (1957), 203–248.
- [Boua] A. Bouaziz, *Sur les caractères des groupes de Lie réductifs non connexes*, J. Funct. Anal. **70** (1987), 1–79.
- [Bour] N. Bourbaki, *Groupes et algèbres de Lie*, Chapters 4–6, Hermann, Paris, 1968.
- [BoLa] J.-P. Bourguignon and M. Lawson, Jr., *Yang-Mills theory: Its physical origins and differential geometric aspects*, Seminar on Differential Geometry (S. T. Yau, ed.), Ann. of Math. Stud., no. 102, Princeton Univ. Press, Princeton, NJ, 1982, pp. 395–422.
- [BoDP] W. Boyce and R. DiPrima, *Elementary differential equations and boundary value problems*, 4th ed., Wiley, New York, 1986.
- [Brau] R. Brauer, *Sur les invariants intégraux des variétés des groupes de Lie simples clos*, C. R. Acad. Sci. Paris **201** (1935), 419–421.
- [BrLa] L. Breen and J. P. Labesse, *Variétés de Shimura et fonctions  $L$* , Publ. Math. Univ. Paris VII, U.E.R. de Math., no. 212, CNRS, Paris, 1979.
- [BrMP] M. Bremner, R. Moody, and J. Patera, *Tables of dominant weight multiplicities for representations of simple Lie algebras*, Marcel Dekker, New York, 1985.
- [BrZe] D. Bressoud and D. Zeilberger, *A proof of Andrews'  $q$ -Dyson conjecture*, Discrete Math. **54** (1985), 201–224.
- [Brie] E. Brieskorn, *Singular elements of semi-simple algebraic groups*, Proc. Internat. Congr. Math. Nice 2, Paris, Gauthier-Villars, 1970, pp. 279–284.
- [Brocl] R. Brockett, *System theory on group manifolds and coset spaces*, SIAM J. Control **10** (1972), 265–284.
- [Brocl2] ———, *Lectures on Lie algebras in systems and filtering*, Stochastic Systems: The Mathematics of Filtering and Identification and Applications (M. Hazewinkel and J. Williams, eds.), Reidel, 1981.
- [BrGo] R. Brooks and C. Gordon, *Isospectral families of conformally equivalent Riemannian metrics*, Bull. Amer. Math. Soc. (N.S.) **23** (1990), 433–436.
- [Brow] W. Browder, *Torsion in  $H$ -spaces*, Ann. of Math. (2) **74** (1961), 24–51.
- [Bruh] F. Bruhat, *Sur les représentations induites des groupes de Lie*, Bull. Soc. Math. France **84** (1962), 666–710.
- [BrTi] F. Bruhat and J. Tits, *Groupes réductifs sur un corps local*. I, II, Inst. Hautes Études Sci. Publ. Math. **41** (1972), 5–25; **60** (1989), 197–376.
- [BrKa] J.-L. Brylinski and M. Kashiwara, *Démonstration de la conjecture de Kazhdan-Lusztig sur les modules de Verma*, C. R. Acad. Sci. Paris Sér. A **291** (1980), 373–376.
- [BuKu] C. Bushnell and P. Kutzko, *The admissible dual of  $GL_N$  via compact open subgroups*, preprint, 1990.
- [CaEH] A. Carey, M. Eastwood, and K. Hannabuss, *Riemann surfaces, Clifford algebras and infinite dimensional groups*, Comm. Math. Phys. **130** (1990), 217–236.
- [Crml] M. Carmeli, *Group theory and general relativity*, McGraw-Hill, New York, 1977.
- [Carm] J. Carmona, *Représentations du groupe de Heisenberg dans les espaces de  $(0, q)$  formes*, Math. Ann. **205** (1973), 89–112.
- [Carr] J. Carrell, *Holomorphic  $C^*$ -actions and vector fields on projective varieties*, Topics in the Theory of Algebraic Groups, Notre Dame Math. Lectures, vol. 10, Univ. of Notre Dame Press, South Bend, IN, 1982.
- [Crtn1] E. Cartan, *Sur la structure des groupes de transformations finis et continus*, These, Paris, Nony (2nd ed., Vuibert, 1933); *Oeuvres complètes*, CNRS, Paris, 1984, pp. 137–253.
- [Crtn2] ———, *Les groupes projectifs qui ne laissent invariante aucune multiplicité plane*, Bull. Soc. Math. France **41** (1913), 53–96; *Oeuvres complètes*, CNRS, Paris, 1984, pp. 355–398.
- [Crtn3] ———, *Les groupes réels simples finis et continus*, Ann. Ser. École Norm. Sup. **31** (1914), 263–355; *Oeuvres complètes*, CNRS, Paris, 1984, pp. 399–491.

- [Crtn4] ———, *Sur certaines formes riemanniennes remarquables des géométries a groupe fondamentale simple*, Ann. Sci. École Norm. Sup. **44** (1927), 345–467.
- [Crtn5] ———, *La topologie des espaces représentatifs des groupes de Lie*, Enseign. Math. **35** (1936), 177–200; *Oeuvres complètes*, CNRS, Paris, 1984, pp. 1307–1331.
- [Crtn6] ———, *La topologie des espaces homogènes clos*, Mem. Sem. Anal. Vect. Moscow **3** (1937), 388–394; *Oeuvres complètes*, vol. 1, CNRS, Paris, 1984, pp. 1331–1339.
- [Crtn7] ———, *Notice sur les travaux scientifiques*, *Oeuvres Complètes*, CNRS, Paris, 1984.
- [Crt1] R. Carter, *Simple groups of Lie type*, Wiley, London, 1972.
- [Crt2] ———, *Finite groups of Lie type*, Wiley-Interscience, Chichester, New York, 1985.
- [Cart1] P. Cartier, *Quantum mechanical commutation relations and theta functions*, Proc. Sympos. Pure Math., vol. 9, Amer. Math. Soc., Providence, RI, 1966, pp. 361–383.
- [Cart2] ———, *Representation of  $p$ -adic groups: a survey*, Automorphic Forms, Representation Theory and  $L$ -functions, Proc. Sympos. Pure Math., vol. 33, part 1, Amer. Math. Soc., Providence, RI, 1979, pp. 111–156.
- [CaCo] L. Casan and D. Collingwood, *Weight filtrations for induced representations of real reductive Lie groups*, Adv. in Math. **73** (1989), 79–144.
- [CaMi] W. Casselman and D. Milicic, *Asymptotic behavior of matrix coefficients of admissible representations*, Duke Math. J. **49** (1982), 869–930.
- [CaOs] W. Casselman and M. Osborne, *The  $n$ -cohomology of representations with an infinitesimal character*, Compositio Math. **31** (1975), 219–227.
- [Cas1] J. Cassels, *An introduction to diophantine approximation*, Cambridge Univ. Press, Cambridge, 1965.
- [ChEb] J. Cheeger and D. Ebin, *Comparison theorems in Riemannian geometry*, North-Holland, Amsterdam, 1975.
- [Cher1] I. Cherednik, *A new interpretation of Gelfand-Tsetlin bases*, Duke Math. J. **54** (1987), 563–577.
- [Cher2] ———, *A unification of Knizhnik-Zamolodchikov and Dunkl operators via affine Hecke algebras*, R.I.M.S. preprint no. 724, Kyoto University, Kyoto, 1990.
- [Chev1] C. Chevalley, *Sur certains groupes simples*, Tôhoku Math. J. **7** (1955), 14–66.
- [Chev2] ———, *Séminaire sur la classification des groupes de Lie algébriques* (mimeographed notes), Paris, 1956–1958.
- [Chev3] ———, *Theory of Lie groups*, Princeton Univ. Press, Princeton, NJ, 1946.
- [Chev4] ———, *Théorie des Groupes de Lie. II*, Hermann, Paris, 1951.
- [Chev5] ———, *Théorie des Groupes de Lie. III*, Hermann, Paris, 1955.
- [Chev6] ———, *Sur la classification des algèbres de Lie simples et de leurs représentations*, C. R. Acad. Sci. Paris **227** (1948), 1136–1138.
- [Chev7] ———, *Invariants of finite groups generated by reflections*, Amer. J. Math. **77** (1955), 778–782.
- [ClMi] L. Clozel and J. Milne, eds., *Automorphic forms, Shimura varieties, and  $L$ -functions. I, II*, Perspect. Math., vols. 10, 11, Academic Press, Boston, 1990.
- [CoPS1] J. Cogdell and I. Piatetski-Shapiro, *Base change for  $SL(2)$* , J. Number Theory **27** (1987), 287–303.
- [CoPS2] ———, *Base change for Saito-Kurokawa representations of  $PGSp(4)$* , J. Number Theory **30** (1988), 298–320.
- [Cole] A. Coleman, *The greatest mathematical paper of all time*, Math. Intelligencer **11** (1989), 29–38.
- [Coll] D. Collingwood, *Representations of rank one Lie groups*, Research Notes in Math., vol. 137, Pitman, Boston, 1985.
- [CoNo] J. Conway and S. Norton, *Monstrous moonshine*, Bull. London Math. Soc. **11** (1979), 308–339.
- [CoSl] J. Conway and N. Sloane, *Sphere packings, lattices and groups*, Grundlehren Math. Wiss., vol. 290, Springer-Verlag, New York, 1988.
- [Corn] J. Cornwell, *Group theory in physics. I, II*, Academic Press, London, 1984.
- [Corw1] L. Corwin, *The unitary dual for the multiplicative group of arbitrary division algebras over local fields*, J. Amer. Math. Soc. **2** (1989), 565–598.
- [Corw2] ———, *A construction of the supercuspidal representations of  $GL_n(F)$ ,  $F$   $p$ -adic*, preprint, 1989.

- [Corw3] ———, *A representation-theoretic criterion for local solvability of left invariant differential operators on nilpotent Lie groups*, Trans. Amer. Math. Soc. **264** (1981), 113–120.
- [CoGr] L. Corwin and F. Greenleaf, *Representations of nilpotent Lie groups and their applications, Part I: Basic theory and examples*, Cambridge Stud. Adv. Math., no. 18, Cambridge Univ. Press, Cambridge, 1990.
- [CoHR] L. Corwin, B. Helffer, and L. Rothschild, *Smoothness and analyticity of first order partial differential equations on nilpotent Lie groups*, Invent. Math. **81** (1985), 205–216.
- [CoHo] L. Corwin and R. Howe, *Computing characters of tamely ramified  $p$ -adic division algebras*, Pacific J. Math. **73** (1977), 461–478.
- [CoSz] L. Corwin and R. Szczerba, *Calculus in vector spaces*, Marcel Dekker, New York, 1979.
- [CoHH] M. Cowling, A. Haagerup, and R. Howe, *Almost  $L^2$ -matrix coefficients*, J. Reine Angew. Math. **387** (1988), 97–110.
- [Coxe] H. Coxeter, *Regular polytopes*, Dover, New York, 1973.
- [Cran1] L. Crane, *2-d physics and 3-d topology*, Comm. Math. Phys. **135** (1991), 615–640.
- [Cran2] ———, *Conformal field theory, spin geometry, and quantum gravity*, preprint, 1990.
- [Crou] P. Crouch, *Solvable approximations to control systems*, SIAM J. Control Optim. **22** (1984), 40–54.
- [Dani1] S. Dani, *On orbits of unipotent flows on homogeneous spaces*, Ergodic Theory Dynamical Systems **4** (1984), 25–34.
- [Dani2] ———, *Orbits of horospherical flows*, Duke Math. J. **53** (1986), 177–188.
- [Dani3] ———, *Bernoullian translations and minimal horospheres*, J. Indian Math. Soc. (N.S.) **40** (1976), 245–284.
- [DaMa1] S. Dani and G. Margulis, *Orbit closures of generic unipotent flows on homogeneous spaces of  $SL(3, \mathbf{R})$* , Math. Ann. **286** (1990), 101–128.
- [DaMa2] ———, *Values of quadratic forms at primitive integral points*, Invent. Math. **98** (1989), 405–424.
- [DJKM1] E. Date, M. Jimbo, M. Kashiwara, and T. Miwa, *Transformation groups for soliton equations*, Proc. R.I.M.S. Symposium (M. Jimbo and T. Miwa, eds.), World Scientific, 1983, pp. 39–120.
- [DJKM2] ———, *Landau, Lifshitz equations; solitons, quasi-periodic solutions and infinite dimensional Lie algebras*, J. Phys. A. **16** (1983), 221–236.
- [Daub] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure. Appl. Math. **41** (1988), 909–996.
- [DeGW] D. DeGeorge and N. Wallach, *Limit formulas for multiplicities in  $L^2(\Gamma/G)$* . I, II, Ann. of Math. (2) **107** (1978), 133–150; Ann. of Math. (2) **109** (1979), 477–495.
- [DLNT] P. Deift, L. Li, T. Nanda, and C. Tomei, *The Toda flow on a generic orbit is integrable*, Bull. Amer. Math. Soc. (N.S.) **11** (1984), 367–368.
- [DeNT] P. Deift, T. Nanda, and C. Tomei, *Differential equations for the symmetric eigenvalue problem*, SIAM J. Numer. Anal. **20** (1983), 1–27.
- [DeLu] P. Deligne and G. Lusztig, *Representations of reductive groups over finite fields*, Ann. of Math. (2) **103** (1976), 103–161.
- [Deva] R. Devaney, *An introduction to chaotic dynamical systems*, 2nd ed., Addison-Wesley, Redwood City, 1989.
- [Dick] L. Dickson, *A new system of simple groups*, Math. Ann. **60** (1905), 137–50.
- [Dixm1] J. Dixmier, *Enveloping algebras*, North-Holland, Amsterdam, 1977.
- [Dixm2] ———, *Les  $C^*$ -algèbres et leur représentations*, Gauthier-Villars, Paris, 1964.
- [Dixm3] ———, *L'application exponentielle dans les groupes de Lie résolubles*, Bull. Soc. Math. France **85** (1957), 113–121.
- [DiMa] J. Dixmier and P. Malliavin, *Factorisations de fonctions et de vecteurs indéfiniment différentiables*, Bull. Soc. Math. France **102** (1978), 305–330.
- [DIRi] R. Dlab and C. Ringel, *Indecomposable representations of graphs and algebras*, Mem. Amer. Math. Soc., vol. 173, Amer. Math. Soc., Providence, RI, 1976, pp. 1–57.
- [Dola] L. Dolan, *Why Kac-Moody subalgebras are interesting in physics*, Applications of Group Theory in Physics and Mathematical Physics, Lectures in Appl. Math., vol. 21, Amer. Math. Soc., Providence, RI, 1985.
- [Donal] S. Donaldson, *An application of gauge theory to the topology of 4-manifolds*, J. Differential Geom. **18** (1983), 269–311.



- [Drag1] A. Dragt, *Lie algebraic theory of geometrical optics and optical aberrations*, J. Opt. Soc. Amer. A **72** (1982), 372–379.
- [Drag2] A. Dragt et al., *Lie algebraic treatment of linear and nonlinear beam dynamics*, Ann. Rev. Nucl. Part. Sci. **38** (1988), 455–496.
- [Drin] V. Drinfeld, *Hopf algebras and the quantum Yang-Baxter equation*, Dokl. Akad. Nauk SSSR **283** (1985), 1060–1064.
- [Duf1] M. Duflo, *Caractères des groupes et des algèbres de Lie résolubles*, Ann. Sci. École Norm. Sup. (4) **3** (1970), 23–74.
- [Duf2] ———, *On the Plancherel formula for almost algebraic real Lie groups*, Lie Group Representations. III, Lecture Notes in Math., vol. 1077, Springer-Verlag, New York, 1984.
- [DuHV] R. Duflo, G. Heckman, and M. Vergne, *Projection d'orbites, formule de Kirillov et formule de Blattner*, Mem. Soc. Math. France (N.S.) **15** (1984), 65–128.
- [DuVe1] M. Duflo and M. Vergne, *La formule de Plancherel des groupes de Lie semi-simples Réels*, Representation of Lie Groups, Adv. Stud. Pure Math., vol. 14, Kyoto, Hiroshima, 1986, pp. 289–336.
- [DuVe2] ———, *Sur le foncteur de Zuckerman*, C. R. Acad. Sci. Paris Ser I. Math. **304** (1987), 467–469.
- [Dunk] C. Dunkl, *Differential difference operators associated to reflection groups*, Trans. Amer. Math. Soc. **311** (1989), 167–183.
- [DyMc] H. Dym and H. McKean, *Fourier series and integrals*, Academic Press, 1972.
- [Dyso1] F. Dyson, *Statistical theory of the energy levels of complex systems. I*, J. Math. Phys. **3** (1962), 140–156.
- [Dyso2] ———, *Missed opportunities*, Bull. Amer. Math. Soc. **78** (1972), 635–653.
- [Eins] A. Einstein, *Zur elektrodynamik bewegter Körpern*, Ann. Physik **17** (1905), 891.
- [Eins2] ———, *Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie*, Sitzungsberichte, Berlin, 1917, p. 42.
- [Emme] J. McL. Emmerson, *Symmetry principles in particle physics*, Clarendon Press, Oxford, 1972.
- [Engl] M. Englefield, *Group theory and the Coulomb problem*, Wiley-Interscience, 1972.
- [EnVa] T. Enright and V. Varadarajan, *On an infinitesimal characterization of the discrete series*, Ann. of Math. (2) **102** (1975), 1–15.
- [FePh] C. Fefferman and D. Phong, *The uncertainty principle and sharp Garding inequalities*, Comm. Pure Appl. Math. **34** (1981), 285–331.
- [FeFu] B. Feigin and F. Fuchs, *Skew symmetric invariant differential operators on a line and Verma modules over the Virasoro algebra*, Funktsional. Anal. i Prilozhen. **16** (1982), 47–63. (Russian)
- [FeFo] W. Feit and P. Fong, *Rational rigidity of  $G_2(p)$  for any prime  $p > 5$* , (Proc. Rutgers Group Theory 1983/84) (M. Aschbacher et al., eds.), Cambridge University Press, Cambridge, 1984, pp. 323–326.
- [Fell1] J. Fell, *Weak containment and induced representations of groups*, Canad. J. Math. **14** (1962), 237–268.
- [Fell2] ———, *Non-unitary dual spaces of groups*, Acta Math. **114** (1965), 267–310.
- [FeDo] J. Fell and R. Doran, *Representations of  $*$ -algebras, locally compact groups, and Banach  $*$ -algebraic bundles. I, II*, Mono. Pure and Appl. Math., vols. 125, 126, Academic Press, Boston, 1988.
- [Flas] H. Flaschka, *The Toda lattice. I, II*, Phys. Rev. B **9** (1974); Progr. Theoret. Phys. **51** (1975), 703–716.
- [Flat] D. Flath, *Decomposition of representations into tensor products*, Automorphic Forms, Representations and L-functions, Proc. Sympos. Pure Math., vol. 33, Part I, Amer. Math. Soc., Providence, RI, 1979.
- [FISZ] M. Flato, P. Sally, and G. Zuckerman, eds., *Applications of group theory in physics and mathematical physics*, Lectures in Appl. Math., vol. 21, Amer. Math. Soc., Providence, RI, 1985.
- [FJJe1] M. Flensted-Jensen, *Discrete series for semisimple symmetric spaces*, Ann. of Math. (2) **111** (1980), 253–311.

- [FJe2] ———, *Analysis on non-Riemannian symmetric spaces*, CBMS Regional Conf. Ser., no. 61, Amer. Math. Soc., Providence, RI, 1986.
- [FoMu] J. Fogarty and D. Mumford, *Geometric invariant theory*, 2nd corr. ed., *Ergeb. Math. Grenzgeb.*, no. 34, Springer-Verlag, Berlin and New York, 1982.
- [FoSt] G. Folland and E. Stein, *Estimates for the  $\bar{\partial}_b$  complex and analysis on the Heisenberg group*, *Comm. Pure. Appl. Math.* **27** (1974), 429–522.
- [Foll1] G. Folland, *Harmonic analysis in phase space*, *Ann. of Math. Stud.*, no. 123, Princeton Univ. Press, Princeton, NJ, 1989.
- [Foll2] ———, *Applications of analysis on nilpotent groups to partial differential equations*, *Bull. Amer. Math. Soc.* **83** (1977), 912–930.
- [FoGe] S. Fomin and I. Gelfand, *Geodesic flows on manifolds of constant negative curvature*, *Uspekhi Mat. Nauk.* **7** (1952), 118–137.
- [Foss] R. Fossum et al., eds., *Invariant theory*, *Contemp. Math.*, vol. 88, Amer. Math. Soc., Providence, RI, 1989.
- [FrUh] D. Freed and K. Uhlenbeck, *Instantons and four-manifolds*, *Math. Sci. Res. Inst. Publ.*, no. 1, Springer-Verlag, New York, 1984.
- [Fren] I. Frenkel, *Representations of affine Lie algebras, Hecke modular forms and Korteweg-deVries type equations*, *Lecture Notes in Math.*, vol. 933, Springer-Verlag, 1982, pp. 71–110.
- [FrKa] I. Frenkel and V. Kac, *Basic representations of affine Lie algebras and dual resonance models*, *Invent. Math.* **62** (1980), 23–66.
- [FrLM] I. Frenkel, J. Lepowsky, and A. Meurman, *Vertex operator algebras and the monster*, *Pure Appl. Math.*, vol. 134, Academic Press, San Diego, CA, 1988.
- [Frie] M. Fried, *Fields of definition of function fields and Hurwitz families—groups as Galois groups*, *Comm. Algebra* **5** (1977), 17–87.
- [Fron] C. Fronsdal, *Infinite multiplets and the hydrogen atom*, *Phys. Rev.* **156** (1967), 1665–1677.
- [Gaal] S. Gaal, *Linear analysis and representation theory*, *Grundlehren Math. Wiss.*, vol. 198, Springer-Verlag, Berlin, 1973.
- [Gabr] P. Gabriel, *Unzerlegbare Darstellungen. I*, *Manuscripta Math.* **6** (1972), 71–103.
- [GaVa] R. Gangolli and V. Varadarajan, *Harmonic analysis of spherical functions on real reductive groups*, *Ergeb. Math. Grenzgeb.* no. 101, Springer-Verlag, Berlin and New York, 1985.
- [GGKM] C. Gardner, J. Greene, M. Kruskal, and R. Miura, *Method for solving the Korteweg-deVries equation*, *Phys. Rev. Lett.* **19** (1967), 1095–1097.
- [Garl] H. Garland, *The arithmetic theory of loop algebras*, *J. Algebra* **53** (1978), 480–551.
- [GaLe] H. Garland and J. Lepowsky, *Lie algebra homology and the Macdonald-Kac formulas*, *Invent. Math.* **34** (1976), 37–76.
- [GaZu] H. Garland and G. Zuckerman, *On unitarizable highest weight modules of Hermitian pairs*, *J. Fac. Sci. Univ. Tokyo Sect. IA Math.* **28** (1982), 877–889.
- [GaGo] F. Garvan and G. Gonnet, *Macdonald's constant term conjectures for exceptional root systems*, *Bull. Amer. Math. Soc. (N.S.)* **24** (1991), 343–347.
- [Gawe] K. Gawedzki, *Conformal field theory*, *Sem. Bourbaki*, no. 704, pp. 177–178; *Asterisque*, 1989, pp. 95–126.
- [Gelb1] S. Gelbart, *Automorphic forms on adèle groups*, *Ann. of Math. Stud.*, no. 83, Princeton Univ. Press, Princeton, NJ, 1975.
- [Gelb2] ———, *Examples of dual reductive pairs*, *Automorphic Forms, Representations and L-functions*, *Proc. Symp. Pure Math.*, vol. 33, part 1, Amer. Math. Soc., Providence, RI, 1979.
- [Gelb3] ———, *An elementary introduction to the Langlands program*, *Bull. Amer. Math. Soc. (N.S.)* **10** (1984), 177–220.
- [GrPS] S. Gelbart and I. Piatetski-Shapiro, et al., *Explicit constructions of automorphic L-functions*, *Lecture Notes in Math.*, vol. 1254, Springer-Verlag, New York, 1987.
- [GeSh] S. Gelbart and F. Shahidi, *Analytic properties of automorphic L-functions*, *Perspect. Math.*, vol. 6, Academic Press, Boston, 1989.
- [Gelf] I. Gelfand, *The cohomology of infinite dimensional Lie algebras; some questions of integral geometry*, *Actes. Internat. Congr. Math. Nice*, vol. 1, 1970, pp. 95–111.
- [GGPS] I. Gelfand, M. Graev, and I. Piatetski-Shapiro, *Representation theory and automorphic functions*, Moscow, 1966.

- [GeNa] I. Gelfand and M. Naimark, *Unitäre Darstellungen Klassischer Gruppen*, Akademie Verlag, 1957.
- [GePo] I. Gelfand and V. Ponomarev, *Problems of linear algebra and classification of quadruples of subspaces in a finite dimensional vector space*, Colloq. Math. Soc. János Bolyai, vol. 5, North-Holland, Amsterdam, 1970, pp. 163-237.
- [Gell] D. Geller, *Analytic pseudodifferential operators for the Heisenberg group and local solvability*, Math. Notes, no. 37, Princeton Univ. Press, Princeton, NJ, 1990.
- [GeLa] P. Gerard and J.-P. Labesse, *The solution of a base change problem*, Automorphic Forms, Representation Theory, and  $L$ -Functions, Proc. Sympos. Pure Math., vol. 33, Amer. Math. Soc., Providence, RI, 1979, pp. 115-134.
- [Gilk] P. Gilkey, *Invariance theory, the heat equation, and the Atiyah-Singer Index Theorem*, Publish or Perish, Wilmington, DE, 1984.
- [Gilm] R. Gilmore, *Lie groups, Lie algebras, and some of their applications*, Wiley-Interscience, 1974.
- [Glea] A. Gleason, *Groups without small subgroups*, Ann. of Math. (2) **56** (1952), 193-212.
- [Gode1] R. Godement, *Sur les relations d'orthogonalité de V. Bargmann*, I, II, C. R. Acad. Sci. Paris **225** (1947), 521-523, 657-659.
- [Gode2] ———, *A theory of spherical functions*, I, Trans. Amer. Math. Soc. **73** (1952), 496-556.
- [Gold] S. Goldberg, *Curvature and homology*, Dover, New York, 1982.
- [GoSc] M. Golubitsky and D. Schaeffer, *Singularities and groups in bifurcation theory*, Appl. Math. Sci., vol. 51, Springer-Verlag, New York, 1985.
- [GoHJ] F. Goodman, P. de la Harpe, and V. Jones, *Coxeter graphs and towers of algebras*, Math. Sci. Res. Inst. Publ., no. 14, Springer-Verlag, New York, 1989.
- [GoWa] R. Goodman and N. Wallach, *Classical and quantum mechanical systems of Toda lattice type*, I, II, III, Comm. Math. Phys. **83** (1982), 355-386; **94** (1984), 177-217; **104** (1986), 473-509.
- [GoMP1] M. Goresky and R. MacPherson, *Intersection homology theory*, Topology **19** (1980), 135-162.
- [GoMP2] ———, *Intersection homology*, Invent. Math. **71** (1983), 77-129.
- [GrSW] M. Green, J. Schwarz, and E. Witten, *Superstring theory*, I, II, Cambridge Univ. Press, Cambridge, 1987.
- [GrSt] P. Greiner and E. Stein, *Estimates for the  $\bar{\partial}$ -Neumann problem*, Math. Notes, No. 19, Princeton Univ. Press, Princeton, NJ, 1977.
- [GrHa] P. Griffiths and J. Harris, *Principles of algebraic geometry*, Wiley-Interscience, New York, 1978.
- [GrMe] D. Gromoll and W. Meyer, *An exotic sphere of non-negative sectional curvature*, Ann. of Math. (2) **100** (1974), 401-406.
- [Grom] M. Gromov, *Soft and hard symplectic geometry*, Proc. Internat. Congr. Math. Berkeley I, Amer. Math. Soc., Providence, RI, 1987, pp. 81-98.
- [GrPS] M. Gromov and I. Piatetski-Shapiro, *Non-arithmetic groups in Lobachevsky spaces*, Inst. Haute Études Sci. Publ. Math. **66** (1988), 93-103.
- [GrRi1] K. Gross and D. St. Richards, *Total positivity, spherical series, and hypergeometric functions of matrix argument*, J. Approx. Theory **59** (1989), 224-246.
- [GrRi2] ———, *Spherical functions of matrix argument. I: Algebraic induction, zonal polynomials, and hypergeometric functions*, Trans. Amer. Math. Soc. **301** (1987), 781-811.
- [GrMP] A. Grossman, J. Morlet, and T. Paul, *Integral transforms associated to square integrable representations*, I, J. Math. Phys. **26** (1985), 2473-2479; II, Ann. Inst. H. Poincaré Phys. Théor. **45** (1986), 293-309.
- [GrBe] L. Grove and G. Benson, *Finite reflection groups*, 2nd ed., Graduate Texts in Math., vol. 99, Springer-Verlag, New York, 1985.
- [GrIa] B. Gruber and F. Iachello, *Symmetries in science*, III, Plenum Press, New York and London, 1989.
- [GuSt1] V. Guillemin and S. Sternberg, *Geometric asymptotics*, Math. Surveys, vol. 14, Amer. Math. Soc., Providence, RI, 1977.
- [GuSt2] ———, *The classification of the complex primitive infinite pseudogroups*, Proc. Nat. Acad. Sci. U.S.A. **55** (1966), 687-690.
- [GuSt3] ———, *Convexity properties of the moment mapping*, Invent. Math. **67** (1982), 491-513.

- [GuSt4] ———, *The moment map and collective motion*, Ann. Physics **127** (1980), 220–253.
- [GüRa] F. Gürsey and L. Radicati, Phys. Rev. Let **13** (1964), 299.
- [Gust] R. Gustafson, *A generalization of Selberg's beta integral*, Bull. Amer. Math. Soc. (N.S.) **22** (1990), 97–105.
- [GuMi1] R. Gustafson and S. Milne, *Schur functions and the invariant polynomials characterizing  $U(n)$  tensor operators*, Adv. in Appl. Math. **4** (1983), 422–478.
- [GuMi2] ———, *Schur functions, Good's identity and hypergeometric series well-poised in  $SU(n)$* , Adv. in Math. **48** (1983), 177–188.
- [GuMi3] ———, *A new symmetry for Beidenharn's  $G$ -functions and classical hypergeometric series*, Adv. in Math. **57** (1985), 209–225.
- [GuNi] M. Guterman and Z. Nitecki, *Differential equations, a first course*, Saunders College Publ., 1984.
- [Habs] L. Habsieger, *La  $q$ -Macdonald-Morris pour  $G_2$* , C. R. Acad. Sci. Paris Sér. I Math. **303** (1986), 211–213.
- [Halm] P. Halmos, *Lectures on ergodic theory*, Chelsea, New York, 1958.
- [Hame] M. Hamermesh, *Group theory and its applications to physical problems*, Dover, New York, 1989.
- [HaCh0] Harish-Chandra, *Infinite irreducible representations of the Lorentz group*, Proc. Roy. Soc. London Ser. A **189** (1947), 372–401.
- [HaCh1] ———, *On some applications of the universal enveloping algebra of a semisimple Lie algebra*, Trans. Amer. Math. Soc. **70** (1951), 28–96.
- [HaCh2] ———, *Representations of a semisimple Lie group on a Banach space*, Trans. Amer. Math. Soc. **75** (1953), 185–243.
- [HaCh3] ———, *Representations of semisimple Lie groups. II, III*, Trans. Amer. Math. Soc. **76** (1954), 26–65; 234–253.
- [HaCh4] ———, *Plancherel formula for the  $2 \times 2$  real unimodular group*, Proc. Nat. Acad. Sci. U.S.A. **38** (1952), 337–342.
- [HaCh5] ———, *The Plancherel formula for complex semisimple Lie groups*, Trans. Amer. Math. Soc. **76** (1954), 485–528.
- [HaCh6] ———, *The characters of semisimple Lie groups*, Trans. Amer. Math. Soc. **83** (1956), 98–163.
- [HaCh7] ———, *On a lemma of F. Bruhat*, J. Math. Pures Appl. **35** (1956), 203–210.
- [HaCh8] ———, *Differential operators on a semisimple Lie algebra*, Amer. J. Math. **79** (1957), 87–120.
- [HaCh9] ———, *Fourier transforms on a semisimple Lie group. I*, Amer. J. Math. **79** (1957), 193–257.
- [HaCh10] ———, *Spherical functions on a semisimple Lie group. I*, Amer. J. Math. **80** (1958), 241–310.
- [HaCh11] ———, *Spherical functions on a semisimple Lie group. II*, Amer. J. Math. **80** (1958), 553–613.
- [HaCh12] ———, *Automorphic forms on a semisimple Lie group*, Proc. Nat. Acad. Sci. U.S.A. **45** (1959), 570–573.
- [HaCh13] ———, *Some results on differential equations*, Harish-Chandra's Collected Works, vol. 3, Springer-Verlag, New York, 1984, pp. 7–48.
- [HaCh14] ———, *Invariant distributions on Lie algebras*, Amer. J. Math. **86** (1964), 271–309.
- [HaCh15] ———, *Invariant differential operators and distributions on a semi-simple Lie algebra*, Amer. J. Math. **86** (1964), 534–564.
- [HaCh16] ———, *Some results on an invariant integral on a semisimple Lie algebra*, Ann. of Math. (2) **80** (1964), 551–593.
- [HaCh17] ———, *Invariant eigendistributions on a semisimple Lie algebra*, Inst. Hautes Études Sci. Publ. Math. **27** (1965), 5–54.
- [HaCh18] ———, *Invariant distributions on a semisimple Lie group*, Trans. Amer. Math. Soc. **119** (1965), 457–508.
- [HaCh19] ———, *Discrete series for semisimple Lie groups. I. Construction of invariant eigendistributions*, Acta Math. **116** (1965), 241–318.
- [HaCh20] ———, *Discrete series for semisimple Lie groups. II. Explicit determination of the characters*, Acta Math. **116** (1966), 1–111.

- [HaCh 21] ———, *Automorphic forms on semisimple Lie groups*, Notes by J. G. Mars, Lecture Notes in Math., vol. 62, Springer-Verlag, Berlin, 1968.
- [HaCh22] ———, *Harmonic analysis on real reductive groups. III. The Maass-Selberg relations and the Plancherel formula*, Ann. of Math. (2) **104** (1976), 117–201.
- [HaVa] P. de la Harpe and A. Valette, *La propriété (T) de Kazhdan pour les groupes localement compacts*, Astérisque, no. 175, Soc. Math. France, Paris, 1989.
- [Harr] M. Harris, *Automorphic forms of  $\bar{\partial}$ -cohomology type as coherent cohomology classes*, J. Differential Geom. **32** (1990), 1–63.
- [Hart] R. Hartshorne, *Algebraic geometry*, Graduate Texts Math., vol. 52, Springer-Verlag, New York, 1977.
- [HaSc] M. Hausner and J. Schwartz, *Lie groups; Lie algebras*, Gordon and Breach, New York, 1968.
- [Hawk1] T. Hawkins, *Elie Cartan and the prehistory of the representation theory of Lie algebras*, unpublished research report.
- [Hawk2] ———, *Geometry, differential equations, and the birth of Lie's theory of groups*, Frank J. Hahn Lectures, Yale University, April, 1990.
- [Hawk3] ———, *Non-Euclidean geometry and Weierstrassian mathematics: the background to Killing's work on Lie algebras*, Studies in the History of Mathematics, MAA Stud. Math., no. 26, Math. Assoc. Amer., Washington, D.C., 1987, pp. 21–36.
- [Hawk4] ———, *Hesse's principle of transfer and the representation of Lie algebras*, Arch. Hist. Exact. Sci. **39** (1988), 41–73.
- [Haze1] M. Hazewinkel, *On deformations, approximations and nonlinear filtering*, Systems Control Lett. **1** (1981/82), 32–36.
- [Haze2] ———, *On Lie algebras of vectorfields, Lie algebras of differential operators and nonlinear filtering*, Geometry Symposium, Utrecht 1980, Lecture Notes in Math., vol. 844, Springer-Verlag, Berlin, 1981, pp. 91–106.
- [HaMa] M. Hazewinkel and S. Marcus, *On Lie algebras and finite dimensional filtering*, Stochastics **7** (1982), 29–62.
- [HeSc1] H. Hecht and W. Schmid, *A proof of Blattner's conjecture*, Invent. Math. **31** (1975), 129–154.
- [HeSc2] ———, *Characters, asymptotics and  $n$ -homology of Harish-Chandra modules*, Acta. Math. **151** (1983), 49–151.
- [HMSW] H. Hecht, D. Milicic, W. Schmid, and J. Wolf, *Localization and standard modules for semisimple Lie groups. I*, Invent. Math. **90** (1987), 297–332.
- [Hecke] E. Hecke, *Modulfunktionen und die Dirichletschen Reihen mit Eulerscher Produktentwicklung. I, II*, Math. Ann. **114** (1937), 1–28; 316–351.
- [Heck1] G. Heckman, *Root systems and hypergeometric functions. II*, Compositio Math. **64** (1987), 353–373.
- [Heck2] ———, *An elementary approach to the hypergeometric shift operators of Opdam*, Invent. Math. (to appear).
- [HeOp] G. Heckman and E. Opdam, *Root systems and hypergeometric functions. I*, Compositio Math. **64** (1987), 329–352.
- [Hed1] G. Hedlund, *On the metrical transitivity of the geodesics on closed surfaces of constant negative curvature*, Amer. J. Math. **35** (1934), 787–858.
- [Hed2] ———, *The dynamics of geodesic flows*, Bull. Amer. Math. Soc. **45** (1939), 241–260.
- [Held] A. Held, ed., *General relativity and gravitation*, Plenum Press, New York and London, 1980.
- [HeNo] B. Helffer and J. Nourrigat, *Hypoéllipticité maximale pour des opérateurs polynômes de champs de vecteurs*, Progr. Math., vol. 58, Birkhäuser, Boston, 1985.
- [Helg1] S. Helgason, *Groups and geometric analysis*, Pure Appl. Math., vol. 113, Academic Press, Orlando, FL, 1984.
- [Helg2] ———, *Differential geometry and symmetric spaces*, Academic Press, New York, 1962.
- [Helg3] ———, *A formula for the radial part of the Laplace-Beltrami operator*, J. Differential Geom. **6** (1972), 411–419.
- [Helg4] ———, *Analysis on Lie groups and homogeneous spaces*, CBMS Regional Conf. Ser. in Math., vol. 14, Amer. Math. Soc., Providence, RI, 1972.

- [Henn1] G. Henniart, *Les conjectures de Langlands locales pour  $GL(n)$* , Journées Arithmétiques de Metz, Asterisque, no. 94, Soc. Math. France, Paris, 1982, pp. 67–85.
- [Henn2] ———, *Représentations des groupes réductifs  $p$ -adiques*, Exposé, Sémin. Bourbaki, no. 736, 1990–1991.
- [Henn3] ———, *La conjecture de Langlands locale pour  $GL(3)$* , I.H.E.S. Notes, 1980.
- [Herm1] H. Hermes, *Nilpotent approximations of control systems and distributions*, SIAM J. Control Optim. **24** (1986), 731–736.
- [Herm2] ———, *Control systems which generate decomposable Lie algebras*, J. Differential Equations **44** (1982), 166–187.
- [HeRo] E. Hewitt and K. Ross, *Abstract harmonic analysis. II*, Grundlehren Math. Wiss., vol. 152, Springer-Verlag, Berlin, 1970.
- [Hilb1] D. Hilbert, *Über die Theorie der algebraischen Formen*, Math. Ann. **36** (1890), 473–534; *Gesammelte Abhandlungen*, Bd. II, Springer-Verlag, Berlin, 1933, pp. 199–257.
- [Hilb2] ———, *Über die vollen Invariantensysteme*, Math. Ann. **42** (1893), 313–373; *Gesammelte Abhandlungen*, Bd. II, Springer-Verlag, Berlin, 1933, pp. 287–344.
- [HiHL] J. Hilgert, K. Hofmann, and J. Lawson, *Lie groups, convex cones and semigroups*, Clarendon Press, Oxford, 1989.
- [Hill] R. Hill, Jr., *Elementary linear algebra*, Academic Press (College Division), Orlando, FL, 1986.
- [Hilr] H. Hiller, *Geometry of Coxeter groups*, Res. Notes in Math. vol. 54, Pitman Advanced Publishing Program, Boston, 1982.
- [Hirz] F. Hirzebruch, *Topological methods in algebraic geometry*, Springer-Verlag, New York, 1966.
- [Hirz1] F. Hirzebruch, *Über vierdimensionale Riemannsche Flächen mehrdeutiger analytischen Funktionen von zwei komplexen Veränderlichen*, Math. Ann. **126** (1953), 1–22.
- [HiZa] F. Hirzebruch and D. Zagier, *Intersection numbers of curves on Hilbert modular surfaces and modular forms of Nebentypus*, Invent. Math. **36** (1976), 57–113.
- [Hoch] G. Hochschild, *The structure of Lie groups*, Holden-Day, San Francisco, 1965.
- [HoPe] W. Hodge and D. Pedoe, *Methods of algebraic geometry. I, II, III*, Cambridge Univ. Press, Cambridge, 1947, 1952, 1954.
- [Hopf1] E. Hopf, *Fuchsian groups and ergodic theory*, Trans. Amer. Math. Soc. **39** (1936), 299–314.
- [Hopf2] H. Hopf, *Ein topologischer Beitrag zur reellen Algebra*, Comment Math. Helv. **13** (1940), 219–239.
- [Hopf3] ———, *Über die Topologie der Gruppen-Mannigfaltigkeiten und ihre Verallgemeinerungen*, Ann. of Math. **42** (1941), 22–52.
- [HoSa] H. Hopf and H. Samuelson, *Ein Satz über die Wirkungsräume geschlossener Liescher Gruppen*, Comment Math. Helv. **13** (1940), 240–251.
- [Hörm] L. Hörmander, *The analysis of linear partial differential operators. I–IV*, Grundlehren Math. Wiss., Springer-Verlag, New York, vols. 256–257, 274–275, 1983–1984.
- [Hott] R. Hotta, *On realization of the discrete series for semi-simple Lie groups*, J. Math. Soc. Japan **23** (1971), 384–407.
- [HoPa] R. Hotta and R. Parthasarathy, *Multiplicity formula for discrete series*, Invent. Math. **26** (1974), 133–178.
- [HoWa] R. Hotta and N. Wallach, *On Matsushima's formula for the Betti numbers of a locally symmetric space*, Osaka J. Math. **12** (1975), 419–431.
- [Howe1] R. Howe, *Remarks on classical invariant theory*, Trans. Amer. Math. Soc. **313** (1989), 539–570.
- [Howe2] ———, *Quantum mechanics and partial differential equations*, J. Funct. Anal. **38** (1980), 188–254.
- [Howe3] ———, *The oscillator semigroup*, The Mathematical Heritage of Hermann Weyl (R. Wells, ed.), Amer. Math. Soc., Providence, RI, 1988, pp. 61–132.
- [Howe4] ———, *On the role of the Heisenberg group in harmonic analysis*, Bull. Amer. Math. Soc. (N.S.) **3** (1980), 821–843.
- [Howe5] ———,  *$\theta$ -series and invariant theory*, Automorphic Forms, Representations and  $L$ -Functions, Proc. Sympos. Pure Math., vol. 33, part 1, Amer. Math. Soc., Providence, RI, 1979, pp. 275–285.

- [Howe6] ———, *Dual pairs in physics: harmonic oscillators, photons, electrons and singletons*, Lectures in Appl. Math., vol. 21, Amer. Math. Soc., Providence, RI, 1985, pp. 179–207.
- [Howe7] ———, *Very basic Lie theory*, Amer. Math. Monthly **90** (1983), 600–623.
- [Howe8] ———, *The classical groups and invariants of binary forms*, The Mathematical Heritage of Hermann Weyl, Proc. Symp. Pure Math., vol. 48 (R. Wells, ed.), Amer. Math. Soc., Providence, RI, 1988, pp. 133–166.
- [Howe9] ———,  $(GL_n, GL_m)$ -duality and symmetric plethysm, Proc. Indian Acad. Sci. Math. Sci. **97** (1987), 85–109.
- [HoMo] R. Howe and C. Moore, *Asymptotic properties of unitary representations*, J. Funct. Anal. **32** (1979), 72–96.
- [HoMy1] R. Howe and A. Moy, *Harish-Chandra homomorphisms for  $p$ -adic groups*, CBMS Regional Conf. Ser. in Math., vol. 59, Amer. Math. Soc., Providence RI, 1985.
- [HoMy2] ———, *Hecke algebra isomorphisms for  $GL(n)$  over a  $p$ -adic field*, J. Algebra **131** (1990), 388–424.
- [HoPS1] R. Howe and I. Piatetski-Shapiro, *A counterexample to the “generalized Ramanujan conjecture” for (quasi-) split groups*, Automorphic Forms, Representations and  $L$ -Functions, Proc. Sympos. Pure Math., vol. 33, part 1, Amer. Math. Soc., Providence, RI, 1979.
- [HoPS2] ———, *Some examples of automorphic forms on  $Sp(4)$* , Duke Math. J. **50** (1983), 55–106.
- [HoTa] R. Howe and E. Tan, *Nonabelian harmonic analysis: Applications of  $SL_2(\mathbf{R})$* , Springer-Verlag, New York (to appear).
- [HoLe] R. Howlett and G. Lehrer, *Induced cuspidal representations and generalized Hecke rings*, Invent. Math. **58** (1980), 37–64.
- [Hu] S.-T. Hu, *Homotopy theory*, Pure Appl. Math., vol. 8, Academic Press, New York, 1959.
- [Huan] J.-S. Huang, *The unitary dual of the universal covering group of  $GL(n, \mathbf{R})$* , Duke Math. J. **61** (1990), 705–745.
- [Hump] J. Humphreys, *Introduction to Lie algebras and representation theory*, Graduate Texts in Math., vol. 9, Springer-Verlag, New York, 1972.
- [Hump2] J. Humphreys, *Arithmetic groups*, Lecture Notes in Math., vol. 789, Springer-Verlag, Berlin, 1980.
- [Huse] D. Husemoller, *Fibre bundles*, McGraw-Hill, New York, 1966.
- [Ibra] N. Ibragimov w, *Transformation groups applied to mathematical physics*, Reidel, Boston, 1985.
- [Igus] J.-I. Igusa, *Theta functions*, Grundlehren Math. Wiss., vol. 194, Springer-Verlag, New York, 1972.
- [Ikeda] A. Ikeda, *On spherical space forms which are isospectral but not isometric*, J. Math. Soc. Japan **35** (1983), 437–444.
- [Iken] G. Ikenberry, *Quantum mechanics, for mathematicians and physicists*, Oxford Univ. Press, New York, 1982.

The items below labelled [ITGTx] refer to the proceedings of sessions of the International Colloquium on Group Theoretical Methods in Physics, which have been convened annually since 1972, in the diverse locations indicated.

- [ITGTI] Marseilles, Joint report of Univ. de Provence, Univ. d'Aix-Marseilles and CNRS, Marseilles, 1972.
- [ITGTII] Nijmegen, Faculty of Science, Univ. of Nijmegen, 1973.
- [ITGTIII] Marseilles, Centre de Physique Theorique, Marseilles, 1974.
- [ITGTIV] Nijmegen (A. Janner, J. Janssen, and M. Boone, eds.), Lecture Notes in Phys., vol. 50, Springer-Verlag, Berlin, 1975.
- [ITGTV] Montreal 1976 (R. Sharp and B. Kolman, eds.), Academic Press, 1977.
- [ITGTVI] Tübingen 1977 (P. Kramer and A. Riecker, eds.), Springer Lecture Notes in Phys., vol. 79, Springer-Verlag, 1978.
- [ITGTVII] Austin 1978 (W. Beiglböck, A. Böhm, and E. Takaseigi, eds.), Springer Lecture Notes in Phys., vol. 94, Springer-Verlag, 1979.
- [ITGTVIII] Kiryat Anavim 1979 (L. Horowitz and Y. Neeman, eds.), Ann. Israel Phys. Soc., vol. 3, Adam Hilger, Bristol, England, 1980.

- [ITGTIX] Cocoyon, Mexico 1980 (K. Wolf, ed.), *Lecture Notes in Phys.*, vol. 135, Springer-Verlag, 1981.
- [ITGTXI] Istanbul (M. Serdaroglu and E. Inonu, eds.), 1982.
- [ITGTXII] Trieste 1983 (G. Denardo, G. Ghirardi, and T. Weber, eds.).
- [ITGTXVII] Saint-Adele, Canada 1988 (Y. Saint-Aubin and L. Vinet, eds.), World Scientific, Singapore, 1988.
- [Iwah1] N. Iwahori, *On the structure of a Hecke ring of a Chevalley group over a finite field*, J. Fac. Sci. Univ. Tokyo **10** (1964), 215–236.
- [Iwah2] N. Iwahori, *Generalized Tits system (Bruhat decomposition) on  $p$ -adic semisimple groups*, Algebraic Groups and Discontinuous Subgroups, Proc. Sympos. Pure Math., vol. 9, Amer. Math. Soc., Providence, RI, 1966, pp. 71–83.
- [IwMa] N. Iwahori and H. Matsumoto, *On some Bruhat decomposition and the structure of the Hecke rings of  $p$ -adic Chevalley groups*, Inst. Hautes Études Sci. Publ. Math. **25** (1965), 5–48.
- [Jaco1] N. Jacobson, *Lie algebras*, Wiley-Interscience, New York, 1962.
- [Jaco2] ———, *Basic algebra*. I, II, Freeman, New York, 1980.
- [JaLa] H. Jacquet and R. Langlands, *Automorphic forms on  $GL(2)$* , Lecture Notes in Math., vol. 114, Springer-Verlag, New York, 1970.
- [JaVe] H. Jakobson and M. Vergne, *Wave and Dirac operators and representations of the conformal group*, J. Funct. Anal. **24** (1977), 52–106.
- [Jame] A. James, *Distributions of matrix varieties and latent roots derived from normal samples*, Ann. Math. Statist. **35** (1964), 475–501.
- [Jamm] M. Jammer, *The conceptual development of quantum mechanics*, Internat. Ser. Pure Appl. Phys., McGraw-Hill, New York, 1966.
- [Jant] J. Jantzen, *Representations of algebraic groups*, Pure Appl. Math., vol. 131, Academic Press, Orlando, FL, 1987.
- [Jimbo] M. Jimbo, *A  $q$ -difference analogue of  $U(\mathfrak{g})$  and the Yang-Baxter equation*, Lett. Math. Phys. **10** (1985), 63–69.
- [JiMi1] M. Jimbo and T. Miwa, *Solitons and infinite dimensional Lie algebras*, Publ. Res. Inst. Math. Sci. **19** (1983), 943–1001.
- [Jone] H. Jones, *Groups, representations and physics*, Adam Hilger, Bristol and New York, 1990.
- [KacM] M. Kac, *Can one hear the shape of a drum?*, Amer. Math. Monthly **73** (1966), 1–23.
- [Kac1] V.G. Kac, *Infinite-dimensional Lie algebras*, 2nd ed., Cambridge Univ. Press, Cambridge, 1985.
- [Kac2] ———, *Simple graded Lie algebras of finite growth*, Functional Anal. Appl. **1** (1967), 328–329.
- [Kac3] ———, *Simple irreducible graded Lie algebras of finite growth*, Math. USSR-Izv. **2** (1968), 1271–1311.
- [Kac4] ———, *Infinite-dimensional algebras, Dedekind's  $\eta$ -function*, Functional Anal. Appl. **8** (1974), 68–70.
- [Kac5] ———, *Infinite-dimensional algebras, Dedekind's  $\eta$ -function, classical Möbius function and the very strange formula*, Adv. in Math. **30** (1978), 85–136.
- [Kac6] ———, *Infinite root systems, representations of graphs and invariant theory*, Invent Math. **56** (1980), 57–92.
- [Kac7] ———, *Infinite root systems, representations of graphs, and invariant theory*. II, J. Algebra **78** (1982), 141–162.
- [KaPe] V. G. Kac and D. Peterson, *Infinite dimensional Lie algebras, theta functions and modular forms*, Adv. in Math. **53** (1984), 125–264.
- [Kant1] I. Kantor, *Simple graded infinite-dimensional Lie algebras*, Dokl. Akad. Nauk SSSR **179** (1968), 534–537; English transl., Soviet Math. Dokl. **9** (1968), 409–412.
- [Kant2] ———, *Graded Lie algebras*, Trudy Sem. Vektor Tenzor. Anal. **15** (1970), 227–266. (Russian)
- [Kap1] I. Kaplansky, *Lie algebras and locally compact groups*, Univ. of Chicago Press, Chicago, IL, 1971.
- [Kap2] ———, *An introduction to differential algebra*, Publ. Inst. Math. Univ. Nancago **5**, Actualités Sci. Indust., vol. 125, Hermann, Paris, 1957.



- [Kash] M. Kashiwara, *Crystallizing the  $q$ -analogue of universal enveloping algebras*, preprint.
- [KaKK] M. Kashiwara, T. Kawai, and T. Kimura, *Foundations of algebraic analysis*, Princeton Univ. Press, Princeton, NJ, 1986.
- [Katz] Y. Katznelson, *An introduction to harmonic analysis*, Dover, New York, 1976.
- [KaLu1] D. Kazhdan and G. Lusztig, *Representations of Coxeter groups and Hecke algebras*, Invent. Math. **53** (1979), 165–184.
- [KaLu2] Y. Kazhdan and G. Lusztig, *Proof of the Deligne-Langlands conjecture for Hecke algebras*, Invent. Math. **87** (1987), 153–215.
- [Kell] J. Kelley, *General topology*, van Nostrand, Princeton, NJ, 1955.
- [KeNa] J. Kelley and I. Namioka et al., *Linear topological spaces*, van Nostrand, Princeton, NJ, 1963.
- [Kill] W. Killing, *Die Zusammensetzung der stetigen, endlichen, Transformationsgruppen*. I, II, III, IV, Math. Ann. **31** (1888), 252–290; **33** (1888), 1–48; **34** (1889), 57–122; **36** (1890) 161–189.
- [Kiri] A. Kirillov, *Representations of nilpotent Lie groups*, Russian Math. Surveys **17** (1962), 53–104.
- [Kiri2] ———, *Elements of the theory of representations*, Grundlehren Math. Wiss., vol. 220, Springer-Verlag, Berlin, 1976.
- [Kirw1] F. Kirwan, *Cohomology of quotients in symplectic and algebraic geometry*, Math. Notes, no. 31, Princeton Univ. Press, Princeton, NJ, 1984.
- [Kirw2] ———, *Convexity properties of the moment mapping*. III, Invent. Math. **77** (1984), 547–552.
- [Kirw3] ———, *An introduction to intersection homology theory*, Pitman Res. Notes Math. Ser., vol. 187, Longman Sci. Tech., Harlow, 1988.
- [KILa] S. Kleiman and D. Laksov, *Schubert calculus*, Amer. Math. Monthly **79** (1972), 1061–1082.
- [Klei1] F. Klein, *The icosahedron and the solution of equations of the fifth degree*, Dover, New York, 1956.
- [Klei2] ———, *Vergleichende Betrachtungen über neuere geometrische Forschungen* (1872), Mathematische Abhandlungen, Springer-Verlag, Berlin, 1921.
- [Knapp1] A. Knapp, *Lie groups, Lie algebras, and cohomology*, Math. Notes, vol. 34, Princeton Univ. Press, Princeton, NJ, 1988.
- [Knapp2] ———, *Representation theory of semisimple groups; An overview based on examples*, Princeton Univ. Press, Princeton, NJ, 1986.
- [KnVo] A. Knapp and D. Vogan, *Duality theorems in relative Lie algebra cohomology*, duplicated notes.
- [KnZu] A. Knapp and G. Zuckerman, *Classification of irreducible tempered representations of semisimple groups*, Ann. of Math. (2) **116** (1982), 389–501.
- [KoZi] H. Koch and E. Zink, *Zur Korrespondenz von Darstellungen der Galois-gruppen und der zentralen Divisionsalgebren über lokalen Körpern (der zahme Fall)*, Akad. Wissenschaften der DDR, Zentral Institut für Mathematik und Mechanik, Report R-D3/79 Berlin, 1979.
- [Koda] K. Kodaira, *On compact analytic surfaces*. II, Ann. of Math. (2) **77** (1963), 563–626.
- [KoTe] M. Koike and I. Terada, *Young-diagrammatic methods for the representation theory of the classical groups of type  $B_n$ ,  $C_n$ ,  $D_n$* , J. Algebra **107** (1987), 466–511.
- [Kolc] E. Kolchin, *Differential algebraic groups*, Mono. Pure Appl. Math., vol. 114, Academic Press, Orlando, FL, 1985.
- [Kolm] B. Kolman, ed., *A survey of Lie groups and Lie algebras with applications and computational methods*, SIAM, Philadelphia, PA, 1972.
- [Körn] T. Korner, *Fourier analysis*, Cambridge Univ. Press, Cambridge, 1988.
- [Kost1] B. Kostant, *Quantization and unitary representations, Part I, Prequantization*, Lecture Notes in Math., vol. 170, Springer-Verlag, New York, 1970, pp. 87–208.
- [Kost2] ———, *The solution to a generalized Toda lattice and representation theory*, Adv. in Math. **34** (1979), 195–338.
- [Kost3] ———, *On Macdonald's  $\eta$ -function formula, the Laplacian and generalized exponents*, Adv. in Math. **20** (1976), 179–212.
- [Kost4] ———, *Lie algebra cohomology and the generalized Borel-Weil theorem*, Ann. of Math. (2) **74** (1961), 329–387.

- [Kost5] ———, *A formula for the multiplicity of a weight*, Trans. Amer. Math. Soc. **93** (1959), 53–73.
- [Kost6] ———, *Orbits, symplectic structures, and representation theory*, Proc. U.S.-Japan Sem. on Differential Geometry, Kyoto, 1965.
- [Kstk] C. Kostka, *Über den Zusammenhang zwischen einigen Formen von symmetrischen Funktionen*, Crelle's J. **93** (1882), 89–123.
- [KuMi1] S. Kudla and J. Millson, *Intersection numbers of cycles on locally symmetric spaces and Fourier coefficients of holomorphic modular forms in several complex variables*, Inst. Hautes Études Sci. Publ. Math. **71** (1990), 121–172.
- [KuMi2] ———, *The theta correspondence and harmonic forms. I*, Math. Ann. **274** (1986), 353–378.
- [KuMi3] ———, *The theta correspondence and harmonic forms. II*, Math. Ann. **277** (1987), 267–314.
- [Kuma] S. Kumaresan, *On the canonical  $\mathfrak{k}$ -types in the irreducible unitary  $\mathfrak{g}$ -modules with non-zero relative cohomology*, Invent. Math. **59** (1980), 1–11.
- [Kuro] N. Kurokawa, *Examples of eigenvalues of Hecke operators on Siegel cusp forms of degree 2*, Invent. Math. **49** (1978), 149–165.
- [Kutz1] P. Kutzko, *On the supercuspidal representations of  $GL_2$ . I, II*, Amer. J. Math. **100** (1978), 43–60, 705–716.
- [Kutz2] ———, *The Langlands conjecture for  $GL_2$  of a local field*, Ann. of Math. (2) **112** (1980), 381–412.
- [KuMo] P. Kutzko and A. Moy, *On the local Langlands conjecture in prime dimension*, Ann. of Math. (2) **121** (1985), 495–517.
- [Labe] J.-P. Labesse, *The present state of the trace formula*, Automorphic Forms, Shimura Varieties and  $L$ -Functions I (L. Clozel and J. Milne, eds.), Perspect. Math., vol. 10, Academic Press, Boston, 1990.
- [LaLa] J.-P. Labesse and R. Langlands,  *$L$ -indistinguishability for  $SL(2)$* , Canad. J. Math. **31** (1979), 726–785.
- [LaSe1] V. Lakshmibai and C. Seshadri, *Geometry of  $G/P$ . II*, Proc. Indian Acad. Sci. Math. Sci. **87A** (1978), 1–54.
- [LaSe2] ———, *Geometry of  $G/P$ . V*, J. Algebra **100** (1986), 461–557.
- [LaMS] V. Lakshmibai, C. Musili, and C. Seshadri, *Geometry of  $G/P$ . III, IV*, Proc. Indian Acad. Sci. Math. Sci. **88A** (1978), 93–177; Bull. Amer. Math. Soc. (N.S.) **1** (1979), 279–362.
- [Lamo] K. Lamotke, *Solids and isolated singularities*, F. Vieweg & Sohn, Braunschweig/Wiesbaden.
- [Lang1] S. Lang,  $SL_2(\mathbb{R})$ , Addison-Wesley, Reading, MA, 1975.
- [Lang2] ———, *Real analysis*, 2nd ed., Addison-Wesley, Reading, MA, 1983.
- [Lang3] ———, *Algebra*, 2nd ed., Addison-Wesley, Menlo Park, 1984.
- [Lang4] ———, *Algebraic number theory*, Graduate Texts in Math., vol. 110, Springer-Verlag, New York, 1986.
- [Lang5] ———, *Introduction to modular forms*, Grundlehren Math. Wiss., vol. 222, Springer-Verlag, New York, 1976.
- [Lgl1] R. Langlands, *On the functional equations satisfied by Eisenstein series*, Lecture Notes in Math., vol. 544, Springer-Verlag, New York, 1976.
- [Lgl2] ———, *Euler products*, Yale Univ. Press, 1967.
- [Lgl3] ———, *Problems in the theory of automorphic forms*, Lectures in Modern Analysis and Applications, Lecture Notes in Math., vol. 170, Springer-Verlag, New York, pp. 18–86.
- [Lgl4] ———, *On the classification of irreducible representations of real algebraic groups*, preprint.
- [Lgl5] ———, *Base change for  $GL_2$ : The theory of Saito-Shintani with applications*, Notes, Institute for Advanced Study, Princeton, NJ, 1975.
- [Lgl6] ———, *Automorphic representations, Shimura varieties and motives, Ein Märchen*, Automorphic Forms, Representations, and  $L$ -Functions, Proc. Sympos. Pure Math. vol. 33, part 2, Amer. Math. Soc., Providence, RI, 1979, pp. 205–246.
- [Lgl7] ———, *Dimensions of spaces of automorphic forms*, Algebraic Groups and Discontinuous Subgroups, Proc. Sympos. Pure Math., vol. 9, Amer. Math. Soc., Providence, RI, 1966.

- [Lgld8] ———, *Les débuts d'une formule des traces stable*, Publ. Math. Univ. Paris VII, U.E.R. de Math., Tour 45–55, Paris.
- [Lgld9] ———, *Eisenstein series, the trace formula, and the modern theory of automorphic forms*, Number Theory, Trace Formulas, and Discrete Groups (Aubert et al., eds.), Academic Press, Boston, 1989, pp. 125–155.
- [LaSh] R. Langlands and D. Shelstad, *On the definition of transfer factors*, Math. Ann. **278** (1989), 219–271.
- [LaBe] A. Lascoux and M. Berger, *Variétés Kähleriennes Compactes*, Lecture Notes in Math., vol. 154, Springer-Verlag, Berlin, 1970.
- [Lau] H. Laufer, *Normal two-dimensional singularities*, Ann. of Math. Stud., no. 71, Princeton Univ. Press, Princeton, NJ, 1971.
- [Lepo1] J. Lepowsky, *Lie algebras and combinatorics*, Proc. Internat. Congr. Math., Helsinki, 1978.
- [Lepo2] ———, *A Lie theoretic interpretation and proof of the Rogers-Ramanujan identities*, Adv. Math. **45** (1982), 21–72.
- [Lepo3] ———, *Algebraic results on representations of semisimple Lie groups*, Trans. Amer. Math. Soc. **176** (1973), 1–44.
- [LeMi] J. Lepowsky and S. Milne, *Lie algebraic approaches to classical partition identities*, Adv. in Math. **29** (1978), 15–59.
- [Lewy] H. Lewy, *An example of a smooth linear partial differential equation without solution*, Ann. of Math. (2) **66** (1957), 155–158.
- [Lich] D. Lichenberg, *Unitary symmetry and elementary particles*, 2nd ed., Academic Press, New York, 1978.
- [LiEn] S. Lie and F. Engel, *Theorie der Transformationsgruppen*. I, II, III, Teubner, Leipzig, 1888, 1890, 1893.
- [LiVe] G. Lion and M. Vergne, *The Weil representation, Maslov index and Theta series*, Progr. Math., vol. 6, Birkhäuser, Boston, 1980.
- [Litt] D. Littlewood, *The theory of group characters and matrix representations of groups*, Clarendon Press, Oxford, 1940.
- [Loeb] E. M. Loeb, ed., *Group theory and its applications*. I, II, III, Academic Press, 1975.
- [Looi] E. Looijenga, *Isolated singular points in complete intersections*, London Math. Soc. Lecture Notes 77, Cambridge Univ. Press, Cambridge, 1984.
- [Loom] L. Loomis, *An introduction to abstract harmonic analysis*, Van Nostrand, Princeton, NJ, 1953.
- [LoSt] L. Loomis and S. Sternberg, *Advanced calculus*, Addison-Wesley, Reading, MA, 1968.
- [Loos] O. Loos, *Symmetric spaces*. I, II, Benjamin, New York, 1969.
- [LuVo] G. Lusztig and D. Vogan, *Singularities of closures of  $K$ -orbits on flag manifolds*, Invent. Math. **71** (1983), 365–379.
- [Lus1] G. Lusztig, *Characters of reductive groups over a finite field*, Ann. of Math. Stud., no. 107, Princeton Univ. Press, Princeton, NJ, 1984.
- [Lus2] ———, *Irreducible representations of finite classical groups*, Invent. Math. **43** (1977), 125–175.
- [Lus3] ———, *Green polynomials and singularities of unipotent classes*, Adv. in Math. **42** (1981), 169–178.
- [Lus4] ———, *Canonical bases arising from quantized enveloping algebras*, Trans. Amer. Math. Soc. **3** (1990), 447–498.
- [Lus5] ———, *Some examples of square integrable representations of semisimple  $p$ -adic groups*, Trans. Amer. Math. Soc. **277** (1983), 623–653.
- [Macd1] I. G. Macdonald, *Symmetric functions and Hall polynomials*, Oxford Univ. Press, 1979.
- [Macd2] ———, *Affine root systems and Dedekind's  $\eta$ -function*, Invent. Math. **15** (1972), 91–143.
- [Macd3] ———, *Some conjectures for root systems*, SIAM J. Math. Anal. **13** (1982), 988–1007.
- [Macd4] ———, *A new class of symmetric functions*, Actes 20me Seminaire Loth., Pub. I.R.M.A., Strasbourg, 1988, pp. 131–171.
- [Mack1] G. Mackey, *Unitary group representations in physics, probability, and number theory*, Benjamin/Cummings, Reading, MA, 1978.

- [Mack2] ———, *Harmonic analysis as the exploitation of symmetry—a historical survey*, Bull. Amer. Math. Soc. (N.S.) **3** (1980), 543–698.
- [Mack3] ———, *Mathematical foundations of quantum mechanics*, Benjamin, New York, 1963.
- [Mack4] ———, *Unitary representations of group extensions. I*, Acta Math. **99** (1958), 265–311.
- [Malc] A. Malcev, *On a class of homogeneous spaces*, Izv. Akad. Nauk SSSR Ser. Mat. **13** (1949), 9–32; English transl. in Amer. Math. Soc. Transl. **39** (1951), 1–33.
- [Mant] L. Mantini, *An  $L^2$ -cohomology construction of negative spin mass zero equations for  $U(p, q)$* , J. Math. Anal. Appl. **136** (1988), 419–449.
- [Marg1] G. Margulis, *Non-uniform lattices in semisimple algebraic groups*, Lie groups and their representations (I. M. Gelfand, ed.), Wiley, New York, 1975.
- [Marg2] ———, *Arithmeticity of irreducible lattices in semisimple groups of rank greater than 1*, appendix to Russian translation of M. Raghunathan, *Discrete subgroups of Lie groups*, Mir, Moscow 1977.
- [Marg3] ———, *Discrete subgroups and ergodic theory*, Number Theory, Trace Formulas, and Discrete Groups (Aubert et al., eds.), Academic Press, Boston, 1989, pp. 377–398.
- [MWRSS] J. Marsden, A. Weinstein, T. Ratiu, R. Schmid, and R. Spencer, *Hamiltonian systems with symmetry, coadjoint orbits and plasma physics*, Atti Accad. Sci. Torino (Suppl.), vol. 117, Symposium Modern Developments in Analytical Mechanics, Academy of Sciences of Turin, 1983, pp. 289–340.
- [Masl] V. Maslov, *Théorie des perturbations et méthodes asymptotiques*, Dunod, Gauthiers-Villars, Paris, 1972.
- [Mass] W. Massey, *A basic course in algebraic topology*, Graduate Texts in Math., vol. 127, Springer-Verlag, New York, 1991.
- [Mats] Y. Matsushima, *A formula for the Betti numbers of compact, locally symmetric Riemannian manifolds*, J. Differential Geom. **1** (1967), 99–109.
- [Matz] B. Matzat, *Konstruktive Galoistheorie*, Lecture Notes in Math., vol. 1284, Springer-Verlag, Berlin, 1987.
- [Meht] M. Mehta, *Random matrices and the statistical theory of energy levels*, Academic Press, New York, 1967.
- [Meyel] Y. Meyer, *Ondelettes*, Hermann, Paris.
- [Mill1] W. Miller, Jr., *Lie theory and special functions*, Academic Press, New York, 1968.
- [Mill2] ———, *Symmetry groups and their applications*, Academic Press, New York, 1972.
- [Miln1] J. Milnor, *Eigenvalues of the Laplace operators on certain manifolds*, Proc. Nat. Acad. Sci. U.S.A. **51** (1964), 542.
- [Miln2] ———, *Morse theory*, Ann. of Math. Stud., no. 51, Princeton Univ. Press, Princeton, NJ, 1963.
- [Miln3] ———, *On the 3-dimensional Brieskorn manifolds  $m(p_q, r)$* , Knots, Groups and 3-Manifolds (L. P. Neuwirth, ed.), Ann. of Math. Stud., no. 84, Princeton Univ. Press, Princeton, NJ, 1974.
- [Mink] H. Minkowski, address to 80th Assembly of German National Scientists and Physicians, Cologne (21 Sept. 1908); reprinted in *The principle of relativity*, Dover, New York, 1923.
- [Moeg] C. Moeglin, *Correspondence de Howe pour les paires réductives duales: Quelques calculs dans le cas archimédien*, J. Funct. Anal. **85** (1989), 1–85.
- [MoVW] C. Moeglin, M.-F. Vigneras, and J.-L. Waldspurger, *Correspondences de Howe sur un corps  $p$ -adique*, Lecture Notes in Math., vol. 1291, Springer-Verlag, Berlin, 1987.
- [Moer] P. van Moerbeke, *The spectrum of Jacobi matrices*, Invent. Math. **371** (1976), 45–81.
- [Moli] T. Molien, *Über die Invarianten der linearen Substitutionsgruppe*, Sitzungsber. Königl. Preuss. Akad. Wiss., 1897, pp. 1152–1156.
- [MoZi1] D. Montgomery and L. Zippin, *Small subgroups of finite-dimensional groups*, Ann. of Math. (2) **56** (1952), 213–241.
- [MoZi2] ———, *Topological transformation groups*, Wiley-Interscience, New York, 1955.
- [Mood1] R. Moody, *Lie algebras associated with generalized Cartan matrices*, Bull. Amer. Math. Soc. **23** (1967), 217–221.
- [Mood2] ———, *A new class of Lie algebras*, J. Algebra **10** (1968), 211–230.
- [Mood3] ———, *Macdonald identities and Euclidean Lie algebras*, Proc. Amer. Math. Soc. **48** (1975), 43–52.

- [MoPa] R. Moody and V. Patera, *Fast recursion formula for weight multiplicities*, Bull. Amer. Math. Soc. (N.S.) **7** (1982), 237–242.
- [Moor] C. Moore, *Representations of solvable and nilpotent groups and harmonic analysis on nil and solvmanifolds*, Proc. Sympos. Pure Math., vol. 26, Amer. Math. Soc., Providence, RI, 1973, pp. 3–44.
- [Moor2] ———, *Ergodicity of flows on homogeneous spaces*, Amer. J. Math. **88** (1966), 154–178.
- [MoorR] R. Moore, *Measurable, continuous and smooth vectors for semigroups and group representations*, Mem. Amer. Math. Soc., vol. 78 Amer. Math. Soc., Providence, RI, 1968.
- [MoSe] G. Moore and N. Seiberg, *Classical and quantum conformal field theory*, Comm. Math. Phys. **123** (1989), 177–254.
- [Morr] W. Morris, II, *Constant term identities for finite and affine root systems: conjectures and theorems*, Thesis, University of Wisconsin, 1982.
- [MoVe] H. Moscovici and A. Verona, *Harmonically induced representations of nilpotent Lie groups*, Invent. Math. **48** (1978), 61–73.
- [Mose] J. Moser, *Three integrable Hamiltonian systems connected with isospectral deformations*, Adv. in Math. **16** (1975), 197–220.
- [Mosh1] M. Moshinsky, *Groups in physics: collective model of the nucleus; canonical transformations in quantum mechanics*, Presses Univ. Montreal, Montreal, 1979.
- [Mosh2] ———, *The harmonic oscillator in modern physics: from atoms to quarks*, Gordon and Breach, New York, 1969.
- [Most1] G. Mostow, *Strong rigidity of locally symmetric spaces*, Ann. of Math. Stud., no. 78, Princeton Univ. Press, Princeton, NJ, 1973.
- [Most2] ———, *Discrete subgroups of Lie groups*, Élie Cartan et les Mathématiques d'aujourd'hui, Astérisque, numéro hors série, Soc. Math. France, Paris, 1985, pp. 289–308.
- [Most3] ———, *Braids, hypergeometric functions, and lattices*, Bull. Amer. Math. Soc. (N.S.) **16** (1987), 225–246.
- [MoTa] G. Mostow and T. Tamagawa, *On the compactness of arithmetically defined homogeneous spaces*, Ann. of Math. (2) **76** (1962), 446–463.
- [Moy1] A. Moy, *Representations of  $U(2, 1)$  over a  $p$ -adic field*, J. Reine Angew. Math. **372** (1987), 178–208.
- [Moy2] ———, *Representations of  $GSp_4$  over a  $p$ -adic field*. I, II, Compositio Math. **66** (1988), 237–328.
- [Moy3] ———, *Local constants and the tame Langlands correspondence*, Amer. J. Math. **108** (1986), 863–930.
- [More] C. Moreno, *The strong multiplicity one theorem for  $GL_n$* , Bull. Amer. Math. Soc. (N.S.) **11** (1984), 180–182.
- [Mumf] D. Mumford, *Tata lectures on theta*. I, II, Progr. Math., vols. 28, 43, Birkhäuser, Boston, 1983, 1984.
- [Murp] G. Murphy, *A new construction of Young's seminormal representations of the symmetric group*, J. Algebra **69** (1981), 287–297.
- [Nach] L. Nachbin, *The Haar integral*, Van Nostrand, Princeton, NJ, 1965; reprinted Krieger, Huntington, NY, 1976.
- [NaSt] M. Naimark and A. Stern, *Theory of group representations*, Grundlehren Math. Wiss., vol. 246, Springer-Verlag, New York, 1982.
- [Neum] J. von Neumann, *Die eindeutigkeit der Schrödingerschen Operatoren*, Math. Ann. **104** (1931), 570–578.
- [Niwa] S. Niwa, *Modular forms of half-integral weight and the integral of certain theta functions*, Nagoya Math. J. **56** (1975), 147–161.
- [OkOz] ———, *On square integrable  $\bar{\partial}$ -cohomology spaces attached to Hermitian symmetric spaces*, Osaka J. Math. **4** (1967), 95–110.
- [OlOr] G. Olafsson and B. Orsted, *The holomorphic discrete series for affine symmetric spaces*. I, J. Funct. Anal. **81** (1988), 126–154.
- [OIPe] M. Olshanzky and A. Perelomov, *Explicit solution of the classical generalized Toda models*, Invent. Math. **54** (1979), 261–269.
- [Olve] P. Olver, *Applications of Lie groups to differential equations*, Graduate Texts in Math., vol. 107, Springer-Verlag, New York, 1986.
- [Olve2] ———, *Classical invariant theory and the equivalence problem for particle Lagrangians*, Bull. Amer. Math. Soc. (N.S.) **18** (1988), 21–26.

- [OnVi] A. Onischik and E. Vinberg, *Lie groups and algebraic groups*, Springer-Verlag, Berlin, 1990.
- [Opda1] E. Opdam, *Root systems and hypergeometric functions*. III, *Compositio Math.* **67** (1988), 21–99.
- [Opda2] ———, *Root systems and hypergeometric functions*. IV, *Compositio Math.* **67** (1988), 191–209.
- [Opda3] ———, *Some applications of hypergeometric shift operators*, *Invent. Math.* **981** (1989), 1–18.
- [Oppe] Oppenheim, *The minimum of indefinite quaternary quadratic forms*, *Ann. of Math.* (2) **32** (1931), 271–298.
- [Oshi1] T. Oshima, *Fourier analysis on semisimple symmetric spaces*, *Non commutative Harmonic Analysis and Lie Groups*, (J. Carmona and M. Vergne, eds.) *Lecture Notes in Math.*, vol. 880, Springer-Verlag, Berlin, 1981, pp. 357–369.
- [Oshi2] ———, *Eigenspaces of invariant differential operators in affine symmetric space*, *Invent. Math.* **57** (1980), 1–81.
- [Oshi3] ———, *Asymptotic behavior of spherical functions on semisimple symmetric spaces*, *Representations of Lie Groups*, Kyoto, Hiroshima, 1986; *Adv. Stud. Pure Math.*, vol. 14, North-Holland, Amsterdam, 1988, pp. 561–601.
- [OsMa] T. Oshima and T. Matsuki, *A description of discrete series for semisimple symmetric spaces*, *Group Representations and Systems Differential Equations*, *Adv. Stud. Pure Math.*, vol. 4, North-Holland, Amsterdam, 1984, pp. 331–390.
- [Ovsi] L. Ovsiannikov, *Group analysis of differential equations*, Academic Press, New York, 1982.
- [OvRo] S. Ovsienko and A. Roiter, *About the Schur problems for DGG*, *Matrix Problems*, 1977.
- [Pala] R. Palais, *A global formulation of the Lie theory of transformation groups*, *Mem. Amer. Math. Soc.*, vol. 22, Amer. Math. Soc., Providence, RI, 1957.
- [Part1] R. Parthasarathy, *Dirac operator and the discrete series*, *Ann. of Math.* (2) **93** (1971), 1–42.
- [Part2] ———, *A generalization of the Enright-Varadarajan modules*, *Compositio Math.* **36** (1978), 53–73.
- [PaRV] K. Parthasarathy, R. Ranga Rao, and V. Varadarajan, *Representations of complex semisimple Lie groups and Lie algebras*, *Ann. of Math.* (2) **85** (1967), 383–429.
- [PaRo] C. Patton and M. Rossi, *Unitary structures in cohomology*, *Trans. Amer. Math. Soc.* **290** (1985), 235–258.
- [Penn] R. Penney, *Lie cohomology of representations of nilpotent Lie groups and holomorphically induced representations*, *Trans. Amer. Math. Soc.* **261** (1980), 33–51.
- [Penr] R. Penrose, *Twistor theory, its aims and achievements*, *Quantum Gravity* (C. J. Isham, R. Penrose, and D. W. Sciama, eds.), Clarendon Press, Oxford, 1975, pp. 268–407.
- [PeWa] R. Penrose and R. Ward, *Twistors for flat and curved space-time*, *General Relativity and Gravitation* (A. Held, ed.), Plenum Press, New York and London, 1980.
- [Pere] A. Perelomov, *Integrable systems of classical mechanics and Lie algebras*. I, Birkhäuser-Verlag, Basel, 1990.
- [PeWe] F. Peter and H. Weyl, *Die Vollständigkeit der primitiven Darstellungen einer geschlossenen kontinuierlichen Gruppe*, *Math. Ann.* **97** (1927), 737–755.
- [Pete] K. Peterson, *Ergodic theory*, *Cambridge Surv. Adv. Math.*, vol. 2, Cambridge Univ. Press, Cambridge, 1983.
- [Piat1] I. Piatetski-Shapiro, *Multiplicity one theorems*, *Automorphic Forms, Representation Theory and L-Functions*, *Proc. Sympos. Pure Math.*, vol. 33, part 1, Amer. Math. Soc., Providence, RI, 1979, pp. 209–212.
- [Plat] V. Platonov, *The problem of strong approximation and the Kneser-Tits conjecture*, *Math. USSR Izv.* **3** (1969), 1139–1147; Addendum, *ibid.* **4** (1970), 784–786.
- [Pont] L. Pontrjagin, *Topological groups*, Princeton Univ. Press, Princeton, NJ, 1939.
- [Pras] G. Prasad, *Strong approximation for semi-simple groups over function fields*, *Ann. of Math.* **105** (1977), 553–572.
- [PrRa] G. Prasad and M. Raghunathan, *On the congruence subgroup problem: determination of the “metaplectic kernel,”* *Invent. Math.* **71** (1983), 21–42.

- [PrSe] A. Pressley and G. Segal, *Loop groups*, Oxford Math. Monographs, Clarendon Press, Oxford, 1986.
- [Proc1] R. Proctor, *Solution of two difficult combinatorial problems with linear algebra*, Amer. Math. Monthly **89** (1982), 721–734.
- [Prze] T. Przebinda, *The oscillator duality correspondence for the pair  $O(2, 2)$ ,  $Sp(2, \mathbb{R})$* , Mem. Amer. Math. Soc., vol. 403, Amer. Math. Soc., Providence, RI, 1989.
- [Puka1] L. Pukanszky, *Leçons sur les représentations des groupes*, Mon. Soc. Math. France, vol. 2, Dunod, Paris, 1967.
- [Puka2] ———, *On characters and the Plancherel formula of nilpotent groups*, J. Funct. Anal. **1** (1967), 255–280.
- [Puka3] ———, *Representations of solvable Lie groups*, Ann. Sci. École Norm. Sup. **4** (1971), 464–608.
- [Puka4] ———, *Characters of algebraic solvable groups*, J. Funct. Anal. **3** (1969), 435–494.
- [Puka5] ———, *On the theory of exponential groups*, Trans. Amer. Math. Soc. **126** (1967), 487–507.
- [Puka6] ———, *The primitive ideal space of solvable Lie groups*, Invent. Math. **22** (1973), 75–118.
- [Rade] H. Rademacher, *Topics in analytic number theory*, Grundlehren Math. Wiss., vol. 169, Springer-Verlag, Berlin, 1973.
- [Ragh] M. Raghunathan, *Discrete subgroups of Lie groups*, Ergeb. Math. Grenzgeb., no. 68, Springer-Verlag, New York, 1972.
- [Ragh2] M. Raghunathan, *On the congruence subgroup problem*, Inst. Hautes Études Sci. Publ. Math. **46** (1976), 107–161.
- [Rama] D. Ramakrishnan, *A regulator for curves via the Heisenberg group*, Bull. Amer. Math. Soc. (N.S.) **5** (1981), 191–195.
- [Ratn1] M. Ratner, *Strict measure rigidity for unipotent subgroups of solvable groups*, Invent. Math. **101** (1990), 449–482.
- [Ratn2] ———, *On measure rigidity of unipotent subgroups of semisimple groups*, Acta. Math. **165** (1990), 229–309.
- [Ratn3] ———, *Raghunathan's topological conjecture and distributions of unipotent flows*, preprint, 1990.
- [Ratn4] ———, *Distribution rigidity for unipotent actions on homogeneous spaces*, Bull. Amer. Math. Soc. (N.S.) **24** (1991), 321–325.
- [RaSW] J. Rawnsley, W. Schmid, and J. Wolf, *Singular unitary representations and indefinite harmonic theory*, J. Funct. Anal. **51** (1983), 1–114.
- [ReSe] A. Reiman and M. Semenov-Tjan-Shanskii, *Reduction of Hamiltonian systems, affine Lie algebras and Lax equations. II*, Invent. Math. **63** (1981), 423–432.
- [ReTu] N. Reshetikin and V. Turaev, *Invariants of 3-manifolds via link polynomials and quantum groups*, Invent. Math. **103** (1991), 547–597.
- [Rham] G. deRham, *Sur l'analyse situs des variétés à  $n$  dimensions*, J. Math. Pures Appl. **10** (1931), 115–200.
- [Rief] M. Rieffel, *Unitary representations of group extensions: an algebraic approach to the theory of Mackey and Blattner*, Adv. Math. Suppl. Stud., Academic Press, Orlando, FL, 1979, pp. 43–82.
- [RiNa] F. Riesz and B. Sz.-Nagy, *Leçons d'analyse fonctionnelle*, 4th ed., Acad. Science de Hongrie, Budapest; Gauthier-Villars, Paris, 1965.
- [Ring] C. Ringel, *The rational invariants of tame quivers*, Invent. Math. **58** (1980), 217–239.
- [Rock] C. Rockland, *Hypoellipticity on the Heisenberg group: representation-theoretic criteria*, Trans. Amer. Math. Soc. **240** (1978), 1–52.
- [Roga1] J. Rogawski, *Automorphic representations of unitary groups in three variables*, Ann. of Math. Stud., no. 123, Princeton Univ. Press, Princeton, NJ, 1990.
- [Roga2] J. Rogawski, *On modules over the Hecke algebra of a  $p$ -adic group*, Invent. Math. **79** (1985), 443–465.
- [Rose] J. Rosenberg, *Realization of square-integrable representations of unimodular Lie groups on  $L^2$ -cohomology spaces*, Trans. Amer. Math. Soc. **126** (1980), 1–32.
- [Ross] W. Rossman, *Kirillov's character formula for reductive Lie groups*, Invent. Math. **48** (1978), 207–220.

- [Ross2] ———, *Limit characters of reductive Lie groups*, Invent. Math. **61** (1980), 53–68.
- [RoSt] L. Rothschild and E. Stein, *Hypoelliptic differential operators and nilpotent groups*, Acta. Math. **137** (1976), 247–320.
- [SaWu] R. Sachs and H. Wu, *General relativity for mathematicians*, Graduate Texts in Math., vol. 48, Springer-Verlag, New York, 1977.
- [Sait] H. Saito, *Automorphic forms and algebraic extensions of number fields*, Lectures in Math., vol. 8, Kinokuniya, Tokyo, 1975.
- [Sall] P. Sally, *Unitary and uniformly bounded representations of the two-by-two unimodular groups over local fields*, Amer. J. Math. **90** (1968), 406–443.
- [Sata] I. Satake, *Spherical functions and Ramanujan conjecture*, Algebraic Groups and Discontinuous Subgroups, Proc. Sympos. Pure Math., vol. 9, Amer. Math. Soc., Providence, RI, 1965.
- [Satt] D. Sattinger, *Group-theoretic methods in bifurcation theory*, Lecture Notes in Math., vol. 762, Springer-Verlag, New York, 1972.
- [SaWe] D. Sattinger and O. Weaver, *Lie groups and algebras, with applications to physics, geometry, and mechanics*, Appl. Math. Sci., vol. 61, Springer-Verlag, New York, 1985.
- [Schl] H. Schlichtkrull, *Hyperfunctions and harmonic analysis on symmetric spaces*, Progr. Math., vol. 49, Birkhäuser, Boston, 1984.
- [Schm1] W. Schmid, *On a conjecture of Langlands*, Ann. of Math. (2) **93** (1971), 1–42.
- [Schm2] ———, *Some properties of square integrable representations of semisimple Lie groups*, Ann. of Math. (2) **102** (1975), 535–564.
- [Schm3] ———,  *$L_2$  cohomology and the discrete series*, Ann. of Math. (2) **103** (1976), 294–375.
- [Scho] B. Schoeneberg, *Elliptic modular functions*, Grundlehren Math. Wiss., vol. 203, Springer-Verlag, Berlin, 1974.
- [Schu] I. Schur, *Über eine Klasse von Matrizen, die sich einer gegebenen Matrix zuordnen lassen*, Dissertation, Gesammelte Abhandlungen, Bd. 1, Springer-Verlag, Berlin, 1973.
- [Segl] G. Segal, *The definitions of conformal field theory*, Links between Geometry and Mathematical Physics, MPI 87-58 preprint, pp. 13–17.
- [Sega1] I. Segal, *Foundations of the theory of dynamical systems of infinitely many degrees of freedom*, I, Mat.-Fys. Medd. Dansk. Vid. Selsk. (12) **31** (1959), 39pp.
- [Sega2] ———, *The complex wave representation of the free boson field*, Topics in Functional Analysis, Adv. Math. Suppl. Stud., vol. 3, Academic Press, New York, 1978.
- [Sega3] ———, *Mathematical problems of relativistic physics*, Lecture in Appl. Math., vol. 2, Amer. Math. Soc., Providence, RI, 1963.
- [Sega4] ———, *Mathematical cosmology and extragalactic astronomy*, Pure. Appl. Math., vol. 68, Academic Press, New York, 1976.
- [Selb1] A. Selberg, *On discontinuous groups in higher dimensional symmetric spaces*, Internat. Colloq. on Function Theory, Tata Institute, Bombay, 1960.
- [Selb2] A. Selberg, *Harmonic analysis and discontinuous groups in weakly symmetric spaces with applications to Dirichlet series*, J. Indian Math. Soc. **20** (1956), 47–87.
- [Seli] G. Seligman, *Modular Lie algebras*, Ergeb. Math. Grenzgeb., no. 40, Springer-Verlag, Berlin, 1967.
- [Serr1] J.-P. Serre, *Algebres de Lie semi-simples complexes*, Benjamin, New York, 1966.
- [Serr2] ———, *Lie algebras and Lie groups*, Benjamin, New York, 1965.
- [Serr3] ———, *Représentations linéaires et espaces homogènes kähleriens des groupes de Lie compacts*, Sem. Bourbaki, no. 100, Paris, 1954.
- [Sesh] C. Seshadri, *Geometry of  $G/P$ , I. I*, in C. P. Ramanujan: A Tribute, Tata Inst. Fund. Res. Stud. in Math., Springer-Verlag, Berlin, 1978, pp. 207–239.
- [Shal] D. Shale, *Linear symmetries of free boson fields*, Trans. Amer. Math. Soc. **103** (1962), 149–167.
- [Shal] J. Shalika, *Representations of the two-by-two unimodular group over local fields*, Thesis, Johns Hopkins University, 1966.
- [Shan] R. Shankar, *Principles of quantum mechanics*, Plenum Press, New York and London, 1980.
- [Shel] D. Shelstad,  *$L$ -indistinguishability for real groups*, Math. Ann. **259** (1982), 385–430.



- [ShTo] G. Shephard and J. Todd, *Finite unitary reflection groups*, *Canad. J. Math.* **6** (1954), 274–304.
- [Shi] J.-Y. Shi, *The Kazhdan-Lusztig cells in certain affine Weyl groups*, *Lecture Notes in Math.*, vol. 1179, Springer-Verlag, Berlin, 1986.
- [Shmz1] H. Shimizu, *On discontinuous groups operating on the product of the upper half planes*, *Ann. of Math.* (2) **77** (1963).
- [Shmz2] ———, *On zeta functions of quaternion algebras*, *Ann. of Math.* (2) **81** (1965).
- [Shim1] G. Shimura, *Introduction to the arithmetic theory of automorphic forms*, Princeton Univ. Press, Princeton, NJ, 1971.
- [Shim2] ———, *On modular forms of half-integral weight*, *Ann. of Math.* (2) **97** (1973), 440–481.
- [Shin1] T. Shintani, *On certain square integrable irreducible unitary representations of some  $p$ -adic linear groups*, *J. Math. Soc. Japan* **20** (1968), 522–565.
- [Shin2] ———, *On construction of holomorphic cusp forms of half-integral weight*, *Nagoya Math. J.* **58** (1975), 83–126.
- [Silb1] A. Silberger,  *$PGL_2$  over the  $p$ -adics: its representations, spherical functions and Fourier analysis*, *Lecture Notes in Math.*, vol. 166, Springer-Verlag, Berlin, 1970.
- [Silb2] A. Silberger, *Introduction to harmonic analysis on reductive  $P$ -adic groups*, *Math. Notes* **23**, Princeton Univ. Press, Princeton, NJ, 1979.
- [Sitt] W. de Sitter, *On Einstein's theory of gravitation and its astronomical consequences*, *Monthly Notices Roy. Astronom. Soc.* **76** (1916), 699; **78** (1917), 3.
- [Slod] P. Slodowy, *Simple singularities and simple algebraic groups*, *Lecture Notes in Math.*, vol. 815, Springer-Verlag, Berlin, 1980.
- [Snia] J. Sniatycki, *Geometric quantization and quantum mechanics*, Springer-Verlag, New York, 1980.
- [Sour] J.-M. Souriau, *Structure des systemes dynamiques*, Dunod, Paris, 1970.
- [SpVo] B. Speh and C. Vogan, *Reducibility of generalized principal series representations*, *Acta Math.* **145** (1980), 227–299.
- [Spr1] T. Springer, *Reductive groups*, *Automorphic Forms, Representations, and  $L$ -Functions*, *Proc. Sympos. Pure Math.*, vol. 33, part 1, Amer. Math. Soc., Providence, RI, 1979, pp. 3–28.
- [Spr2] ———, *Linear algebraic groups*, *Progr. Math.*, vol. 9, Birkhäuser, Boston, 1981.
- [Stan] R. Stanley, *Invariants of finite groups and their applications to combinatorics*, *Bull. Amer. Math. Soc. (N.S.)* **1** (1979), 475–511.
- [Stan] N. Stanton, *The heat equation in several complex variables*, *Bull. Amer. Math. Soc. (N.S.)* **11** (1984), 65–84.
- [Stee] N. Steenrod, *The topology of fibre bundles*, Princeton Univ. Press, Princeton, NJ, 1951.
- [Ste] E. Stein, *Analysis in matrix space and some new representations of  $SL(n, \mathbb{R})$* , *Ann. of Math.* (2) **86** (1967), 461–490.
- [Stnb1] R. Steinberg, *Generators for simple groups*, *Canad. J. Math.* **14** (1962), 277–283.
- [Stnb2] ———, *Générateurs, relations et revêtements des groupes algébriques*, *Colloq. sur la Théorie des Groupes Algébriques*, Bruxelles, 1962, pp. 113–127.
- [Stem] J. Stembridge, *A short proof of Macdonald's conjecture for the root systems of type A*, *Proc. Amer. Math. Soc.* **102** (1988), 777–786.
- [Ster] S. Sternberg, *Lectures on differential geometry*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
- [Sthr] D. Sternheimer, *Phase-space representations, Applications of group theory in physics and mathematical physics*, *Lectures in App. Math.*, vol. 21, Amer. Math. Soc., Providence, RI, 1985.
- [StWi] H. Strade and R. Wilson, *Classification of simple Lie algebras over algebraically closed fields of finite characteristic*, *Bull. Amer. Math. Soc. (N.S.)* **24** (1991), 357–362.
- [Stra] G. Strang, *Linear algebra and its applications*, 2nd ed., Academic Press, New York, 1980.
- [Sugi] M. Sugiura, *Unitary representations and harmonic analysis, an introduction*, Kodansha, Tokyo and Wiley, New York, 1975.
- [Suna] T. Sunada, *Gelfand's problem on unitary representations associated with discrete subgroups of  $PSL_2(\mathbb{R})$* , *Bull. Amer. Math. Soc. (N.S.)* **12** (1985), 237–238.
- [Suss] H. Sussman, *Lie brackets and local controllability; a sufficient condition for scalar input systems*, *SIAM J. Control Optim.* **21** (1983), 686–713.
- [Swee] M. Sweedler, *Hopf algebras*, Benjamin, New York, 1969.

- [Sylv1] J. Sylvester, *Tables of the generating functions and ground forms of the binary quantics of the first ten orders*, Collec. Math. Papers III, Chelsea, 1973, pp. 283–311.
- [Sylv2] J. Sylvester, *Tables of the generating functions and ground forms of the binary duodecemic, with some remarks, and tables of the irreducible syzygies of certain quantics*, Collec. Math. Papers III, Chelsea, 1973, pp. 489–508.
- [Syme1] W. Symes, *Systems of Toda type, inverse spectral problems and representation theory*, Invent. Math. **59** (1980), 13–51.
- [Syme2] ———, *The QR algorithm and scattering for the non-periodic Toda lattice*, Phys. D **4** (1982), 13–51.
- [Tadi] M. Tadic, *Classification of unitary representations of irreducible representations of general linear group (non-archimedean case)*, Ann. Sci. École Norm. Sup. **19** (1986), 335–382.
- [Tate] J. Tate, *Number theoretic background*, Automorphic Forms, Representation Theory, and L-Functions, Proc. Sympos. Pure Math., vol. 33, part 2, Amer. Math. Soc., Providence, RI, 1979, pp. 3–26.
- [Taub] C. Taubes, *Self-dual Yang-Mills connections on non-self-dual 4-manifolds*, J. Differential Geom. **17** (1982), 139–170.
- [Tera] I. Terada, *A Robinson-Schensted-type correspondence for a dual pair on spinors*, preprint, 1990.
- [Tits] J. Tits, *Liesche Gruppen und Algebren*, Springer-Verlag, Berlin, 1983.
- [Thim] A. Thimm, *Integrable geodesic flows on homogeneous spaces*, Ergodic Theory Dynamical Systems **1** (1981), 495–517.
- [Thir] W. Thirring, *Principles of quantum electrodynamics*, Pure Appl. Phys., vol. 3, Academic Press, New York, 1958.
- [Thms] E. Thomas, *Steenrod squares and H-spaces*, II, Ann. of Math. (2) **81** (1965), 473–495.
- [Thom] J. Thompson, *Some finite groups which appear as  $\text{Gal}(L/K)$  where  $K \subset \mathbb{Q}(\mu_n)$* , J. Algebra **89** (1984), 437–449.
- [Thom] T. Thompson, *From error-correcting codes through sphere packings to simple groups*, Carus Math. Monographs, no. 21, Math. Assoc. Amer., 1983.
- [TiBo] L. Timothy and B. Bona, *State space analysis: an introduction*, McGraw-Hill, San Francisco, 1968.
- [Toda1] M. Toda, *Wave propagation in anharmonic lattices*, J. Phys. Soc. Jap. **23** (1967), 501–506.
- [Toda2] ———, *Studies of a non-linear lattice*, Phys. Rep. **8** (1975), 1–125.
- [ToWa1] Y. Tong and S. Wang, *Harmonic forms dual to geodesic cycles in quotients of  $\text{SU}(p, q)$* , Math. Ann. **258** (1982), 298–318.
- [ToWa2] ———, *Period integrals in non-compact quotients of  $\text{SU}(p, 1)$* , Duke Math. J. **52** (1985), 649–688.
- [Tora] P. Torasso, *Quantification géométrique et représentations de  $\widetilde{\text{SL}}_3(\mathbb{R})$* , C. R. Acad. Sci. Paris Sér. A **291** (1980), A185–A188.
- [TrPo] E. Trubowitz and J. Poschel, *Inverse spectral theory*, Pure. Appl. Math., vol. 130, Academic Press, Boston, 1987.
- [TuVi] V. Turaev and O. Viro, *State sum invariants of 3-manifolds and quantum 6j-symbols*, preprint, 1990.
- [Tunn1] J. Tunnel, *On the local Langlands conjectures for  $\text{GL}(2)$* , Invent. Math. **46** (1978), 179–200.
- [Tunn2] ———, *Artin's conjecture for representations of octahedral type*, Bull. Amer. Math. Soc. (N.S.) **5** (1981), 173–175.
- [Unte] A. Unterberger, *Oscillateur harmonique et opérateurs pseudo-différentiels*, Ann. Inst. Fourier (Grenoble) **29** (1979), 201–221.
- [Vara] V. Varadarajan, *Lie groups, Lie algebras, and their representations*, Graduate Texts in Math., vol. 102, Springer-Verlag, New York, 1984.
- [Vara] V. Varadarajan, *Harmonic analysis on real reductive groups*, Lecture Notes in Math., vol. 576, Springer-Verlag Berlin, 1977.
- [Verg1] M. Vergne, *Étude de certaines représentations induites d'un groupe de Lie résoluble exponentiel*, Ann. Sci. École Norm. Sup. **3** (1970), 353–384.
- [Verg2] ———, *On Rossmann's character formula for discrete series*, Invent. Math. **54** (1979), 11–14.

- [Vign] M.-F. Vigneras, *Variétés riemanniennes isospectrales et non-isométriques*, Ann. of Math. (2) **112** (1980), 21–33.
- [Vile] N. Vilenkin, *Special functions and the theory of group representations*, Transl. Math. Monographs, no. 22, Amer. Math. Soc., Providence, RI, 1968.
- [Voga1] D. Vogan, *Unitary representations of reductive Lie groups*, Ann. of Math. Stud., no. 118, Princeton Univ. Press, Princeton, NJ, 1987.
- [Voga2] ———, *Representations of real reductive Lie groups*, Progr. Math., vol. 15, Birkhäuser, Boston, 1981.
- [Voga3] ———, *Unitarizability of certain series of representations*, Ann. of Math. (2) **120** (1984), 141–187.
- [Voga4] ———, *Algebraic structure of the representations of semisimple Lie groups. I*, Ann. of Math. (2) **109** (1979), 1–60.
- [Voga5] ———, *The unitary dual of  $GL(n)$  over an archimedean field*, Invent. Math. **83** (1986), 449–505.
- [Voga6] ———, *Noncommutative algebras and unitary representations*, The Mathematical Heritage of Hermann Weyl, Proc. Sympos. Pure Math., vol. 48, Amer. Math. Soc., Providence, RI, 1988, pp. 35–60.
- [Voga7] ———, *Irreducible characters of semisimple Lie groups. III, Proof of the Kazhdan-Lusztig conjectures in the integral case*, Invent. Math. **71** (1983), 381–417.
- [Voga8] ———, *Irreducible characters of semisimple Lie groups. IV, Character multiplicity duality*, Duke Math. J. **49** (1982), 943–1073.
- [VoZu] D. Vogan and G. Zuckerman, *Unitary representations with continuous cohomology*, Compositio Math. **53** (1984), 51–90.
- [Waer] B. van der Waerden, *Group theory and quantum mechanics*, Grundlehren Math. Wiss., vol. 214, Springer-Verlag, Berlin, 1974.
- [Wald1] J.-L. Waldspurger, *Correspondence de Shimura*, J. Math. Pures Appl. **59** (1980), 1–132.
- [Wald2] ———, *Sur les coefficients de Fourier des formes modulaires de poids demi-entier*, J. Math. Pures Appl. **60** (1981), 375–384.
- [Wall1] N. Wallach, *Symplectic geometry and Fourier analysis*, Math.-Sci. Press, Brookline, MA, 1977.
- [Wall2] ———, *Real reductive groups. I*, Pure Appl. Math., vol. 132, Academic Press, San Diego, CA, 1988.
- [Wall3] ———, *Harmonic analysis on homogeneous spaces*, Marcel Dekker, New York, 1973.
- [Wall4] ———, *On the Enright Varadarajan modules: a construction of the discrete series*, Ann. Sci. École Norm. Sup. **9** (1976), 81–102.
- [Wall5] ———, *On the unitarizability of derived functor modules*, Invent. Math. **78** (1984), 131–141.
- [Wall6] ———, *Compact homogeneous Riemannian manifolds with strictly positive curvature*, Ann. of Math. (2) **96** (1972), 277–295.
- [Wang] S. Wang, *Correspondence of modular forms to cycles associated to  $O(p, q)$* , J. Differential Geom. **22** (1985), 151–222.
- [Warn] G. Warner, *Harmonic analysis on semi-simple Lie groups. I, II*, Grundlehren Math. Wiss., vols. 188, 189, Springer-Verlag, New York, 1972.
- [Weil1] A. Weil, *Sur certains groupes d'opérateurs unitaires*, Acta Math. **111** (1964), 143–211.
- [Weil2] ———, *Basic number theory*, Grundlehren Math. Wiss., vol. 144, Springer-Verlag, Berlin, 1973.
- [Weil3] ———, *Variétés Kähleriennes*, Hermann, Paris, 1958.
- [Weil4] ———, *L'intégration dans les groupes topologiques et ses applications*, Actualités Sci. Indust., no. 551, Hermann, Paris, 1938.
- [Wein] A. Weinstein, *Symplectic geometry*, Bull. Amer. Math. Soc. (N.S.) **5** (1981), 1–13.
- [Weis] B. Weisfeiler, *Abstract homomorphisms of big subgroups of algebraic groups*, Topics in the Theory of Algebraic Groups, Notre Dame Math. Lectures, vol. 10, Univ. of Notre Dame Press, South Bend, IN, 1982.
- [Well] R. Wells, *Complex manifolds and mathematical physics*, Bull. Amer. Math. Soc. (N.S.) **1** (1979), 296–336.

- [Weyl1] H. Weyl, *Theorie der Darstellungen kontinuierlicher halbeinfachen Gruppen durch lineare Transformationen*. I, II, III und Nachtrag: Math. Z. **23** (1925), 271–309; Math. Z. **24** (1926), 328–376; Math. Z. **24** (1926), 377–395.
- [Weyl2] ———, *The classical groups*, Princeton Univ. Press, Princeton, NJ, 1946.
- [Weyl3] ———, *The theory of groups and quantum mechanics*, Dover, New York, 1931.
- [Wigr] E. Wigner, *On unitary representations of the inhomogeneous Lorentz group*, Ann. of Math. (2) **40** (1939), 149–204.
- [Witt] E. Witten, *Non-abelian bosonization in two dimensions*, Comm. Math. Phys. **92** (1984), 455–472.
- [Wolb] A. Wolbart, *Symmetry and quantum systems*, Van Nostrand, Reinhold, New York, 1977.
- [Yama1] H. Yamabe, *On the conjecture of Iwasawa and Gleason*, Ann. of Math. (2) **58** (1953), 48–54.
- [Yama2] ———, *A generalization of a theorem of Gleason*, Ann. of Math. (2) **58** (1953), 351–365.
- [Yau] S.-T. Yau, *Mathematical aspects of string theory*, World Scientific, Singapore, 1987.
- [ZaDe] L. Zadeh and C. Desoer, *Linear system theory*, McGraw-Hill, New York, 1963.
- [Zeil1] D. Zeilberger, *A proof of the  $G_2$  case of Macdonald's root system—Dyson conjecture*, SIAM J. Math. Anal. **18** (1989), 880–883.
- [Zeil2] ———, *A unified approach to Macdonald's root system conjecture*, SIAM J. Math. Anal. **19** (1988), 987–1013.
- [Zeil3] ———, *Kathy O'Hara's constructive proof of the unimodality of the Gaussian polynomials*, Amer. Math. Monthly **96** (1989), 590–601.
- [Zhel1] D. Zhelobenko, *Compact Lie groups and their representations*, Transl. Math. Mono., vol. 40, Amer. Math. Soc., Providence, RI, 1973.
- [Zhel2] ———, *The analysis of irreducibility in a class of elementary representations of a complex semisimple Lie group*, Math. USSR-Izv. **2** (1968), 105–128.
- [Zill] D. Zill, *Differential equations with boundary-value problems*, Prindle, Weber and Schmidt, Boston, 1986.
- [Zimm1] R. Zimmer, *Ergodic theory and semisimple groups*, Monographs Math., vol. 81, Birkhäuser, Boston, 1984.
- [Zimm2] ———, *Ergodic theory, group representations, and rigidity*, Bull. Amer. Math. Soc. (N.S.) **6** (1982), 383–416.

DEPARTMENT OF MATHEMATICS, YALE UNIVERSITY, NEW HAVEN, CONNECTICUT 06520

## From Quantum Theory to Knot Theory and Back: a von Neumann Algebra Excursion

V.F.R. JONES

In the usual formulation of quantum mechanics, the pure states of a physical system are given by the one-dimensional subspaces of a complex Hilbert space  $\mathcal{H}$ , with inner product  $\langle \cdot, \cdot \rangle$ . (If  $\mathcal{H}$  is realized as a concrete  $L^2$  space of functions, a function of unit norm in the one-dimensional subspace is called the wave function.) The observables of the system are given by selfadjoint (possibly unbounded) operators on  $\mathcal{H}$  and the expected value of measuring an observable  $A$  for a system in a state  $\xi$  is  $\langle A\xi, \xi \rangle$ . Two observables  $A$  and  $B$  can be simultaneously measured with arbitrary precision only if they commute, i.e.,  $AB = BA$ . The importance of this relation justifies the following notation.

**DEFINITION.** If  $X$  is a set of operators on  $\mathcal{H}$ , the commutant  $X'$  is the set of all *bounded* operators on  $\mathcal{H}$  which commute with any element of  $X$ .

The time evolution of the system is governed by a privileged observable  $H$  and is given by the one-parameter unitary group  $\exp(itH)$ . Any symmetry group of the system will be implemented by a projective unitary representation of the group on  $\mathcal{H}$ .

The above formalism was well understood and clearly enunciated by von Neumann in [vN1]. The whole picture can be almost deduced on more or less philosophical grounds from logical relationships between the results of experiments. See, for instance, [Pi]. One thing that is quite tricky from this “logical” point of view but which is an essential part of the picture is the description of a composite system made up of two or more subsystems. The formalism is that the Hilbert space describing the composite system is the tensor product of the Hilbert spaces describing the components. This is quite natural from the point of view of wave functions.

I have already said enough to justify, from a physical point of view, the study of von Neumann algebras, for which I now give three equivalent

---

1980 *Mathematics Subject Classification* (1985 Revision). Primary 46L10, 81P05, 81R10, 57M25, 16S99.

definitions, all on a given Hilbert space  $\mathcal{H}$ .

DEFINITION 1. A von Neumann algebra is the commutant of some self-adjoint set of operators on  $\mathcal{H}$ .

DEFINITION 2. A von Neumann algebra  $M$  is a selfadjoint algebra of operators on  $\mathcal{H}$  equal to the commutant of its commutant:  $M = M''$ .

DEFINITION 3. A von Neumann algebra  $M$  is a selfadjoint algebra of bounded operators on  $\mathcal{H}$ , containing the identity and such that if a net  $A_n$ , in  $M$ , converges to  $A$  in the sense that  $\langle A_n \xi, \eta \rangle \rightarrow \langle A \xi, \eta \rangle$  for all  $\xi$  and  $\eta$ , then  $A \in M$  (the net is said to converge weakly).

Definitions 1 and 2 are superficially equivalent while the equivalence between 2 and 3 is the famous "density theorem" of von Neumann [vN2].

We can immediately see four (related) uses for von Neumann algebras.

- (1) In the mathematical formulation of quantum mechanics (e.g., sub-systems).
- (2) To study unbounded operators by bounded ones using the commutant.
- (3) To study unitary group representations.
- (4) As a generalization of finite dimensional semisimple algebras where the selfadjointness of the algebra should guarantee some kind of semisimplicity.

All the above were stated motivations in the 1936 paper of Murray and von Neumann [MvN], where von Neumann algebras were first introduced under the name "rings of operators."

Abelian von Neumann algebras are easy to understand. Take a  $\sigma$ -finite measure space  $(X, \mu)$  and form the Hilbert space  $L^2(X, \mu)$ . The  $L^\infty$  functions act on  $L^2$  by multiplication and form a maximal abelian  $*$ -algebra of bounded operators. Thus  $L^\infty(X, \mu)$  is an abelian von Neumann algebra. The general abelian von Neumann algebra can be obtained from this one simply by changing the "multiplicity," for instance by letting it act diagonally on  $L^2(X, \mu) \oplus L^2(X, \mu) \oplus \cdots$ , and then reducing to some invariant subspace. In any case, as an abstract  $*$ -algebra any abelian von Neumann algebra is isomorphic to  $L^\infty(X, \mu)$ .

Now if we take an arbitrary von Neumann algebra  $M$ , its center  $Z(M)$  is abelian. Along with the decomposition of  $Z(M)$  as  $L^\infty(X, \mu)$  goes a decomposition of all of  $M$  as a "direct integral"  $M = \int_X^\oplus M(\eta) d\mu(\eta)$  into a measurable family  $M(\eta)$  of von Neumann algebras with trivial center. An arbitrary element of  $M$  becomes a measurable function  $\eta \mapsto a(\eta) \in M(\eta)$ , modulo sets of measure zero. As one can imagine, the details of the direct integral decomposition are somewhat painful (see [vN3]) but in principle it reduces the study of arbitrary von Neumann algebras to that of ones with trivial center. Note also how we are quickly moving away from the notion of a concrete von Neumann algebra to an abstract notion.

So what does a von Neumann algebra  $M$  with trivial center look like? In finite dimensions the structure theory of semisimple algebras says that such

an algebra is  $*$ -isomorphic to a full matrix algebra  $M = M_n(\mathbb{C})$ . Moreover, as a concrete algebra on a Hilbert space it corresponds to a tensor product factorization  $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$  with  $M = \{x \otimes 1 \mid x \text{ on } \mathcal{H}_1\}$  (and  $M' = \{1 \otimes y \mid y \text{ on } \mathcal{H}_2\}$ ). The same factorization result holds in infinite dimensions if  $M$  is abstractly isomorphic to the algebra  $\mathcal{B}(\mathcal{H})$  of all bounded operators on some Hilbert space. Thus Murray and von Neumann called a von Neumann algebra with trivial center a *factor*, and quickly restricted their attention to these.

In the language of quantum mechanics, a factor corresponds to a physical system for which there is no observable which can be simultaneously measured, to arbitrary precision, with all other observables. In group representation language, a representation whose commutant is a factor is called isotypical, at least in finite dimensions, because Schur's Lemma guarantees that only one type of representation occurs in a decomposition into irreducibles.

Central to the analysis of von Neumann algebras is the understanding of *projections* onto closed subspaces which can be characterized abstractly as operators  $p$  with  $p^2 = p = p^*$ . It is in terms of projections that Murray and von Neumann gave their first classification of factors into types I, II, and III. I shall give an alternative version emphasizing *traces*, though even so the role of projections will be important.

A factor  $M$  is said to be of *type I* if it is abstractly isomorphic to some  $\mathcal{B}(\mathcal{H})$ . It then admits a trace defined on a certain subalgebra of operators, such that the trace of a rank one projection is 1. With an obvious convention concerning  $\infty$ , type I factors (on a separable Hilbert space) are classified up to abstract isomorphism by the single number  $n = \dim \mathcal{H}$ , and up to concrete isomorphism by the pair  $(\dim \mathcal{H}, \dim \mathcal{H}')$ , where  $\mathcal{H} = \mathcal{H} \otimes \mathcal{H}'$  as before. The traces of projections form the set  $I_n: \{0, 1, 2, \dots, n\}$ .

A factor  $M$  is said to be of *type II* if it is not of type I and admits a trace  $\text{tr}$  which is a linear functional  $\text{tr}: M \rightarrow \mathbb{C}$  satisfying

- (1)  $\text{tr}(ab) = \text{tr}(ba)$ ,
- (2)  $\text{tr}(1) = 1$ ,
- (3)  $\text{tr}(aa^*) > 0$  if  $a \neq 0$ ,
- (4)  $\text{tr}$  is weakly continuous.

(As it happens, only condition (1) is important, (2) is a normalization, and (3) and (4) are automatic.)

The traces of projections in a  $\text{II}_1$  factor form the set  $[0, 1]$  and one talks of continuous dimensionality.

It is not true that  $\text{II}_1$  factors give tensor product factorizations, nor is it known how to list all  $\text{II}_1$  factors up to abstract isomorphism. However Murray and von Neumann did show how to associate a number  $\dim_M(\mathcal{H})$  (which they called the coupling constant) to a concrete  $\text{II}_1$  factor on  $\mathcal{H}$ , which completely characterizes  $M$ , once it is known as an abstract  $\text{II}_1$

factor. It corresponds to the integer  $\dim \mathcal{H}'$  in the type I case but because of continuous dimensionality it may be any nonnegative real or  $\infty$ .

A factor  $M$  is said to be of *type*  $\text{II}_\infty$  if it contains a projection  $p$  with  $pMp$  a type  $\text{II}_1$  factor. It is then of the form  $N \otimes \mathcal{B}(\mathcal{H})$  with  $N \cong pMp$ , and admits a trace in a similar sense to the type I case. The traces of projections form the set  $[0, \infty]$ .

A factor  $M$  is said to be of *type* III if it is not of type I or II.

It is easy to construct  $\text{II}_1$  factors (and hence  $\text{II}_\infty$ ) using discrete groups. If  $\Gamma$  is a group all of whose conjugacy classes (except that of the identity) are infinite, then the von Neumann algebra generated by the left regular representation is a type  $\text{II}_1$  factor. Elements are all of the form  $\sum_{\gamma \in \Gamma} c_\gamma U_\gamma$  (where  $(u_\gamma f)(\gamma^1) = f(\gamma^{-1}\gamma^1)$ ,  $f \in l^2(\Gamma, \mathbb{C})$ ) and  $\text{tr}(\sum c_\gamma u_\gamma) = C_1$ . Type III factors can be constructed by replacing the scalars in the above construction by an abelian von Neumann algebra  $A$  carrying an ergodic action of  $\Gamma$  with no invariant measure (on the space  $(X, \mu)$  for which  $A = L^\infty(X, \mu)$ ). In quantum field theory, the von Neumann algebra corresponding to all fields restricted to a bounded region of space time is supposed to be a type III factor and this can be shown in some cases.

To begin to answer more subtle questions, such as "how does the  $\text{II}_1$  factor depend on the group  $\Gamma$  as above," the notion of *hyperfiniteness* is crucial. A von Neumann algebra  $M$  is said to be hyperfinite if there is an increasing net of finite-dimensional  $*$ -subalgebras whose union is weakly dense in  $M$ . Murray and von Neumann showed in [MvN] that there is only one hyperfinite  $\text{II}_1$  factor. A celebrated result of Connes in [Co1] asserts that the factors constructed above using groups are hyperfinite iff  $\Gamma$  is amenable (e.g.,  $\Gamma = \mathbb{Z}$ ). The local algebras of quantum field theory are hyperfinite.

The classification of hyperfinite factors is complete. To explain it would require some Tomita-Takesaki theory so we simply say that Connes gave a finer classification of type III factors in terms of a continuous parameter  $\lambda \in [0, 1]$ . Connes showed that there is exactly one hyperfinite type  $\text{II}_\infty$  factor and one in each class  $\text{III}_\lambda$ ,  $0 < \lambda < 1$ . By results of Krieger, hyperfinite type  $\text{III}_0$  factors are classified by ergodic transformations, and Haagerup proved the uniqueness of the hyperfinite type  $\text{III}_1$  factor in [Ha].

**Subfactors.** With hyperfinite factors classified it is natural to turn to subfactors. There were four good reasons for this in 1976 when I began to look at subfactors.

(1) The extraordinary result of Connes in [Co1] that any subfactor of the hyperfinite  $\text{II}_1$  factor  $R$  is itself hyperfinite, hence isomorphic to  $R$  or finite dimensional.

(2) The growing conviction, expounded in [Co2], that  $\text{II}_1$  factors could be thought of as scalars, and Hilbert spaces carrying representations of them as "vector spaces" over the  $\text{II}_1$  factors. Hence the notation  $\dim_M(\mathcal{H})$  for the Murray-von Neumann coupling constant. By this analogy, the study of



subfactors corresponds to the rich subject of Galois theory.

(3) A nontrivial but barely noticed result of Goldman [Go] which is quite analogous to the fact that a subgroup of index 2 of a group is normal.

(4) Remarkable success, beginning with [Co3], in classifying group actions on the hyperfinite  $\text{II}_1$  factor.

The following invariant of a subfactor (up to conjugacy by automorphisms) was implicit in works mentioned above and was probably thought about by Murray and von Neumann.

**DEFINITION.** If  $N$  is a subfactor of the  $\text{II}_1$  factor  $M$  containing the same identity and if  $M$  acts on  $\mathcal{H}$  so that  $M'$  is a  $\text{II}_1$  factor, the real number  $\frac{\dim_N(\mathcal{H})}{\dim_M(\mathcal{H})}$  is independent of  $\mathcal{H}$  and is called the *index of  $N$  in  $M$*  or the *degree of the extension  $M$  of  $N$* . It is written  $[M:N]$  and trivially  $[M:N] \geq 1$ .

In ring-theoretic terms  $[M:N]$  measures the rank of  $M$  as a left  $N$ -module. Since  $K_0(N) = \mathbb{R}$ , this is a precise statement.

The simplest example is where  $M = N \otimes M_n(\mathbb{C})$ ,  $N$  being identified with  $N \otimes 1$ . Here  $[M:N] = n^2$ .

The next example is where  $N = M^G$ , the fixed point algebra for a group of automorphisms. Provided every element of  $G$  is outer (except the identity), we have  $[M:N] = |G|$ .

There are two extremes for subfactors, *irreducible* and *locally trivial*. An irreducible subfactor is one for which  $N' \cap M = \mathbb{C}1$ , and a locally trivial subfactor is one for which  $[pMp: pNp] = 1$  for all minimal projections in  $N' \cap M$ . The usefulness of the locally trivial subfactors has only recently become apparent in work of Popa ([Pop]).

The subfactors  $N = M^G$  as above are irreducible and show that the numbers 1, 2, and 3 are indices of subfactors. If a subfactor is locally trivial it is easy to see that its index, if different from 1, must be  $\geq 4$ . An obvious construction of locally trivial subfactors shows that the hyperfinite  $\text{II}_1$  factor has subfactors of index  $r$  for any  $r \geq 4$ . Thus one is tempted to ask, "what are the values of the index for subfactors?" The result, proved in [JO1] was a surprise. If  $[M:N] < 4$  then it must be of the form  $4 \cos^2 \pi/n$  for some integer  $n \geq 3$ , and all such values are realized for subfactors of the hyperfinite type  $\text{II}_1$  factor.

The situation for indices of irreducible subfactors is unclear. It seems plausible that there is a gap between 4 and the next value.

Up to now all of our results have been in the *internal* theory of von Neumann algebras. Historically there have been many connections with other subjects. We have already alluded to the connection with foliations in [Co2]. We are about to pursue what has been an extremely faithful connection with several subjects. This came from a rather unlikely source—a detail in the proof of the result on index values of subfactors. So we must begin that proof.

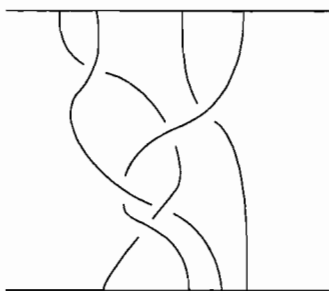
Starting with a subfactor  $N \subseteq M$ , a general result of [U] asserts that there is a "conditional expectation"  $e_N: M \rightarrow N$  which is an  $N-N$  bimodule map preserving 1. It follows that if  $[M:N] < \infty$ , the algebra of linear maps of  $M$  generated by  $e_N$  and  $M$  itself (acting by left multiplication) is again a  $\text{II}_1$  factor, with  $\text{tr}(e_N) = [M:N]^{-1}$ . Let us call this canonical construction  $\langle M, e_N \rangle$ . By results of [MvN] we have  $[\langle M, e_N \rangle: N]$ . One may now continue this process by considering  $M \subset \langle M, e_N \rangle$  and so on. One obtains a tower  $M_i$  of  $\text{II}_1$  factors defined by the second-order difference equation  $M_{i+1} = \langle M_i, e_{M_{i-1}} \rangle$  with initial conditions  $M_1 = N, M_2 = M$ . Letting  $e_i = e_{M_i}$  and noting that the trace  $\text{tr}$  is unique on a  $\text{II}_1$  factor, and hence well defined on  $\bigcup_i M_i$ , we have obtained, among other things, a sequence of operators  $e_i$  satisfying

- (i)  $e_i^2 = e_i^* = e_i$ ,
- (ii)  $e_i e_{i \pm 1} e_i = \tau e_i$  ( $\tau = [M:N]^{-1}$ ),
- (iii)  $e_i e_j = e_j e_i$  if  $|i-j| \geq 2$ ,
- (iv)  $\text{tr}(x e_{n+1}) = \tau \text{tr}(x)$  if  $x$  is a word on  $1, e_1, e_2, \dots, e_n$ .

The last property is called the Markov property because those  $e_i$ 's can be thought of as defining a "noncommutative Markov chain." The proof of the result on restrictions of the index is completed by examining the positivity condition  $\text{tr}(a^* a) > 0$  for  $a \neq 0$  for the algebra generated by the  $e_i$ 's.

Precisely the relations (i), (ii), and (iii) had been noticed and used by Temperley and Lieb in statistical mechanics in [TL], but this connection was not noticed at once. It will become highly significant later on. But the first connection to be observed (by colleagues in Geneva) was a similarity between (ii) and (iii) and Artin's presentation of the braid group. Thus I need to explain about the braid group.

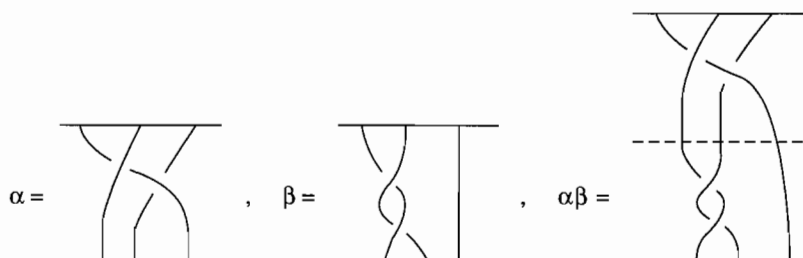
**Braids, knots, and links.** An  $n$ -string *braid* is a way of tying  $n$  points on a horizontal plane to the same points on another horizontal plane so that the height function on any string has no critical points. It is conventional, and apparently quite important, to arrange the  $n$  points on a straight line and to number them 1 to  $n$  in increasing order. Thus one may draw any braid as shown below.



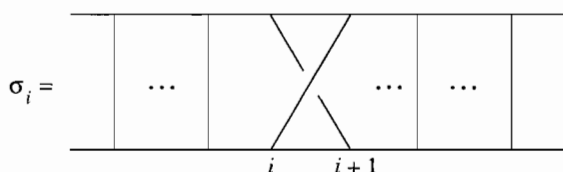
A four string braid.

Braids are considered up to an intuitive equivalence relation which may be rigorously defined using homotopy language.

The significant thing about braids on  $n$  strings is that they form a *group*  $B_n$  under concatenation as illustrated below:



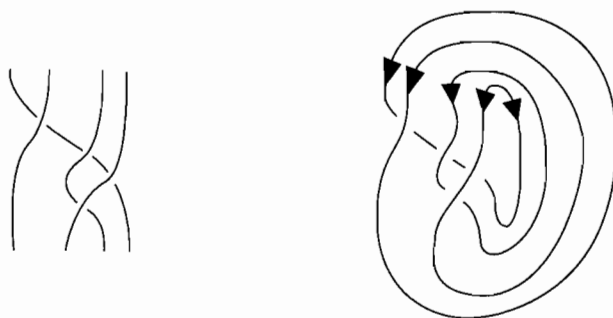
The following elementary braids generate the braid group:



Artin showed in [Ar] that the following is a presentation of  $B_n$ :

$$\langle \sigma_1, \sigma_2, \dots, \sigma_{n-1} \mid \sigma_i \sigma_{i+1} \sigma_i = \sigma_{i+1} \sigma_i \sigma_{i+1}, \sigma_i \sigma_j = \sigma_j \sigma_i \text{ for } |i - j| \geq 2 \rangle.$$

A knot in  $S^3$  is a smoothly embedded circle and a link is a nonintersecting union of knots. All are to be considered up to diffeomorphisms of  $S^3$ . Given a braid  $\alpha \in B_n$ , one may obtain an oriented link  $\hat{\alpha}$  in  $S^3$  as depicted below.



braid  $\alpha \in B_4$

oriented link  $\hat{\alpha}$

In the light of events since 1984, this process should be called taking the trace of the braid and the result as some kind of link valued character of the braid. But it is traditional (see [Bi]) to call  $\hat{\alpha}$  the closure of the braid and we shall use this terminology.

A result of Alexander [Al] asserts that any oriented link can be obtained in this way and Markov (son of the Markov of Markov chains!) showed that two braids  $x \in B_n$  and  $\beta \in B_m$  have the same closures if they can be connected by a sequence of "Markov moves" as follows:

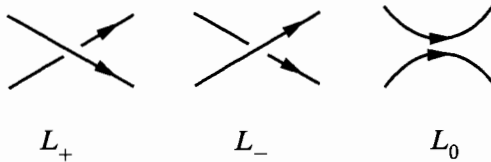
*type I Markov move:*  $\alpha \rightarrow \gamma \alpha \gamma^{-1}$ ,  $\alpha \in B_n$ ,  $\gamma \in B_n$ ,

*type II Markov move:*  $\alpha \leftrightarrow \alpha \sigma_n^{\pm 1}$ ,  $\alpha \in B_n$ , and  $B_n$  is embedded in  $B_{n+1}$  according to the points on the straight line.

The similarity between this business and relations (i)–(iv) in the algebra coming from subfactors gave, after much conversation between the author and topologists, especially Joan Birman, an invariant of oriented links in  $S^3$ , called  $V_L(t)$ , as follows (see [Jo2]).

Define a representation  $\pi$  of  $B_n$ , for all  $n$ , inside the algebra generated by the  $e_i$ 's coming from the subfactor proof by  $\pi(\sigma_i) = t^{1/4}(te_i - (1 - e_i))$ . If  $L$  is a link choose  $\alpha \in B_m$  with  $\hat{\alpha} = L$ . Then  $V_L(t) = (-\sqrt{t} - \frac{1}{\sqrt{t}})^{m-1} \text{tr}(\pi(\alpha))$  depends only on  $L$  by Markov's result.

Property (i),  $e_i^2 = e_i$ , is easily seen to imply that if  $L_+$ ,  $L_-$ , and  $L_0$  are links with diagrams differing only at one crossing as below,



then  $\frac{1}{t}V_{L_+} - tV_{L_-} = (\sqrt{t} - \frac{1}{\sqrt{t}})V_{L_0}$ . The Alexander polynomial  $\Delta_L$  satisfies  $\Delta_{L_+} - \Delta_{L_-} = (\sqrt{t} - \frac{1}{\sqrt{t}})\Delta_{L_0}$ . This relation, together with the normalization  $V_{\bigcirc} \equiv 1$ ,  $\Delta_{\bigcirc} \equiv 1$  suffices to calculate these polynomials. One finds

$$V_{\bigcirc} = t + t^3 - t^4, \quad \Delta_{\bigcirc} = t^{-1} - 1 + t.$$

Immediately after its discovery  $V$  was generalized in [F+] to a two-variable version  $P(l, m)$ , embracing the Alexander polynomial, with defining formula  $l^{-1}P_{L_+} - lP_{L_-} = mP_{L_0}$ . This polynomial  $P$  contains significant information about turning links into braids [Mo], [FW].

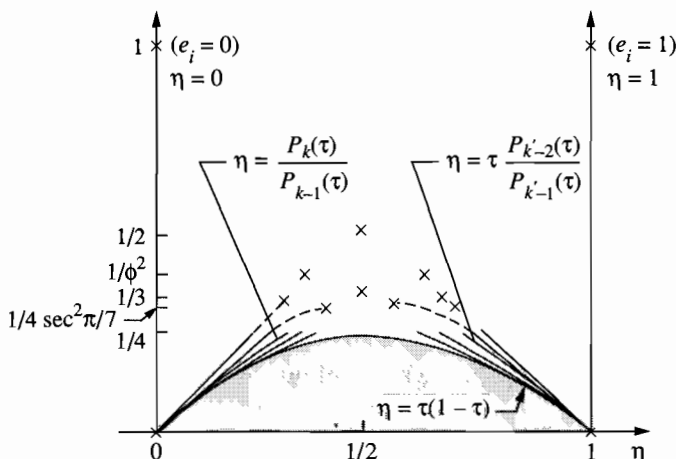
Oceanu's approach to defining  $P_L$  was important for von Neumann algebras for he introduced the Hecke algebra of type  $A_n$  with presentation  $\langle g_1, g_2, \dots, g_{n-1} | g_i^2 = (q-1)g_i + q, g_i g_{i+1} g_i = g_{i+1} g_i g_{i+1}, g_i g_j = g_j g_i \text{ for } |i-j| > 1 \rangle$

$|i - j| \geq 2$ ) and defined  $P$  via a trace on this algebra as we did above with the Temperley-Lieb algebra (which is a Hecke algebra quotient). The trace is defined like condition (iv) above:

$$\text{tr}(w g_n) = z \text{tr}(w)$$

where  $w$  is a word on  $g_1, g_2, \dots, g_{n-1}$  and  $z$  is a free parameter. Ocneanu also showed that the Hecke algebra for  $A_\infty$ , together with this trace, defines a  $\Pi_1$  factor precisely for the following values.

$$\tau^{-1} = 2 + q + q^{-1}, \eta =$$



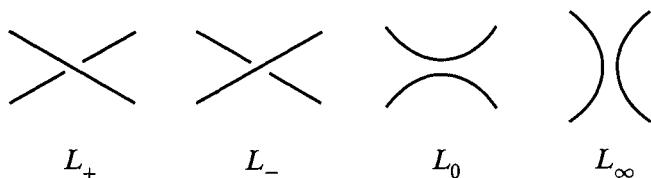
One can define a subfactor by taking the algebra generated by  $g_2, g_3, g_4, \dots$ . Wenzl [We1] calculated the index for  $\tau^{-1} = 4 \cos 2\pi/n$ ,  $\eta$  to be  $\frac{\sin^2 k\pi/n}{\sin^2 \pi/n}$ . He also proved that these subfactors are irreducible.


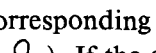
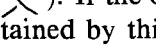
The most spectacular success of the new polynomials has been the proof of some Tait conjectures ([K2], [Mus], [Th]). We have the following remarkable inequality, true for a link diagram  $L$  with  $n$  crossings:

- (a)  $\deg(V_L(t)) \leq n$ ,
- (b)  $\deg(V_L(t)) = n$  iff  $L$  is alternating and reduced.

Here “deg” means, for instance,  $\deg(t + t^3 - t^4) = 4 - 1 = 3$ . Alternating means that over-crossings and under-crossings alternate as any component of the link is followed and “reduced” means that the diagram admits no obvious simplification. It follows immediately from (a) and (b) that if the link  $L$  has a reduced alternating diagram with  $n$  crossings then it has no diagram at all with fewer crossings. For an account of this see [HKW].

**The Kauffman polynomial.** Following the discovery of  $V$  and  $P$ , a polynomial of unoriented links  $Q_L$  was found in [BLM, Ho]. It is defined by the relation  $Q_{L_+} + Q_{L_-} = x(Q_{L_0} + Q_{L_\infty})$ , where  $L_+$ ,  $L_-$ ,  $L_0$ , and  $L_\infty$  are



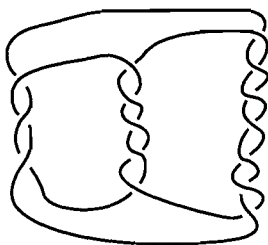
as shown: (and  $Q_\circ \equiv 1$ ). A sample calculation gives  $Q_\circ = 2x^2 + 2x - 3$ . Kauffman saw how to give a two-variable version of  $Q$  by following the inductive scheme for calculating  $Q$ , but whenever a picture like  is encountered it is to be replaced by ; the corresponding polynomial must be multiplied by a factor “ $a$ ” (or  $a^{-1}$  for ). If the original link  $L$  is then given an orientation, the polynomial obtained by this computation may be multiplied by a power of “ $a$ ” to give the Kauffman polynomial  $F_L(a, x)$ . The Kauffman polynomial contains  $V_L$  as a specialization but not  $\Delta$ . See [K2].

There are many open questions about these polynomials. The most obvious are:

(I) Is there a nontrivial knot  $K$  with  $V_K(t) \equiv 1$ ? (Same for  $P$ ,  $Q$ ,  $F$ ).

(II) Can one characterize which Laurent polynomials arise as  $V_K(t)$  for some knot? (Same for  $P$ ,  $Q$ ,  $F$ ).

Both these questions were answered for the Alexander polynomial in a paper by Seifert in 1934 [S]. The knot below has trivial Alexander polynomial.



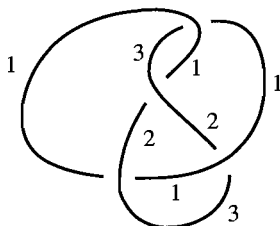
A Laurent polynomial  $p(t)$  is  $\Delta_K(t)$  for some  $K$  iff  $p(1) = 1$  and  $p(t) = p(t^{-1})$ . Seifert solved these problems by understanding the cyclic covering spaces of the knot complement in a very concrete way using an oriented surface whose boundary is the knot.

These problems seem out of reach for  $V$  at the moment. Barring an accidental discovery, a new idea is needed in the understanding of these polynomials. Everything seems to indicate that the place to look is physics, as I shall now explain.

**Statistical mechanical models.** We now begin to return to physics. In classical statistical mechanics a system may be defined by its set of *states* and a function  $E(\sigma)$  which assigns to a state  $\sigma$  its energy. The *partition func-*

tion of the system is then defined to be  $Z = \sum_{\text{states } \sigma} \exp\left(-\frac{E(\sigma)}{kT}\right)$ . One is typically interested in very large but regular systems (e.g., on a square lattice) so the calculation of  $Z$  often proceeds by introducing some algebra of matrices. This is the context in which Temperley and Lieb discovered the interest of relations (i), (ii), and (iii) for the  $e_i$ 's. They used these relations to show the equivalence between two models (self-dual Potts and 6-vertex). In this algebra the partition function will typically be the trace of some matrix representing the system. This suggests looking for a statistical mechanical model to give the polynomials we have discussed. This suggestion was totally compelling in the light of an explicit formula for  $V_L$  given by Kauffman [K1]. Indeed a glance at Chapter 12 of [Ba] shows that Kauffman's "states model" is precisely an intermediate step in an elegant combinatorial proof of Temperley-Lieb equivalence.

Thus one is led to define vertex models and spin models directly on link diagrams (see [Jo3]). The vertex model scheme is as follows. We begin with two sets of "Boltzmann weights"  $w_{\pm}(a, b|x, y)$ , where  $a, b, x, y$  are "spin" indices, say running from 1 to  $N$ . If  $L$  is a link diagram we define a state of it to be a function from the edges of the diagram to  $\{1, 2, \dots, N\}$ . A state of the figure eight is depicted below:



Given a state every crossing is surrounded by a configuration  $\begin{smallmatrix} a & y \\ & x \\ & b \end{smallmatrix}$ . One then takes the product  $\prod_{\text{crossings}} w_{\pm}(a, b|x, y)$ , using the "+" weights at a positive crossing and the "-" weights at a negative one. The partition function  $Z_L$  is then defined to be  $Z_L = \sum_{\text{states}} \prod_{\text{crossings}} w_{\pm}(a, b|x, y)$ . (In fact another term is needed to make the following discussion correct but I shall ignore this for simplicity. See [Tu, Re] or [Jo3].)

One may now inquire as to whether  $w_{\pm}(a, b|x, y)$  may be chosen in such a way that  $Z$  depends only on the link and not on the chosen diagram. The answer is yes and an abundance of examples may be constructed using the theory of quantum groups. Indeed the index sets  $\{1, \dots, N\}$  may be chosen differently for different components of the link. The final result, proved in [Ro, Re] is as follows: let  $\mathfrak{g}$  be a simple complex Lie algebra and let  $L$  be a link with distinguished components  $c_1, c_2, \dots, c_n$ . Then to every way of assigning finite dimensional representations  $\pi_1, \dots, \pi_n$  to  $c_1, \dots, c_n$  there is an invariant  $Z_L^{\mathfrak{g}}(\pi_1, \pi_2, \dots, \pi_n, q)$  defined as the partition function for some choice of  $w_{\pm}(a, b|x, y)$ .

Quantum groups were actually constructed with a view to defining and understanding statistical mechanical models for which the large scale behavior of the partition function can be calculated explicitly. A condition that seems to guarantee this has been emphasized by Baxter and is known as the Yang-Baxter equation. The reason that quantum groups apply to knot theory is that the main condition on the Boltzmann weights that ensures topological invariance of the partition function is a weak form of the Yang-Baxter equation. For references on quantum groups see [Dr, Ji1].

Is this appearance of the statistical mechanical formalism a coincidence? At this stage one cannot give a definite answer to this question but one may begin by seeing if the correspondence can be pushed further. Vertex models are not the only kind used in statistical mechanics. There are also "spin" models, where the spins live on the vertices of a graph, and IRF models, where the energy of a state is the sum of energy contributions from the faces of a graph. Both of these kinds of models adapt beautifully to knot theory. There are two simple spin models which give the  $V$  polynomial and the homology of the 2-fold branched cover of  $S^3$  branched over the link respectively. See [Jo3]. The IRF models give what is probably the most economical model (in terms of number of states) known. A beautiful account occurs in [KR].

I have not yet explained how the invariants coming from quantum groups relate to the polynomials  $V$ ,  $P$ ,  $Q$ , and  $F$ . The simplest case is  $V_L(t)$  which is  $Z_L^{\text{sl}_2}(\pi_2, \pi_2, \dots, \pi_2, t)$ ,  $\pi$  being the 2-dimensional defining representations. Also  $Z_L^{\text{sl}_n}(\pi_n, \pi_n, \dots, \pi_n)$  is  $P_L(q^{(n+1)/2}, q^{1/2} - q^{-1/2})$  and the orthogonal and symplectic algebras, with the defining representation assigned to all components of the link, give an infinite sequence of specializations of the Kauffman polynomial.

Note that if the knot diagram is a closed braid, the partition function  $Z$  is easily seen to be the trace of the braid in a representation of the braid group coming from the  $w_{\pm}(a, b|x, y)$  on  $\bigotimes^n V$ , where  $\sigma_i$  acts on  $\bigotimes^n V$  by

$$\begin{aligned} \sigma_i(v_1 \otimes \cdots \otimes v_i \otimes v_{i+1} \otimes \cdots \otimes v_n) \\ = \sum w(i, j|i+1, j+1)(v_1 \otimes \cdots \otimes v_j \otimes v_{j+1} \otimes \cdots \otimes v_n). \end{aligned}$$

This brings us back to the braid way of calculating the polynomials. In fact one has to modify the trace a little for this to work. One obtains "states" on  $\bigotimes_{n=1}^{\infty} \text{End}(V)$  which give rise naturally to type  $\text{III}_{\lambda}$  factors,  $0 < \lambda < 1$  (see [Pow]).

**Quantum invariant theory.** A central problem in classical invariant theory is to decompose the tensor powers of a given irreducible representation of a group. This is of course equivalent to finding the commutant algebras. The simplest case is for  $\text{sl}_n$  in its defining representation  $V$ . The symmetric group  $S_m$  acts naturally on  $\bigotimes^m V$  and it is well known that the algebra



generated by  $S_m$  is the commutant of the tensor product action of  $\mathfrak{sl}_n$ . Thus the group algebra of  $S_m$  is a universal object for the commutants. For groups other than  $\mathrm{SL}_n$  it is necessary to find other invariants to generate the whole commutant. Brauer gave a universal algebra for the commutants of the orthogonal and symplectic groups in [Br]. An old problem about this algebra was solved by Wenzl using the techniques described in this talk—see [We2].

If we replaced  $\mathfrak{G}$  by the quantum group  $U_h(\mathfrak{G})$  one may ask the same question about the commutant. Jimbo proved in [Ji2] that the commutant of  $U_h(\mathfrak{sl}_n)$  on  $\bigotimes^m V$  is generated by the braid group representation coming from the representation on  $\bigotimes^m V$  we have just described! Thus the  $P$  polynomial sits in a sense in the commutant of  $\mathfrak{sl}_n$ .

The situation for the orthogonal and symplectic algebras is even more interesting. We will see that the Kauffman polynomial plays exactly the same role as the  $P$  polynomial did for  $\mathfrak{sl}_n$ . In [BW, Mus], an algebra, the BMW algebra, was invented to play the role for the Kauffman polynomial that the Hecke algebra plays for  $P$ . The idea, due to Kauffman, was to extend the braid group by allowing objects

$$E_i =$$

as well as  $\sigma_i$ 's in braid words. To mimic the Kauffman polynomial definition, an algebra with generators  $G_i, E_i, i = 1, \dots, n-1$  was defined with relations

$$G_i G_{i+1} G_i = G_{i+1} G_i G_{i+1}, \quad G_i G_j = G_j G_i \quad \text{if } |i-j| \geq 2,$$

$$G_i + G_i^{-1} = x(1 + E_i),$$

$$E_i G_i = G_i E_i = a E_i,$$

$$E_i^2 = (a + a^{-1} - x)x^{-1} E_i,$$

$$E_i G_{i\pm 1}^{\pm 1} E_i = a^{\mp 1} E_i,$$

$$E_i G_{i\pm 1} G_i = E_i E_{i\pm 1}.$$

These relations can be interpreted by drawing pictures. A trace was defined on this algebra using the Kauffman polynomial, with the property that the appropriately normalized trace of a braid is the Kauffman polynomial of the closure of the braid.

By results of Jimbo in [Ji1], the braid group representations coming from the quantum versions of the orthogonal and symplectic groups in their defining representations, factor through the BMW algebra in the obvious way. It is surely true that they generate the commutants of the quantum group representations, though we have not seen a complete proof of this.

It seems important to decide in general how much of the commutant is generated by the braid matrices. For  $\mathfrak{sl}_n$ , they generate the whole commutant in all irreducible representations (see [KR]) but this fails for the adjoint representation of  $\mathfrak{sl}_3$ . On the other hand, if one includes the whole family  $R(\lambda)$ ,  $\lambda$  being the spectral parameter, one will obtain more of the commutant. The significance of this question is that all the classical invariants might appear simply as limits of a very simple picture on the quantum level.

**Quantum field theory.** One reason for the interest in statistical mechanical models and their phase transitions is that at the critical point they are supposed to give nontrivial examples of continuum quantum field theories by letting the lattice spacing tend to zero. In two dimensions the resulting field theory should carry a projective unitary representation of the conformal group. The corresponding (complexified) Lie algebra is the direct sum of two copies of the "Virasoro algebra" with basis  $\{c, L_n; n \in \mathbb{Z}\}$  with Lie bracket  $[L_n, L_m] = (n-m)L_{n+m} + (\frac{m^3-m}{12})\delta_{n,-m}c$ ,  $[c, L_n] = 0$  for all  $n$ . In [FQS] it was shown that if  $c < 1$  in a unitary  $(L_n^* = L_{-n})$  irreducible representation of the Virasoro algebra then  $c = 1 - \frac{6}{n(n+1)}$ . This result is analogous to the result for indices of subfactors, less than 4. This, and other ideas, suggest that the way to make a direct connection between knot theory and von Neumann algebras is via quantum field theory.

Witten, [Wi] has given a definition of the knot polynomials as partition functions for three-dimensional quantum field theories with Chern-Simons action in a way intimately related with the Wess-Zumino-Witten (conformal) quantum field theories. A great advantage of this approach is that it works for links in any oriented three-manifold, and, for the empty link, even gives an invariant of considerable interest. See Witten's lecture in this series.

In a rather different approach, Fredenhagen, Longo, Rehren, and Schroer [FRS, L] have discovered the braiding structure and indices for subfactors to be consequences of an axiomatic approach to "superselection sectors" in algebraic quantum field theory. The equivalence, beyond some bounded region, of a representation of the field algebra with the vacuum representation gives an endomorphism of the local field algebra, and the braiding operators come from the representation of the Lorentz group. Fröhlich has had ideas along similar lines, see [F].

These connections with quantum field theory are currently under active study and it is hoped that a beautiful unified theory will appear.

## REFERENCES

- [Al] J. Alexander, *A lemma on systems of knotted curves*, Proc. Nat. Acad. Sci. U.S.A. **9** (1923), 93–95.
- [Ar] E. Artin, *Theory of braids*, Ann. of Math. (2) **48** (1967), 101–126.
- [Ba] R. Baxter, *Exactly solved models in statistical mechanics*, Academic Press, 1982.
- [Bi] J. Birman, *Braids, links and mapping class groups*, Ann. of Math. Stud., no. 82, Princeton Univ. Press, Princeton, NJ, 1974.
- [BLM] R. Brandt, W. B. R. Lickorish, and K. Millett, *A polynomial invariant for unoriented knots and links*, Invent. Math. **84** (1986), 563–573.
- [Br] R. Brauer, *On algebras which are connected with the semisimple continuous groups*, Ann. of Math. (2) **38** (1937), 854–872.
- [BW] J. Birman and H. Wenzl, *Braids, link polynomials and a new algebra*, Trans. Amer. Math. Soc.
- [Co1] A. Connes, *Classification of injective factors*, Ann. of Math. (2) **106** (1976), 73–115.
- [Co2] —, *Sur la theorie non commutative de l'integration*, Lecture Notes in Math., vol. 725, Springer, 1979, pp. 19–163.
- [Co3] —, *Periodic automorphisms of the hyperfinite factor of type  $II_1$* , Acta Sci. Math. **39** (1977), 39–66.
- [Dr] V. Drinfeld, *Quantum groups*, Proc. Internat Congr. Math., vol. 1, 1986, pp. 798–820.
- [F+] P. Freyd, D. Yetter, J. Hoste, W. Lickorish, K. Millett, and A. Ocneanu, *A new polynomial invariant of knots and links*, Bull. Amer. Math. Soc. (N.S.) **12** (1988), 183–192.
- [FQS] D. Friedan, Z. Qiu, and S. Shenker, *Conformal invariance, unitarity and two dimensional critical exponents*, Vertex Operators in Mathematics and Physics, MSRI Ser., Springer, 1986, pp. 419–447.
- [FRS] K. Fredenhagen, K.-H. Rehren and B. Schroer, *Superselection sectors with braid group statistics and exchange algebras*, Comm. Math. Phys. **125** (1989), 201–226.
- [FW] J. Franks and R. Williams, *Braids and the Jones-Conway polynomial*, Trans. Amer. Math. Soc. **303** (1987), 97–108.
- [F] J. Fröhlich, *Statistics of Fields, the Yang-Baxter equation and the theory of knots and links*, Proceedings Cargèse (G. 't Hooft, et al. eds.), 1987.
- [Go] M. Goldman, *On subfactors of type  $II_1$* , Michigan Math. J. **7** (1960), 167–172.
- [HKW] P. de la Harpe, M. Kervaire, and C. Weber, *On the Jones polynomial*, Enseigne. Math. **32** (1986), 271–335.
- [Ha] U. Haagerup, *Connes bicentralizer problem and the uniqueness of the injective factor of type  $II_1$* , Acta Math. **158** (1987), 95–148.
- [HS] U. Haagerup and J. Schou, *Some new subfactors of the hyperfinite  $II_1$  factor*, preprint, 1989.
- [Ho] C. F. Ho, *A new polynomial for knots and links—preliminary report*, Abstracts Amer. Math. Soc. **6** (1985), 300.
- [Ji1] M. Jimbo, *A  $q$ -difference analogue of  $U(\mathcal{Z})$  and the Yang-Baxter equation*, Lett. Math. Phys. **102** (1986), 537–567.
- [Ji2] —, *A  $q$ -analogue of  $U(\mathfrak{gl}(N+1))$ , Hecke algebra and the Yang-Baxter equation*, Lett. Math. Phys. **11** (1986), 247–252.
- [Jo1] V. Jones, *Index for subfactors*, Invent. Math. **72** (1983), 1–25.
- [Jo2] —, *A polynomial invariant for knots via von Neumann algebras*, Bull. Amer. Math. Soc. (N.S.) **12** (1985), 103–111.
- [Jo3] —, *On knot invariants related to some statistical mechanical models*, Pacific J. Math. **137** (1989), 311–334.
- [K1] L. Kauffman, *State models and the Jones polynomial*, Topology **26** (1987), 395–401.
- [K2] —, *New invariants in the theory of knots*, Amer. Math. Monthly **95** (1988), 195–242.
- [KR] A. N. Kirillov and N. Yu. Reshetikhin, *Representations of the algebra  $U_q(\mathfrak{sl}_2)$ ,  $q$ -orthogonal polynomials and invariants of links*, LOMI, preprint, 1988.
- [L] R. Longo, *Index of subfactors and statistics of quantum fields. I*, Comm. Math. Phys. **126** (1989), 217–247.
- [MvN] F. Murray and J. von Neumann, *On rings of operators*, Ann. of Math. (2) **37** (1936), 116–229.

- [Mo] H. Morton, *Closed braid representations for a link and its 2-variable polynomial*, preprint, Liverpool, 1985.
- [Mus] K. Murasugi, *Jones polynomials and classical conjectures in knot theory*, *Topology* **26** (1987), 187–194.
- [Pi] C. Piron, *Foundations of quantum physics*, Benjamin, 1976.
- [Pow] R. Powers, *Representations of uniformly hyperfinite algebras and their associated von Neumann rings*, *Ann. of Math. (2)* **86** (1967), 138–171.
- [Pop] S. Popa, *Sousfacteurs, actions des groupes et cohomologie*, *C.R. Acad. Sci. Paris Sér. I. Math.* **309** (1989), 771–776.
- [Re] N. Reshetikhin, *Quantized universal enveloping algebras, the Yang-Baxter equation and invariants of links. I, II*, LOMI, preprints, 1988.
- [Ro] M. Rosso, *Groupes quantiques et modèles à vertex de V. Jones en théorie des noeuds*, *C.R. Acad. Sci. Paris Sér. I Math.* **307** (1988), 207–210.
- [S] H. Seifert, *Über das Geschlecht von Knoten*, *Math. Ann.* **110** (1934), 571–592.
- [TL] H. Temperley and E. Lieb, *Relations between the percolation and ...*, *Proc. Roy. Soc. London Ser. A* **322** (1971), 251–280.
- [Th] M. Thistlethwaite, *A spanning tree expansion of the Jones polynomial*, *Topology* **26** (1987), 297–309.
- [Tu] V. Turaev, *The Yang-Baxter equations and invariants of links*, *Invent. Math.* **92** (1988), 527–553.
- [U] H. Umegaki, *Conditional expectations in an operator algebra. I*, *Tôhoku Math. J.* **6** (1954), 358–362.
- [vN1] J. von Neumann, *Mathematical foundations of quantum mechanics* (translated from German), Princeton Univ. Press, Princeton, NJ, 1955.
- [vN2] —, *Über adjungierte Funktionaloperatoren*, *Ann. of Math. (2)* **33** (1932), 294–310.
- [vN3] —, *On rings of operators. Reduction theory*, *Ann. of Math. (2)* **50** (1969), 401–485.
- [We1] H. Wenzl, *Hecke algebras of type  $A_n$  and subfactors*, *Invent. Math.* **92** (1988), 349–383.
- [We2] —, *On the structure of Brauer's centralizer algebras*, *Ann. of Math. (2)* **128** 1988, 173–193.
- [Wi] E. Witten, *Quantum field theory and the Jones polynomial*, *Comm. Math. Phys.* **121** (1989), 351–399.

## Modular Invariance in Mathematics and Physics

VICTOR G. KAC

In this talk I want to discuss some recently discovered beautiful connections of representation theory of infinite-dimensional Lie algebras with the theory of modular functions, and related progress in theoretical physics.

**1. Modular functions.** Consider a finite-dimensional space of complex analytic functions on the upper half-plane, having at worst a pole at  $i\infty$ , and suppose that this space is invariant under transformations

$$(1)_w \quad f(\tau) \mapsto f(\tau + 1) \quad \text{and} \quad f(\tau) \mapsto \tau^{-w} f(-1/\tau).$$

These functions are then called *modular functions* of weight  $w \in (\frac{1}{2}\mathbb{Z})$ .

To illustrate the idea of modular invariance, consider the classical partition function  $p(n)$ , the number of partitions of  $n$  into a sum of positive integers. Its generating series  $1 + \sum_{n=1}^{\infty} p(n)x^n$  is equal to  $1/\phi(x)$ , where  $\phi(x) = \prod_{n \geq 1} (1 - x^n)$ . The key observation is that the closely related *Dedekind  $\eta$ -function*  $\eta(\tau) = q^{1/24} \phi(q)$ , where  $q = e^{2\pi i \tau}$ , is a modular function of weight  $1/2$  since it has the following modular invariance property:

$$(2) \quad \eta(\tau) = (-i\tau)^{-1/2} \eta(-1/\tau).$$

Let  $\tau = i\beta$  ( $\beta$  is the standard notation of the inverse of the temperature in statistical mechanics). Then, by (2), we have asymptotically as  $\beta \downarrow 0$ :

$$(3) \quad \eta(i\beta)^{-1} \sim \beta^{1/2} e^{\pi/12\beta}.$$

Applying the standard Tauberian theorem which relates the asymptotics of a series  $\sum_n a_n x^n$  as  $x \rightarrow 1$  to the asymptotics of  $a_n$  as  $n \rightarrow \infty$ , we obtain the following classical result:

$$p(n) \sim \frac{1}{4\sqrt{3}n} e^{\pi\sqrt{2n/3}} \quad \text{as } n \rightarrow \infty.$$

Similarly, given an integral lattice  $\Lambda$  of rank  $l$  in the euclidean space

---

1980 *Mathematics Subject Classification* (1985 Revision). Primary 17B65, 17B15, 17B81, 17-02; Secondary 17B67, 17B68, 11F03.

Supported in part by NSF grant DMS 8802489.

$\mathbb{R}^l$ , we can count the number of vectors of any given length by studying the associated *theta series*:

$$\theta_{\Lambda}(\tau) = \sum_{\gamma \in \Lambda} q^{|\gamma|^2/2}.$$

This is a modular function of weight  $l/2$ . Indeed, the linear span of "generalized" theta series  $\{\theta_{\Lambda+a}(\tau) | a \in \Lambda^* \bmod \Lambda\}$  is invariant with respect to transformations  $(1)_{l/2}$  (this is obvious for the first transformation, and follows from the Poisson summation formula for the second). For example, taking  $\Lambda = \mathbb{Z}^l$ , one can study the number of ways  $n$  can be represented as a sum of  $l$  squares.

Note that by Euler's identity

$$(4) \quad \phi(q) = \sum_{n \in \mathbb{Z}} (-1)^n q^{(3n^2+n)/2},$$

it follows that

$$\eta(\tau) = \sum_{n \in \mathbb{Z}} (-1)^n q^{3(n+1/6)^2/2}$$

is a difference of two (generalized) theta series, hence is a modular function (of weight  $1/2$ ).

Another famous example is the *modular invariant*  $j = \theta_{E_8}^3 / \eta^{24}$  (which parametrizes elliptic curves), where  $E_8$  is the only even unimodular lattice in eight dimensions. The function  $j(\tau)$  is invariant under transformations  $(1)_0$ , and, moreover, generates the ring of all such functions.

On recent developments in number theory based on modular invariance (including Ribet's proof of Fermat's last theorem modulo the Taniyama-Weil conjecture), see the address of B. Gross.

Note that a modular function  $f$  of weight  $-w$  has asymptotics similar to (3), namely, for some numbers  $d$  and  $g$  one has:

$$(5) \quad f(i\beta) \sim d\beta^{w/2} e^{\pi g/12\beta} \quad \text{as } \beta \downarrow 0.$$

**2. Infinite-dimensional Lie algebras.** The most important infinite-dimensional Lie algebras  $\mathfrak{g}$  possess a derivation (Hamiltonian)  $H$  satisfying the following three properties:

- (i)  $\mathfrak{g} = \bigoplus_{j \in \mathbb{Z}} \mathfrak{g}_j$ , where  $\mathfrak{g}_j = \{x \in \mathfrak{g} | [H, x] = jx\}$  and  $\mathfrak{g}_1 \neq 0$ ;
- (ii)  $\mathfrak{g}$  has only trivial graded (with respect to (i)) ideals;
- (iii)  $\dim \mathfrak{g}_j < \text{const.}$

According to the Kac-Mathieu classification theorem, there are only two possibilities for such algebras:

- (a) loop algebras (the Lie algebra of regular maps from  $\mathbb{C}^\times$  to complex simple finite-dimensional Lie algebras) and their "twisted" analogs;
- (b) the Lie algebra of regular vector fields on  $\mathbb{C}^\times$ .

In our experience, one loses most (if not all) interesting representations, unless one considers central extensions of the Lie algebras in question. We

explain below the construction of (â) the *affine Kac-Moody algebras*  $\hat{\mathfrak{g}}$ , which are universal central extensions of (nontwisted) loop algebras, and of (b) the *Virasoro algebra*  $\text{Vir}$ , the universal central extension of (b).

EXAMPLE (â).  $\hat{\mathfrak{g}} = (\bigoplus_{n \in \mathbb{Z}} \mathfrak{g}_{(n)}) \oplus \mathbb{C}k$ , where, for all  $n$ , we have  $\mathfrak{g}_{(n)} = \mathfrak{g}$ , a simple Lie subalgebra of  $\mathfrak{gl}_N(\mathbb{C})$ , with the following commutation relations:

$$(6) \quad [x_{(m)}, y_{(n)}] = [x, y]_{(m+n)} + m\delta_{m, -n}(x, y)k, \quad [k, \hat{\mathfrak{g}}] = 0;$$

here  $x, y \in \mathfrak{g}$  and  $(x, y) = \text{const tr } xy$  (const = 1 for  $\mathfrak{g} = \mathfrak{sl}_N(\mathbb{C})$ ). The Hamiltonian  $H = H_0 + \text{ad } z_{(0)}$ , where  $[x_{(m)}, H_0] = mx_{(m)}$  and  $z \in \mathfrak{g}$  is a real diagonal matrix. Geometric interpretation:

$$\hat{\mathfrak{g}}/\mathbb{C}k = \bigoplus \mathfrak{g}_{(n)} = \text{Map}(\mathbb{C}^\times, \mathfrak{g}) = \mathbb{C}[t, t^{-1}] \otimes \mathfrak{g}, \quad x_{(n)} = t^n x.$$

EXAMPLE (b).  $\text{Vir} = (\bigoplus_{n \in \mathbb{Z}} \mathbb{C}L_n) \oplus \mathbb{C}c$ , with commutation relations

$$(7) \quad [L_m, L_n] = (m - n)L_{m+n} + \frac{m^3 - m}{12}\delta_{m, -n}c, \quad [c, \text{Vir}] = 0.$$

The Hamiltonian  $H = L_0$ . Geometric interpretation:

$$\text{Vir}/\mathbb{C}c = \text{Vect } \mathbb{C}^\times, \quad L_n = -t^{n+1} \frac{d}{dt}.$$

**3. Positive energy representations.** It has been clear for some time now, both to mathematicians and to physicists, that the most interesting representations of an infinite-dimensional Lie algebra are the *positive energy representations*. These are the representations in a vector space  $V$  for which the Hamiltonian  $H$  is diagonalizable and  $\text{Spec } H$  is a set of real numbers bounded below. The function  $\text{tr}_V q^H$  is called the *character* of this representation. An important and still somewhat mysterious fact is that in all known examples, the characters of irreducible positive energy representations have asymptotics of the form (5) (i.e., behave as modular functions at the high temperature limit), and that the numbers  $d$ ,  $w$ , and  $g$ , called respectively the *asymptotic dimension*, the *weight*, and the *growth* of  $V$ , have a group theoretical interpretation. For example, the asymptotic dimension  $d$  has all the properties of usual dimension, although it is an irrational number in general. Actually, for this reason,  $d$  is a much more powerful invariant than the usual dimension.

In all known examples of irreducible positive energy representations of infinite-dimensional Lie algebras, the number  $w$  is nonnegative (and the numbers  $d$  and  $g$  are positive), in sharp contrast with the finite-dimensional case, when  $\text{tr}_V q^H \sim e\beta^{-d}$  as  $\beta \downarrow 0$  and  $d$  is a nonnegative integer (called the Bernstein-Gelfand-Kirillov dimension).

Sometimes a character becomes a modular function when multiplied by  $q^{-a/24}$ , where  $a$  is some number (the *modular anomaly*); the result is then called the *modified character*. The corresponding representation is said to

be *modular invariant*. (Note that modular invariance implies energy positivity.) It is the modular invariant representations that have played a fundamental role in the recent development of representation theory of infinite-dimensional Lie algebras and groups on the one hand, and of quantum field theory on the other.

**4. A toy example.** An affine algebra may be viewed as a “nonabelian” generalization of the *oscillator algebra* (= Heisenberg algebra)  $\hat{a}$ , the affine algebra associated to the 1-dimensional Lie algebra  $\mathfrak{a} = \mathbb{C}s$ . We have:

$$[s_{(m)}, s_{(n)}] = m\delta_{m, -n}k, \quad [s_{(m)}, k] = 0.$$

It is very easy to describe all irreducible positive energy representations  $V$  of  $\hat{a}$ : Either  $k = 0$ , then  $V$  is the trivial 1-dimensional representation, or  $k \neq 0$ , then  $V = \mathbb{C}[x_1, x_2, \dots]$ , and  $s_{(m)} = \partial/\partial x_m$ ,  $s_{(-m)} = kmx_m$  for  $m > 0$ ;  $s_0 = \mu$ ;  $H_0 = \sum_{j \geq 1} jx_j \partial/\partial x_j$ . These are the canonical commutation relations representations. All of them are modular invariant with modified character  $\eta(\tau)^{-1}$ , so that  $d = w = g = a = 1$ . This example, which is very important in quantum field theory, is, from a purely mathematical point of view, not just a toy model. It also serves as a basis for rather nontrivial constructions, such as the vertex operator construction and its variations. Along these lines one constructs the modular invariant representation of the famous Monster group, whose modified character is  $j(\tau)$ .

**5. Modular invariant representations of the affine algebra  $\hat{g}$ .** There are some general experimental facts (most of them conjectures) about an irreducible positive energy representation  $V$  of  $\hat{g}$ :

- (a)  $\text{rank } \mathfrak{g} \leq g(V) \leq \dim \mathfrak{g}$ ;
- (b)  $0 \leq w(V) \leq \dim \mathfrak{g}$ ;
- (c)  $w(V) = 0 \Rightarrow V$  is modular invariant;
- (d)  $g(V) < \dim \mathfrak{g} \Leftrightarrow V$  is modular invariant with  $w(V) < \dim \mathfrak{g}$ .

In our experience, the “smaller” a representation is, the more interesting it becomes. Since the growth is a reasonable measure of the “size” of a representation, we may conclude that the most interesting are the irreducible modular invariant representations of zero weight. An amazing fact is that even in the simplest case  $\mathfrak{g} = \mathfrak{sl}_2(\mathbb{C})$ , the theory of such representations is remarkably rich (in sharp contrast with the “abelian case” when no such representation exists).

Positive energy representations  $V$  of  $\hat{g}$  are parametrized by the central charge  $k$  (= the eigenvalue of the operator  $k$ ) and by the (finite-dimensional irreducible) representation  $\lambda$ , on the minimal energy subspace  $V_h$  of the Hamiltonian  $H_0 + z$ , of the algebra of “internal symmetries”  $\{x \in \mathfrak{g}_{(0)} | [z, x] = 0\}$ . (In unitary theories one always takes  $z = 0$ , otherwise one needs to take  $z \neq 0$  to avoid divergences.)

The first problem I am going to address is to determine for which pairs  $(k, \lambda)$  the corresponding representations  $V(k, \lambda)$  is modular invariant of



zero weight. An easy necessary condition is

$$(8) \quad k + h' \text{ is a nonnegative rational number,} \\ \text{where } h' \text{ is the dual Coxeter number of } \mathfrak{g} \text{ (} h' = N \text{ for } \mathfrak{g} = \mathfrak{sl}_N(\mathbb{C}) \text{).}$$

Consider also the following condition:

$$(9) \quad \langle \hat{\lambda} + \hat{\rho}, \alpha \rangle \in \mathbb{Q} \setminus \{0, -1, -2, \dots\} \quad \text{for all } \alpha \in R, \\ \text{where } R \text{ is the set of "positive real coroots" of } \hat{\mathfrak{g}}.$$

Provided that (8) and (9) hold, one has an explicit character formula (due to Kac and Wakimoto) of the following form:

$$(10) \quad \mathrm{tr}_{V(k, \lambda)} q^H = \sum_{y \in W^{k, \lambda}} \pm \Theta_{y, k, \lambda} / \prod_{\alpha} ,$$

where  $\Theta_{y, k, \lambda}$  are certain theta series and  $\prod_{\alpha}$  is certain infinite product over all "positive roots" of  $\hat{\mathfrak{g}}$ . Here  $\hat{\lambda} + \hat{\rho}$  is the "shifted highest weight" of  $V(k, \lambda)$  (which is simply expressed in terms of  $k$  and  $\lambda$ ) and  $W^{k, \lambda}$  is the associated subgroup of the "Weyl group"  $W$  of  $\mathfrak{g}$ .

But if  $\lambda$  is the trivial 1-dimensional representation of  $\hat{\mathfrak{g}}$  and  $k = 0$ , then  $\dim V(k, \lambda) = 1$ , hence  $\mathrm{tr}_V q^H = 1$  and we obtain from (10) an identity of the form:

$$(11) \quad \prod_{\alpha} = \sum_{y \in W} \pm \Theta_y.$$

These are the well-known Macdonald identities.

It follows from (10) and (11) that, provided that (8) and (9) hold, the representation  $V(k, \lambda)$  is modular invariant; moreover, the following condition guarantees that  $w(V(k, \lambda)) = 0$ :

$$(12) \quad \text{the } \mathbb{Q}\text{-span of the set } \{\alpha \in R \mid \langle \hat{\lambda} + \hat{\rho}, \alpha \rangle \in \mathbb{Z}\} \text{ contains } R.$$

The main *conjecture* asserts that conditions (8), (9), and (12) give us the complete list of all modular invariant representations of  $\hat{\mathfrak{g}}$  of weight 0 with  $k \neq -h'$ . This conjecture has been checked for  $\widehat{\mathfrak{sl}}_2$ , Vir, and in some other cases.

**6. The example of  $\widehat{\mathfrak{sl}}_2$ .** To avoid technicalities, I have given only a flavor of the main results for general affine algebras. Now I will state these results explicitly in the simplest case of the affine algebra  $\widehat{\mathfrak{sl}}_2$ .

The complete list of zero weight modular invariant irreducible representations  $V(k, \lambda)$  is as follows: the central charge  $k$  is a rational number  $u'/u$ , in lowest terms, such that

$$(13) \quad k + 2 \geq 2/u;$$

all possible  $\lambda$  (= the maximal eigenvalue of the element  $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$  on the subspace of minimal energy) for this  $k$  are

$$(14) \quad \lambda = n - s(k + 2),$$

where  $n, s \in \mathbb{Z}$ ,  $0 \leq n \leq u(k+2) - 2$ ,  $0 \leq s \leq u - 1$ .

The character formula (10) (taking into account (11)) for such a representation is:

$$(15) \quad \text{tr}_{V(k, \lambda)} q^{H_0 + \frac{1}{2}z(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}) - (\frac{1}{24}a - \frac{1}{4}z^2)} = \frac{(\Theta_{b_+, b}(\tau, \frac{z}{u}) - \Theta_{b_-, b}(\tau, \frac{z}{u}))}{(\Theta_{1,2}(\tau, z) - \Theta_{-1,2}(\tau, z))}.$$

Here  $\Theta_{n,m}(\tau, z) = \theta_{(n/m+z)/2+\mathbb{Z}}(m\tau)$  is a Jacobi-Riemann theta function, and

$$b = u^2(k+2), \quad b_{\pm} = u(\pm(n+1) - s(k+2)), \quad a = 3 - 6b_{+}^2/b.$$

One also has:

$$d(V(k, \lambda)) = \sqrt{\frac{2}{b}} \sin \frac{\pi(n+1)}{u(k+2)} \sin \frac{\pi(s+z)}{u} / \sin \pi z;$$

$$g(V(k, \lambda)) = 3 - 6/b.$$

Given  $k$ , the space spanned by all modified characters of modular invariant representations is invariant with respect to modular transformations  $(1)_0$ . Explicitly, denoting the left-hand side of (15) by  $\chi_{\lambda}$ , one has

$$(16) \quad \chi_{\lambda} \left( -\frac{1}{\tau}, -\tau z \right) = \left( \exp \frac{i\pi k z^2 \tau}{2} \right) \left( \frac{2}{b} \right)^{1/2} \\ \times \sum_{\lambda'} \frac{e^{-i\pi b_{+} b'_{-}/b} - e^{-i\pi b_{+} b'_{+}/b}}{2i} \chi_{\lambda'}(\tau, z),$$

where  $\lambda'$  runs over the list (14).

Results similar to these hold for all affine algebras (see Kac-Peterson and Kac-Wakimoto).

All experimental facts (a)–(d) stated at the beginning of this section hold for  $\widehat{\mathfrak{sl}}_2$ . In particular, the following statements are equivalent for an irreducible positive energy representation  $V$  of  $\widehat{\mathfrak{sl}}_2$ :

- (i)  $g(V) < 3$ ;
- (ii)  $w(V) = 0$ ;
- (iii)  $V$  is modular invariant of weight 0.

The Macdonald identity (11) is, in this case, the celebrated Jacobi triple product identity:

$$\prod_{n=1}^{\infty} (1 - u^n v^n) (1 - u^n v^{n-1}) (1 - u^{n-1} v^n) = \sum_{n \in \mathbb{Z}} (-1)^n u^{n(n+1)/2} v^{n(n-1)/2}.$$

Letting  $u = q$ ,  $v = q^2$  we deduce Euler's identity (4), and letting  $u = v = -q$ , we obtain a product expansion  $\theta_{\mathbb{Z}}(\tau) = \eta^2((\tau+1)/2)/\eta(\tau+1)$ .

**7. Some applications.** The most popular representation is the modular invariant representation of minimal possible growth  $l = \text{rank } \mathfrak{g}$ . This is the

basic representation  $V = V(1, \lambda = \text{trivial})$  (and the ones obtained from  $V$  by a simple twist) of the affine algebra  $\hat{\mathfrak{g}}$ , where  $\mathfrak{g}$  is one of the simple classical Lie algebras  $\mathfrak{sl}_{l+1}(\mathbb{C})$ ,  $\mathfrak{so}_{2l}(\mathbb{C})$ , or the exceptional ones  $E_6, E_7, E_8$ . The modified character of this representation is given by an especially simple formula (Kac):

$$(17) \quad \text{tr}_{V_Q} q^{H_0 + z - (l/24 - (z, z)/2)} = \frac{\theta_{Q+z}(\tau)}{\eta(\tau)^l},$$

where  $Q$  is the root lattice of  $\mathfrak{g}$ .

Note that since  $H_0$  commutes with  $\mathfrak{g} \subset \hat{\mathfrak{g}}$ , the  $H_0$ -eigenspace decomposition  $V = \bigoplus_{j \in \mathbb{Z}_+} V_j$  is invariant with respect to  $\mathfrak{g}$ . Thus we obtain a series of finite-dimensional representations of  $\mathfrak{g}$  ( $\dim V_0 = 1$ ,  $\dim V_1 = \dim \mathfrak{g}$ , ...) such that the generating series  $\sum_{j \in \mathbb{Z}_+} (\dim V_j) q^{j-1/24}$  is equal to  $\theta_Q/\eta^l$ . In particular, for  $E_8$  this series is equal to  $j(\tau)^{1/3}$ . Moreover, replacing  $(\dim V_j)$  by  $\text{tr}_{V_j} g$ , where  $g$  is a finite order element of the Lie group  $G$  with Lie algebra  $\mathfrak{g}$ , we obtain a modular function of weight 0.

In a remarkable parallel development, McKay, Thompson, Conway, and Norton observed that the sporadic finite simple Monster group  $F_1$  has a series of finite-dimensional representations  $V_0, V_1 = 0, V_2, V_3, \dots$ , such that the generating series  $\sum_{j \in \mathbb{Z}_+} (\dim V_j) q^{j-1}$  equals  $j(\tau) - 744$  and that, moreover, the series  $\sum_{j \in \mathbb{Z}_+} (\text{tr}_{V_j} g) q^{j-1}$  is a modular function of weight 0 and "genus" 0 for all  $g \in F_1$ ; almost all modular functions of "genus" 0 occur in this way. The latter phenomenon is yet to be explained.

The basic representation admits a large variety of explicit constructions. One starts with a regular loop  $s(t) \in \mathbb{C}[t, t^{-1}] \otimes \mathfrak{g} (= \hat{\mathfrak{g}}/Ck)$  (i.e., for any  $t_0 \in \mathbb{C}^\times$ , the centralizer of  $s(t_0)$  in  $\mathfrak{g}$  consists of commuting diagonalizable elements); its centralizer  $\bar{\mathfrak{s}}$  is commutative and its preimage  $\mathfrak{s}$  in  $\hat{\mathfrak{g}}$  is a Heisenberg algebra. Let  $\bar{S}$  denote the centralizer of  $\bar{\mathfrak{s}}$  in  $G(\mathbb{C}[t, t^{-1}])$ . An important property of the basic representation (proved by Kac and Peterson) is its irreducibility with respect to the pair  $(\mathfrak{s}, \bar{S})$ . (For example, taking  $s(t)$  to be a regular constant diagonal matrix, we obtain the homogeneous Heisenberg subalgebra and the irreducibility theorem follows from (17).) This allows one to identify  $V$  with the space of the oscillator representation. The oscillator representation extends to the basic representation of  $\hat{\mathfrak{g}}$  by making use of the vertex operators of string theory. The vertex operators are characterized by the property of being eigenvectors with respect to  $\mathfrak{s}$  and  $S$ , and are, up to a simple factor, of the form

$$\left( \exp \sum \lambda_i x_i \right) \left( \exp \sum \mu_i \frac{\partial}{\partial x_i} \right).$$

The vertex operator construction attached to the homogeneous Heisenberg subalgebra (the Frenkel-Kac-Segal construction) in the case  $\mathfrak{g} = E_8$  is an important part of the "heterotic" string model used in compactification

from 26 to 10 dimensions by Gross-Harvey-Martinec-Rohm. A twist of this construction applied to Griess's algebra produces the series of representations of the Monster group mentioned above (Frenkel-Lepowsky-Meurman and Borcherds).

Another beautiful application of the basic representation, to the soliton theory, was discovered by Sato, Date, Jimbo, Kashiwara, and Miwa. An analysis of their work shows that their approach is based on the following simple observation: Let  $G$  be a group acting on a vector space  $V$ , let  $v_0 \in V$ , and let  $\Omega$  be an operator on  $V \otimes V$  commuting with the (diagonal) action of  $G$  and such that  $v_0 \otimes v_0$  is its eigenvector with eigenvalue  $\lambda$ . Then an element  $f$  of the orbit  $G \cdot v_0$  satisfies the equation

$$(18) \quad \Omega(f \otimes f) = \lambda f \otimes f.$$

For example, we can take  $\Omega$  to be the "Casimir operator" and  $v_0$  to be the vacuum vector.

Applying this construction to  $G = \text{SL}_2(\mathbb{C}[t, t^{-1}])$  and  $V$  its basic (projective) representation, we obtain two quite different systems of partial differential equations depending on the construction of  $V$ . If the construction is based on the homogeneous Heisenberg subalgebra of  $\widehat{\text{sl}}_2$ , then equation (18) turns into the NLS hierarchy, the simplest equations being (after a change of functions) the coupled nonlinear Schrödinger equations on functions  $g(t, x)$  and  $g^*(t, x)$ :

$$g_t = g_{xx} - 2g^2 g^*, \quad g_t^* = -g_{xx}^* + 2g g^{*2}.$$

If the construction is based on the other Heisenberg subalgebra, associated to  $s(t) = \begin{pmatrix} 0 & 1 \\ t & 0 \end{pmatrix}$  (the principal construction), the equation (18) turns into the KdV hierarchy, the simplest equation being (after a change of functions) the celebrated KdV equation:

$$u_t = \frac{3}{2}uu_x + \frac{1}{4}u_{xxx}.$$

Moreover, the  $N$ -soliton solutions of these equations can be obtained by making use of vertex operators. For example, if  $X(z)$  is the vertex operator of the principal construction

$$X(z) = \left( \exp 2 \sum_{\substack{j \geq 1 \\ j \text{ odd}}} z^j x_j \right) \left( \exp -2 \sum_{\substack{j \geq 1 \\ j \text{ odd}}} \frac{z^{-j}}{j} \frac{\partial}{\partial x_j} \right),$$

then  $N$ -soliton solutions of the KdV equation are given by the following formula (in which  $x = x_1$ ,  $t = x_3$ ):

$$u(t, x) = 2(\log((1 + a_N X(z_N)) \cdots (1 + a_1 X(z_1)) \cdot 1))_{xx}.$$

These solutions describe the interaction of  $N$  solitary waves.

This theory has an important connection to the theory of theta-functions. Namely, equation (18) in the case of the basic representation of the group

$GL_\infty$  is nothing else but the KP hierarchy, which characterizes theta functions of algebraic curves among all theta functions (Arbarello-De Concini, Mulase, Shiota).

Apart from the basic representation, the only positive energy irreducible representations of growth  $< 2l$  have central charge  $k$  equal to 2. This case plays an important role in various supersymmetric theories.

The modular invariance constraint is an important ingredient of the representation theory itself. The basic observation here is that given an affine algebra  $\hat{\mathfrak{g}}$  and its affine subalgebra  $\hat{\mathfrak{g}}_1$  the branching coefficients of an irreducible unitary (and hence modular invariant) positive energy representation of  $\hat{\mathfrak{g}}$  restricted to  $\hat{\mathfrak{g}}_1$  are coefficients of modular functions, called *branching functions*, with well studied transformation properties. In the simplest example  $\mathfrak{g} = \mathfrak{sl}_2(\mathbb{C})$  and  $\mathfrak{g}_1 = \mathbb{C} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ , these modular functions turn out to be the classical indefinite Hecke modular forms divided by  $\eta^2$  (Kac-Peterson). This has been exploited by Date, Jimbo, Miwa, and their collaborators to calculate explicitly the local height probabilities in solvable lattice models. It should be pointed out, however, that even in this simplest example, the branching functions of a nonunitary modular invariant representation are not always modular functions (Lu), and their transformation properties are unknown.

The  $SL_2(\mathbb{Z})$ -invariance of the space of modified characters of modular invariant representations with given central charge  $k$  has important applications to 2-dimensional quantum field theories, since it allows one to compute explicitly their partition functions (Gepner, Capelli-Itzykson-Zuber, ...).

In general, a conformally invariant 2-dimensional quantum field theory (CFT) produces a finite set of modular functions  $f_0, \dots, f_N$  whose  $\mathbb{C}$ -span is invariant with respect to transformations  $(1)_0$  and such that

$$f_i(\tau) = \sum_{n \in \mathbb{Z}_+} a_n^{(i)} q^{n-r_i}, \quad \text{where } a_n^{(i)} \in \mathbb{Z}_+, \quad r_i \in \mathbb{Q}, \quad \text{and } r_0 > r_1 \geq r_2 \geq \dots$$

These data contain the most important information of this field theory, such as conformal anomaly, conformal dimensions, the partition function (Cardy), the fusion rules (Verlinde), etc. These properties of  $f_0, \dots, f_N$  alone turn out to be very restrictive, and, for small  $N$ , allow an effective classification algorithm (Mathur-Mukhi-Sen). Given  $\hat{\mathfrak{g}} \supset \hat{\mathfrak{g}}_1$  and  $k \in \mathbb{Z}_+$ , taking branching functions of all unitary modular invariant representations of  $\hat{\mathfrak{g}}$  with central charge  $k$  with respect to  $\hat{\mathfrak{g}}_1$  we get a finite set of functions satisfying the above properties (Kac-Peterson), obtaining thus a large variety of CFT.

**8. Modular invariant representations of Vir.** Positive energy irreducible representations of Vir are parametrized by two numbers, the *conformal anomaly*  $c$  (= the eigenvalue of  $c$ ) and the minimal energy  $h$  of the Hamiltonian  $L_0$ . A complete list of pairs  $(c, h)$  such that the corresponding

representation  $V_{c,h}$  is modular invariant of weight 0 is as follows:

$$(19) \quad c = 1 - \frac{6(m-n)^2}{mn}, \quad h = \frac{(mr - ns)^2 - (m-n)^2}{4mn},$$

where  $m, n, r, s$  are positive integers such that  $(m, n) = 1$ ,  $r < n$ ,  $s < m$ ,  $sn \leq rm$ .

The character formula for arbitrary  $V_{c,h}$  was given by Feigin and Fuchs. For representations from the list (19) this formula turns into

$$(20) \quad \text{tr}_{V_{c,h}} q^{L_0 - c/24} = (\theta_{(mr - ns)/2mn + \mathbb{Z}}(mn\tau) - \theta_{(mr + ns)/2mn + \mathbb{Z}}(mn\tau)) / \eta(\tau).$$

For these representations we have

$$(21) \quad g(V_{c,h}) = 1 - \frac{6}{mn}.$$

The following properties of a representation  $V = V_{c,h}$  are equivalent:

- (i)  $g(V) < 1$ ;
- (ii)  $w(V) = 0$ ;
- (iii)  $V$  is modular invariant.

According to Belavin, Polyakov, and Zamolodchikov each  $c$  from the list (19) corresponds to a solvable 2-dimensional statistical model, the  $h$ 's corresponding to the critical exponents of this model.

Note that  $g(V) \leq 1/2$  in two cases only:  $c = 1/2$ , which corresponds to the Ising model and  $c = -22/5$ , which corresponds to the Lee-Yang edge singularity model (Cardy). The Ising model is the first member of the sequence of unitary statistical models (Friedan-Qiu-Shenker) and the Lee-Yang model is the first member of a very interesting sequence of nonunitary models with

$$c = 1 - \frac{3(m-2)}{m} \quad (m = 5, 7, 9, \dots),$$

$$h = h_s = -\frac{(m-s-1)(s-1)}{2m} \quad \left(s = 1, 2, \dots, \frac{m-1}{2}\right).$$

In this case the character admits a simple product decomposition:

$$(22) \quad \text{tr}_{V_{c,h}} q^{L_0 - h_s} = \prod_{\substack{n \geq 1 \\ n \not\equiv 0, \pm s \pmod{m}}} (1 - q^n)^{-1}.$$

Note that this product, in the case  $m = 5$ , is precisely the product part of the celebrated Rogers-Ramanujan identities (and in the cases  $m > 5$  that of their Gordon generalizations). A natural conjecture is that these identities provide bases of representation spaces; in particular, in the case  $m = 5$ , vectors  $\cdots L_{-j_3} L_{-j_2} L_{-j_1} |h_s\rangle$  with  $j_1 + j_2 + \cdots = n$ ,  $j_2 \geq j_1 + 2$ ,  $j_3 \geq j_2 + 2$ ,  $\dots$  and  $j_1 \geq 2$  (resp.  $\geq 1$ ) when  $s = 0$  (resp.  $= 1$ ), should form a basis of the subspace of energy  $h_s + n$ . In the same spirit, every modular invariant representation of Vir should produce a Rogers-Ramanujan type identity.

Further development of the Belavin-Polyakov-Zamolodchikov approach to the 2-dimensional conformal field theory gave the following simple construction of the partition function  $Z(\tau)$  on a torus (Cardy). Fix the conformal anomaly  $c$ , and let  $\chi_{(h)}(\tau) = \text{tr}_{V_{c,h}} q^{L_0 - c/24}$  be the modified character. Then  $Z(\tau)$  must be a (real analytic) function of the form

$$(23) \quad Z(\tau) = \sum_{h, h'} a_{h, h'} \chi_{(h)}(\tau) \overline{\chi_{(h')}(\tau)},$$

where  $a_{h, h'}$  are nonnegative integers,  $a_{0,0} = 1$ , and  $Z(\tau)$  is modular invariant (i.e., invariant under transformations  $(1)_0$ ).

If  $(c, h)$  is from the list (19), then all modified characters (with given  $c$ ) form a basis of a vector space invariant under transformations  $(1)_0$  and the corresponding matrices are unitary in this basis. It follows that  $Z(\tau) = \sum_h |\chi_{(h)}(\tau)|^2$  is a partition function. A complete classification of partition functions (23) with  $c < 1$  (and hence from the list (19)) is obtained in two steps. First, the problem is reduced to a similar problem for  $\widehat{\mathfrak{sl}}_2$  and unitary characters (Gepner). Fix a positive integer  $k$ , and for an integer  $\lambda$  from (14), i.e.,  $0 \leq \lambda \leq k$ , let

$$\chi_\lambda(\tau) = \text{tr}_{V(k, \lambda)} q^{H_0 - a/24}.$$

Then we have the following special case of (16) (Kac-Peterson):

$$(24) \quad \chi_\lambda\left(-\frac{1}{\tau}\right) = \left(\frac{2}{k+2}\right)^{1/2} \sum_{\lambda'=0}^k \left(\sin \pi \frac{(\lambda+1)(\lambda'+1)}{k+2}\right) \chi_{\lambda'}(\tau).$$

The problem of determining all partition functions of the form (23) with  $\chi_{(h)}$  replaced by  $\chi_\lambda$  was solved by Cappelli-Itzykson-Zuber. A remarkable fact (which is yet to be explained) is that these partition functions are labeled by Dynkin diagrams, so that the constants  $a_{\lambda, \lambda'}$  in (23) are expressed in terms of "exponents" associated to these diagrams.

The transformation law (24) has reappeared recently in the beautiful work of Witten on Jones polynomials (see addresses of Witten and Jones).

**9. The Ising model.** This is a very instructive example of an explicit construction of all modular invariant irreducible representations of the Virasoro algebra with conformal anomaly  $c = 1/2$ .

Fix  $\delta = 0$  (the "Ramond sector") or  $\delta = 1/2$  (the "Neveu-Schwarz sector"), and let  $U_\delta$  be the algebra over  $\mathbb{C}$  on anticommuting indeterminates  $\xi_j$  ( $j \in \delta + \mathbb{Z}_+$ ). Define the following operators on  $U_\delta$ :

$$\begin{aligned} \psi_0 &= 2^{-1/2} \left( \xi_0 + \frac{\partial}{\partial \xi_0} \right); & \psi_n &= \frac{\partial}{\partial \xi_n}, & \psi_{-n} &= \xi_n \quad \text{for } n > 0; \\ L_0 &= \frac{1-2\delta}{16} + \frac{1}{2} \sum_{j \in \delta + \mathbb{Z}_+} j \psi_{-j} \psi_j, & L_k &= \frac{1}{2} \sum_{j \in \delta + \mathbb{Z}} j \psi_{-j} \psi_{j+k} \quad \text{for } k \neq 0. \end{aligned}$$

Using that  $[\psi_m, L_k] = (m + \frac{1}{2}k)\psi_{m+k}$ , we find

$$[L_m, L_n] = (m - n)L_{m+n} + \delta_{m, -n} \frac{m^3 - m}{24}.$$

Thus, we have constructed a representation of Vir on  $U_\delta$  with conformal anomaly  $c = 1/2$ . This representation is not irreducible: the subspaces  $U_\delta^+$  and  $U_\delta^-$  of elements of even and odd degree respectively are invariant. One can show that

$$(25) \quad U_{1/2}^+ = V_{1/2, 0}, \quad U_{1/2}^- = V_{1/2, 1/2}, \quad U_0^+ = U_0^- = V_{1/2, 1/16},$$

obtaining thereby all modular invariant irreducible representations of Vir with  $c = 1/2$ . In fact, we see from (25) that

$$(26) \quad \chi_{(0)} \pm \chi_{(1/2)} = q^{-1/48} \prod_{n \in \mathbb{Z}_+} (1 \pm q^{n+1/2}), \quad \chi_{(1/16)} = q^{1/24} \prod_{n \in \mathbb{Z}_+} (1 + q^{n+1}).$$

This can be rewritten in terms of the  $\eta$ -function:

$$(27) \quad \begin{aligned} \chi_{(0)} + \chi_{(1/2)} &= \frac{\eta(\tau)^2}{\eta(\frac{1}{2}\tau)\eta(2\tau)}, & \chi_{(0)} - \chi_{(1/2)} &= \frac{\eta(\frac{1}{2}\tau)}{\eta(\tau)}, \\ \chi_{(1/16)} &= \frac{\eta(2\tau)}{\eta(\tau)}. \end{aligned}$$

It is clear from (26) and (27) that the 3-dimensional subspace spanned by  $\chi_{(0)}$ ,  $\chi_{(1/2)}$ , and  $\chi_{(1/16)}$  is invariant with respect to the modular transformations  $(1)_0$ .

**10. Modular invariance versus unitarity.** Let  $\omega$  be a conjugate linear anti-automorphism of  $\hat{\mathfrak{g}}$  (resp. Vir) defined by  $\omega(x_{(n)}) = {}^t\bar{x}_{(-n)}$ ,  $\omega(k) = k$  (resp.  $\omega(L_n) = L_{-n}$ ,  $\omega(c) = c$ ). A representation of  $\hat{\mathfrak{g}}$  or Vir in a vector space  $V$  is called unitary if  $V$  carries a positive definite Hermitian form for which the operators  $g$  and  $\omega(g)$  are adjoint ( $g \in \hat{\mathfrak{g}}$  or Vir). (For example, taking in the Ising model all monomials in  $\xi_i$  for an orthonormal basis, we see that the representations of Vir in  $U_\delta$  are unitary).

Unitary irreducible positive energy representations of an affine algebra  $\hat{\mathfrak{g}}$  are all modular invariant (with weight 0). We have seen that modular invariance is given by certain rationality conditions ((8) and (9)). Unitarity turns out to be given by certain integrality conditions. For example, the modular invariant representations  $V(k, \lambda)$  of  $\hat{\mathfrak{sl}}_2$ , given by (13) and (14), are unitary if and only if  $k$  and  $\lambda$  are nonnegative integers and  $\lambda \leq k$ . Representation  $V_{c,h}$  of Vir for  $c \geq 1$  is always unitary (Kac), and, for  $c < 1$  is unitary if and only if  $(c, h)$  is one of the pairs from the list (19) such that  $|m - n| = 1$  (Friedan-Qiu-Shenker, Goddard-Kent-Olive, Kac-Wakimoto).

The unitary representations of the affine algebra  $\hat{\mathfrak{g}}$  always give rise to a representation of the corresponding group  $\hat{G}$  (= a central extension of  $G(\mathbb{C}[t, t^{-1}])$  by  $\mathbb{C}^\times$ ). This is not true for nonunitary modular invariant



representations. The representation theory of affine algebras is thus much richer than that of the corresponding groups.

The unitary positive energy representations have been much studied in the past decade both by mathematicians and physicists. One may expect that the more universal class of modular invariant representations will keep both groups busy for years to come. I have discussed here only the genus one case and have not even touched the superalgebra case!

## BIBLIOGRAPHY

### Books:

- G. E. Andrews, *The theory in partitions*, Encyclopedia of Mathematics, vol. 2, 1976.  
 I. Frenkel, J. Lepowsky, and A. Meurman, *Vertex operator algebras and the Monster*, Academic Press, 1989.  
 M. B. Green, J. H. Schwarz, and E. Witten, *Superstring theory*, Cambridge Univ. Press, 1987.  
 V. G. Kac, *Infinite-dimensional Lie algebras*, Progr. in Math., vol. 44, Birkhäuser, Boston, 1983; Second ed., Cambridge Univ. Press, 1985.  
 V. G. Kac and A. K. Raina, *Bombay lectures on highest weight representations*, Adv. Ser. in Math. Phys., vol. 2, World Scientific, 1987.  
 M. Knopp, *Modular functions in analytic number theory*, Markham, Chicago, 1970.  
 A. Pressley and G. Segal, *Loop groups*, Oxford Univ. Press, 1986.

### A selection of papers:

- E. Arbarello and C. De Concini, *Another proof of Novikov's conjecture on periods of abelian integrals on Riemann surfaces*, Duke Math. J. **54** (1987), 163–178.  
 A. A. Belavin, A. M. Polyakov, and A. B. Zamolodchikov, *Infinite conformal symmetry in two-dimensional quantum field theory*, Nuclear Phys. B **241** (1984), 333–380.  
 A. Capelli, C. Itzykson, and J. B. Zuber, *The A-D-E classification of minimal and  $A_1^{(1)}$  conformal invariant theories*, Comm. Math. Phys. **112** (1987), 1–26.  
 J. Cardy, *Operator content of two-dimensional conformally invariant theories*, Nuclear Phys. B **270** (1986), 186–204.  
 J. H. Conway and S. P. Norton, *Monstrous moonshine*, Bull. London Math. Soc. **11** (1979), 308–339.  
 E. Date, M. Jimbo, M. Kashiwara, and T. Miwa, *Transformation groups for soliton equations*, Proc. RIMS Sympos. (M. Jimbo and T. Miwa, eds.), World Scientific, 1983, pp. 39–120.  
 E. Date, M. Jimbo, A. Kuniba, T. Miwa, and M. Okado, *Exactly solvable SOS models: local height probabilities and theta function identities*, Nuclear Phys. B **290** (1987), 231–273.  
 B. L. Feigin, and D. B. Fuchs, *Verma modules over the Virasoro algebra*, Lecture Notes in Math., vol. 1060, Springer, 1984, pp. 230–245.  
 D. Friedan, Z. Qiu, and S. Shenker, *Conformal invariance, unitarity and two dimensional critical exponents*, Publ. MSRI, No. 3, 1985, pp. 419–449.  
 D. Gepner and E. Witten, *String theory on group manifold*, Nuclear Phys. B **278** (1986), 493–520.  
 P. Goddard and D. Olive, *Kac-Moody and Virasoro algebras in relation to quantum physics*, Internat. J. Modern Phys. A **1** (1986), 303–414.  
 V. G. Kac, *Highest weight representations of infinite dimensional Lie algebras*, Proc. Internat. Congr. Math., Helsinki, 1978, pp. 299–304.  
 V. G. Kac and D. H. Peterson, *Infinite dimensional Lie algebras, theta functions and modular forms*, Adv. in Math. **53** (1984), 125–264.  
 ———, *112 constructions of the basic representation of the loop group of  $E_8$* , Proc. Conf. Anomalies, geometry, topology (Argonne, 1985), World Scientific, 1985, pp. 276–298.  
 V. G. Kac and M. Wakimoto, *Modular and conformal invariance constraints in representation theory of affine algebras*, Adv. in Math. **70** (1988), 156–236.

——, *Modular invariant representations of infinite dimensional Lie algebras and superalgebras*, Proc. Nat. Acad. Sci. U.S.A. **85** (1988), 4956–4960.

——, *Exceptional hierarchies of soliton equations*, Proc. Sympos. Pure Math., Amer. Math. Soc., Providence, RI, 1989.

O. Mathieu, *Classification des algèbres de Lie graduées simples de croissance  $\leq 1$* , Invent. Math. **86** (1986), 371–426.

S. D. Mathur, S. Mukhi, and A. Sen, *On the classification of rational conformal field theories*, Phys. Lett. B **213** (1988), 303–308.

E. Verlinde, *Modular transformations and the operator algebra in 2D conformal field theory*, Nuclear Phys. B **300** (1988), 360–375.

E. Witten, *Physics and geometry*, Proc. Internat. Congr. Math., Berkeley, CA, 1986.

DEPARTMENT OF MATHEMATICS, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, CAMBRIDGE, MASSACHUSETTS 02139

# **Mathematical Fluid Dynamics: the Interaction of Nonlinear Analysis and Modern Applied Mathematics**

ANDREW J. MAJDA

## CONTENTS

1. The elementary mathematical structure of fluid flows
  - (A) The equations for compressible flow
  - (B) Nonlinear sound waves and vorticity waves
2. Physical phenomena and mathematical theory for nonlinear sound waves:  
Recent progress and future directions
  - (A) The zero diffusion approximation: the mathematical theory of conservation laws in a single space variable
  - (B) Structure and stability of wave patterns in several space variables
  - (C) Nonlinear geometric optics
  - (D) Nonlinear diffraction at caustics and boundaries
3. Vorticity waves and turbulence
  - (A) The equations of compressible and incompressible flow
  - (B) An outstanding open problem: Breakdown for the 3-D Euler equations
  - (C) Current turbulence theories and modern mathematical physics
4. Examples of the interaction between large scale computing and modern mathematical theory: Vortex sheets in two distinct regimes of fluid motion
  - (A) The nonlinear evolution of vortex sheets for 2-D incompressible flow
  - (B) The nonlinear development of instability for supersonic vortex sheets

---

1980 *Mathematics Subject Classification* (1985 Revision). Primary 76L05, 35L65.

Research partially supported by N.S.F. DMS 8702864, A.R.O. DAAL03-89-K-0013, O.N.R. N00014-89-J-1044, D.A.R.P.A. N00014-86-K-0759.

©1992 American Mathematical Society  
0-8218-0167-8 \$1.00 + \$.25 per page

**Introduction.** It is not difficult to understand the importance of research in fluid mechanics. The world that surrounds us involves both air and water, two primary examples of fluids. Problems of understanding fluid flow continue to have a fundamental role in both technological and basic scientific research. Such problems include the accurate tracking of hurricane paths, the nature of blood flow through the heart, the efficient mixing of fuel in internal combustion engines, and the flight of aircraft through the atmosphere at a wide variety of speeds. For example, these speeds range from the very high re-entry velocities of the space shuttle to the flight conditions near the speed of sound for typical commercial jumbo jets to the hazardous swirling flows generated by slow-moving airplanes landing at airports.

Applied mathematicians have been primary contributors to the theoretical understanding and explanation of phenomena in fluid flows through their research in mathematical fluid dynamics. It might surprise some readers that over the last two centuries, many luminaries in pure mathematics have made decisive contributions to mathematical fluid mechanics including Euler, Riemann, Weyl, Friedrichs, Courant, von Neumann, Kolmogorov, Hopf, and Leray among others. For example, a rather large portion of von Neumann's collected works involves his research in fluid mechanics. As an area of mathematical research, fluid mechanics provides rich examples of nonlinear partial differential equations with features of hyperbolic, elliptic, and mixed type depending on the various regimes of fluid motion being investigated. Such examples will be pointed out throughout this paper. The wonderful collection of experimental photographs in the book of Van Dyke [1] provides dramatic experimental confirmation for the diversity of nonlinear phenomena which occur in fluid flow in various regimes; we often refer to specific photographs from this book throughout the various sections of this paper when we discuss the mathematical structure.

The rapid development of high speed computers over the last forty years has irreversibly altered the way in which applied mathematicians do research in fluid mechanics. In fact, one common contemporary point of view states that detailed mathematical analysis in applied mathematics will rapidly become obsolete since soon there will be a "Cray supercomputer in every kitchen." In this paper, I will strongly advocate just the opposite viewpoint. The rapid evolution of applied mathematics through large scale computation reveals new phenomena in fluid flow, in some instances beyond the capability of experimental measurements. Explaining and controlling these complex phenomena requires new mathematical ideas from nonlinear analysis, differential equations, probability theory, and geometry which simultaneously interact in a highly interdisciplinary fashion with both methods for computation and more traditional tools of applied mathematics such as asymptotic methods. Thus, this increasingly broad interactive mode of research blends results from

- (0.1) (1) Large Scale Numerical Computation  
(2) Asymptotic Modelling  
(3) Qualitative Modelling  
(4) Rigorous Mathematical Analysis for Prototype Problems

in explaining phenomena in fluid flow. Such a new mode of modern applied mathematics in close alliance with developments in theoretical mathematics is likely to reach a mature state by the beginning of the twenty-first century in other disciplines as well as mathematical fluid dynamics.

In the remainder of this paper, I will give several specific examples of this interdisciplinary interaction. The first section is an introduction to the elementary structure of fluid flow. I discuss the physical phenomena and mathematical theory for nonlinear sound waves and vorticity/turbulence respectively in §§2 and 3. A discussion of several interesting directions for contemporary and future research is included in both sections. Finally in §4, I discuss two specific and current examples of the interaction of large scale computing and modern mathematical theory in describing strikingly new phenomena in fluid flows.

## 1. The elementary mathematical structure of fluid flows.

(A) *The equations for compressible flow.* The equations of compressible fluid flow in  $N$  space dimensions are a system of  $N + 1$  equations for  $N + 1$  unknowns given by

$$(1.1) \quad \begin{aligned} \frac{\partial \rho}{\partial t} + \operatorname{div}(\vec{m}) &= 0, \\ \frac{\partial \vec{m}}{\partial t} + \operatorname{div} \left( \frac{\vec{m} \otimes \vec{m}}{\rho} \right) + M^{-2} \nabla p(\rho) \\ &= \nu \operatorname{div} \left( \nabla \left( \frac{\vec{m}}{\rho} \right) + {}^t \nabla \left( \frac{\vec{m}}{\rho} \right) - \frac{2}{3} I \operatorname{div} \left( \frac{\vec{m}}{\rho} \right) \right). \end{aligned}$$

In (1.1),  $\rho$  is the fluid density while  $\vec{m} = \rho \vec{v}$  is the momentum with  $\vec{v} = {}^t(v_1, \dots, v_N)$  the fluid velocity. The first equation in (1.1) expresses conservation of mass while the remaining  $N$  equations express conservation of momentum, i.e., a continuous analogue of Newton's law,  $\vec{F} = m\vec{a}$ . There are two forces exerted by a fluid on itself. The first is given by the pressure gradient, the term  $\nabla p(\rho)$  on the left-hand side of (1.1) with  $p(\rho)$ , the pressure, a given nonlinear function of density determined from thermodynamics. For ideal fluids such as air or water,  $p = A\rho^\gamma$  with constants  $A$ ,  $\gamma$  satisfying  $A > 0$  and  $\gamma > 1$ . The second force exerted by the fluid is the frictional force of relative motion of fluid molecules represented by the term on the right-hand side of (1.1)— $\nu$  is the coefficient of viscosity. The nondimensional number  $M$  is the Mach number and is a constant which we explain below. The size of  $M$  has enormous significance for the phenomena

observed in various regimes of fluid motion; however, we postpone a detailed discussion until §§3 and 4 of this paper which illustrate this point in detail. The  $N \times N$  matrix  $\vec{w} \otimes \vec{w}$  is the tensor product,  $(\vec{w} \otimes \vec{w})_{ij} = w_i w_j$ , while  $\nabla \vec{w}$  is the  $N \times N$  matrix with rows given by  $\nabla w_i$  and the divergence operator applied to a matrix acts on each row. This completes the definition of the equations for fluid flow. I have intentionally begun this section with this complex set of nonlinear equations. In fact, the reader with a more sophisticated knowledge of fluid dynamics recognizes that despite the complexity of these equations, I have ignored energy effects and entropy changes so that the equations in (1.1) are the system describing isentropic compressible flows [2, 3]. In the remainder of this section, I will use elementary mathematical ideas to display the structure of solutions of these equations which confirms some of our physical intuition. This provides a first illustration of the fashion in which mathematical ideas contribute to understanding fluid flow and also provides some necessary background for the remainder of this paper.

Our intuition suggests that the frictional effect measured by the coefficient of viscosity is extremely small in air or rapidly moving water. Experiments confirm that the coefficient of viscosity in nondimensional terms often has the size  $\nu = O(10^{-4})$  to  $\nu = O(10^{-8})$ . These are the regimes of fluid flow we discuss in this paper. Since the terms on the right-hand side of (1.1) are complicated but multiplied by a small factor  $\nu$ , formally we make the *zero-viscosity approximation* and set  $\nu = 0$  to arrive at the *Compressible Euler Equations*:

$$(1.2) \quad \begin{aligned} \frac{\partial \rho}{\partial t} + \operatorname{div}(\vec{m}) &= 0, \\ \frac{\partial \vec{m}}{\partial t} + \operatorname{div} \left( \frac{\vec{m} \otimes \vec{m}}{\rho} \right) + M^{-2} \nabla p(\rho) &= 0. \end{aligned}$$

How good are such zero-viscosity approximations? This is an important current topic of mathematical research in fluid mechanics which we discuss in §§2 and 3 below.

Besides their obvious importance in describing fluid flow, the equations in (1.1), (1.2) are the primary example of a general structure useful for application to many other systems of equations describing a wide variety of complex physical phenomena. The equations in (1.1) and (1.2) are prototypical examples of  $m \times m$  systems of conservation laws in  $N$  space variables with the form,

$$(1.3) \quad \frac{\partial \vec{u}}{\partial t} + \sum_{j=1}^N \frac{\partial}{\partial x_j} F_j(\vec{u}) = \nu \sum_{j=1}^N \frac{\partial}{\partial x_j} (D_j(\nabla \vec{u})).$$

Here  $\nu > 0$  is a diffusion coefficient and typically satisfies  $\nu \ll 1$  in applications. Invoking the zero-diffusion approximation by setting  $\nu = 0$ , one

obtains the system

$$(1.4) \quad \frac{\partial \vec{u}}{\partial t} + \sum_{j=1}^N \frac{\partial}{\partial x_j} F_j(\vec{u}) = 0.$$

Here for  $1 \leq j \leq N$ ,  $F_j(\vec{u})$  and  $D_j(\nabla \vec{u})$  are given nonlinear mappings from  $R^m$  and  $R^{m^2}$  respectively to  $R^m$ . A partial list of other physical problems besides fluid flow modeled by such systems of equations includes nonlinear elasticity, magneto-fluid dynamics, combustion, super fluids, secondary oil recovery and the equations for meteorology among others. Thus, a systematic mathematical approach to understanding properties of solutions of the equations of fluid flow as a special case of the more general systems in (1.3), (1.4) contributes to increased potential understanding for a wide variety of diverse physical phenomena. Such an approach to (1.3), (1.4) emphasizing the common mathematical structure of the important physical examples has been pioneered by Friedrichs and P. Lax [4, 5]. I illustrate this general approach next and apply it to the equations of fluid flow.

(B) *Nonlinear sound waves and vorticity waves.* I concentrate on the inviscid equations in (1.2) and (1.4) for the remainder of this section. Any constant vector  $u_0 \in R^m$  is a special solution of (1.4). Consider perturbed smooth solutions with the form  $u^\varepsilon = u_0 + \varepsilon u' + O(\varepsilon^2)$ . Substituting this form into (1.4) and collecting the leading power in  $\varepsilon$  has the consequence that  $u'(x, t)$  satisfies the linearized equation,

$$(1.5) \quad u'_t + \sum_{j=1}^N A_j(u_0) u'_{x_j} = 0, \quad x \in R^N, \quad t > 0, \\ u'(x, 0) = u_0(x),$$

where  $A_j(u) = \partial F_j / \partial u$  are the  $m \times m$  Jacobian matrices of  $F_j$  and  $x = (x_1, \dots, x_N)$ . A minimum requirement on the system in (1.4) is that the problem in (1.5) is well-posed, i.e., solutions exist for reasonable initial data and small changes in the initial data lead to small changes in the solution. This condition is satisfied provided that for every unit direction  $\vec{n} = (n_1, \dots, n_N)$ , the  $m \times m$  matrix  $A(\vec{n}) = \sum_{j=1}^N A_j n_j$  has  $m$  real eigenvalues and complete eigenvectors; thus there are  $m$  real numbers  $\{\lambda_j\}_{j=1}^m$  and corresponding nonzero vectors  $\{r_j\}_{j=1}^m$  with

$$(1.6) \quad A(\vec{n}) r_j = \lambda_j r_j, \quad 1 \leq j \leq m.$$

In general I assume that  $r_j$  is a smooth function of  $\vec{n}$ . An immediate consequence of (1.6) is that the linearized equation in (1.5) has exact plane wave solutions,

$$(1.7) \quad u' = \sigma(x \cdot \vec{n} - \lambda_j t) r_j$$

for any  $j$  with  $1 \leq j \leq m$  with  $\sigma(s)$  an arbitrary smooth function. For this reason, the eigenvalues  $\{\lambda_j\}_{j=1}^m$  are called the wave speeds.

For the equations of fluid flow in (1.2) it is extremely easy to guess the nature of the wave speeds. During a lecture in a relatively still room, the fluid remains still but the *lecturer is heard* through the *propagation of sound waves* in the fluid. On the other hand, a windmill generates swirling of the air without producing much noise; this generation of *swirling* is the *propagation of vorticity waves* in the fluid. Thus, physical intuition leads one to expect that for fluid flow the wave speeds in (1.6) should represent the propagation of sound waves and vorticity waves. This intuition is confirmed through the following fact:

**PROPOSITION.** *For a given constant state  $(\rho_0, \vec{v}_0)$ ,  $\vec{v}_0 \in R^N$ , there are always  $N + 1$  real wave speeds at  $(\rho_0, \vec{v}_0)$  for the system in (1.2) provided that the nonlinear function  $p(\rho)$  satisfies  $\frac{dp}{d\rho} > 0$ . Two of the wave speeds,  $\lambda_{\pm}$ , associated with sound waves, are given by*

$$(1.8A) \qquad \lambda_{\pm} = v_0 \cdot \vec{n} \pm M^{-1} C_0,$$

where  $C_0 = (\frac{dp}{d\rho}(\rho_0))^{1/2}$ . The other  $N - 1$  wave speeds, associated with fluid swirling, are the vorticity waves which move with the fluid and have the speed

$$(1.8B) \qquad \lambda_0 = v_0 \cdot \vec{n}.$$

Next I explain the meaning of the constant  $M$ , the Mach number. This number represents the ratio of a typical representative fluid velocity magnitude,  $V$ , to the typical representative sound speed,  $C$ , for the class of fluid flows under consideration, i.e.,

$$(1.9) \qquad M = V/C.$$

The Mach number provides a nondimensional reference for the strength of the fluid velocity versus the speed of sound. The air in the wake of an auto moving at 20 m.p.h. satisfies  $M \ll 1$  while for a commercial flight on a jumbo jet,  $M$  satisfies  $M \approx 1$ .

Next, I generalize the exact solutions in (1.7) for the linear problem in (1.5) to the nonlinear problem in (1.4) (see [3] for the details). I replace the straight line  $\sigma r_k$  in (1.7) by the curve  $\sigma \mapsto U(\sigma)$  mapping  $R^1$  to  $R^m$ . I also replace the linear argument of  $\sigma$  with the more general form  $\sigma(x \cdot \vec{n}, t)$  and seek solutions of (1.4) with the special form

$$(1.10) \qquad u = U(\sigma(x \cdot \vec{n}, t)).$$

Such solutions are called nonlinear simple waves. The reader readily verifies that  $U(\sigma(x \cdot \vec{n}, t))$  is a solution of (1.4) provided that

$$(1.11A) \qquad \text{the function } U(\sigma) \text{ satisfies the nonlinear O.D.E. } U_{\sigma} = r_k(U(\sigma)) \text{ for some } k;$$



$$(1.11B) \quad \text{with } \lambda_k(\sigma) = \lambda_k(U(\sigma)), \sigma(y, t) \text{ satisfies the scalar nonlinear equation } \sigma_t + \lambda_k(\sigma)\sigma_y = 0.$$

In most physical examples including the equations of fluid flow, the nonlinear O.D.E. in (1.11A) can be integrated explicitly.

THE NONLINEAR BEHAVIOR OF SOUND WAVES. The following Proposition summarizes the nature of these exact solutions for the nonlinear sound waves in fluid flow.

PROPOSITION. Assume that the nonlinear pressure function  $p(\rho)$  satisfies the convexity condition  $p_{\tau\tau} > 0$  with  $\tau = \frac{1}{\rho}$ . Then there are exact solutions of the fluid equations in (1.2) given by the recipe in (1.10) and (1.11), where  $\sigma(y, t)$  satisfies the much simpler scalar nonlinear equation

$$(1.12) \quad \sigma_t + (\tfrac{1}{2}\sigma^2)_y = 0, \quad \sigma(y, 0) = \sigma_0(y).$$

Thus, the simpler equation in (1.12) provides a simplified quantitative model for the propagation of special smooth solutions of (1.2) associated with nonlinear sound waves. The equation in (1.12) is the celebrated inviscid Burgers equation. Next I summarize the well-known properties of both smooth and discontinuous solutions of the inviscid Burgers equation (see [3, 6]) and contrast these features with those of linear equations. Since the form in (1.10) yields a solution of (1.2) only as long as  $\sigma$  remains smooth, I rely on the discussion of discontinuous solutions of (1.12) to provide qualitative insight into the nonlinear propagation of sound waves in a fluid—this is a simple example of qualitative-quantitative modelling of complex physical phenomena by much simpler equations.

The linear equation  $\sigma_t + c\sigma_y = 0$ ,  $\sigma(y, 0) = \sigma_0(y)$  has the explicit solution  $\sigma_0(y - ct)$  and the following properties:

- (A) If  $\sigma_0(y) \in C_0^\infty$ ,  $\sigma(y, t) \in (C^\infty)$  for all  $t$ .
- (B) The initial data  $\sigma_0$  is discontinuous at the point  $y_0$  if and only if  $\sigma(y, t)$  is discontinuous at the points  $y = y_0 + ct$  along the characteristic curve emanating from  $y$ .

In contrast, the solution of the nonlinear equation  $\sigma_t + (\tfrac{1}{2}\sigma^2)_y = 0$ ,  $\sigma(y, 0) = \sigma_0(y)$  has the following properties:

$$(1.13A) \quad \text{For every } \sigma_0(y) \in C_0^\infty, \text{ there is a critical time } T_* > 0 \text{ so that } |\sigma'(y, t)| \nearrow \infty \text{ as } t \uparrow T_0; \text{ thus, solutions never stay smooth for all time.}$$

For discontinuous initial data with the form

$$(1.13B) \quad \sigma_0(y) = \begin{cases} \sigma_L, & y < 0, \\ \sigma_R, & y \geq 0: \end{cases}$$

CASE I. If  $\sigma_L > \sigma_R$ ,  $\sigma(y, t)$  is given by

$$\sigma(y, t) = \begin{cases} \sigma_L, & y < \frac{\sigma_L + \sigma_R}{2}t, \\ \sigma_R, & y \geq \frac{\sigma_L + \sigma_R}{2}t. \end{cases}$$

Thus these solutions have a propagating discontinuity which does *not* move at characteristic wave speed.

CASE II. If  $\sigma_L < \sigma_R$ ,  $\sigma(y, t)$  is given by

$$\sigma(y, t) = \begin{cases} \sigma_L, & y < \sigma_L t, \\ y/t, & \sigma_L t \leq y \leq \sigma_R t, \\ \sigma_R, & y > \sigma_R t. \end{cases}$$

Thus, a discontinuity in the initial data disappears for all positive times.

The phenomenon in (1.13A) corresponds to the formation of *shock waves* in a fluid through the nonlinear propagation of sound waves (see Chapter 3 of [3]). The results in Case I of (1.13B) indicate that shock waves do not travel at the speed of sound; this is well known experimentally and mathematically (see [2]). Finally, the smooth solutions with discontinuous initial data given in Case II of (1.13B) are called centered *rarefaction waves* in the general theory because such solutions were discovered first for the fluid equations by Riemann and correspond physically to rarefaction of the gas, i.e., the density of the gas drops by following the gas particles through such a wave. None of the above features occur in linear equations as illustrated by the simple example described above.

NONLINEAR VORTICITY WAVES. Next I apply the construction in (1.10), (1.11) to the wave speeds  $\lambda_0 = v_0 \cdot n$  associated with vorticity. For simplicity I consider two space dimensions and  $\vec{n} = (1, 0)$ . The construction in (1.10) yields exact solutions of the fluid equations with the form

$$(1.14) \quad \rho = \rho_0, \quad \vec{v} = {}^t(0, v_2(x_1)).$$

These nonlinear vorticity waves are shear layers; the velocity points along the  $x_2$ -axis but remains a function of the perpendicular variable,  $x_1$ . Such solutions involve only motion of the fluid and no propagation of sound. Furthermore these exact solutions are *incompressible* because the velocity in (1.14) satisfies  $\text{div } \vec{v} = 0$ . These solutions have *nonzero vorticity* which is defined by  $\vec{w} = \text{curl } \vec{v}$  because

$$\text{curl } \vec{v} = -\frac{\partial v_2}{\partial x_1} \neq 0.$$

The propagation and amplification of vorticity produce the dramatic effects of turbulence (see §3).

**2. Physical phenomena and mathematical theory for nonlinear sound waves: Recent progress and future directions.** Here I describe the current developments in four interesting research areas involving the behavior of nonlinear

sound waves. The first two areas discussed involve rigorous mathematical analysis, while the third involves a combination of rigorous analysis, quantitative asymptotic modelling, and small scale numerical computation. In the fourth topic I describe some of the new insights into the multi-dimensional interaction of nonlinear sound waves which have been produced through recent ingenious large scale numerical simulation. I end this section with some brief speculation on an interdisciplinary attack mingling the four tools in (0.1) which might explain the observed phenomena in the last topic.

(A) *The zero diffusion approximation: the mathematical theory of conservation laws in a single space variable.* Before discussing the specific nature of results in this problem for the equations of fluid flow, I define the basic mathematical issues for a general system in a single space variable (where the mathematical issues are already very difficult!). Thus, I consider solutions  $u^\nu$  of the  $m \times m$  system of equations

$$(2.1A) \quad u_t^\nu + F(u^\nu)_x = \nu(D(u^\nu))_x, \quad t > 0, \quad x \in R^1,$$

with given initial data

$$(2.1B) \quad u^\nu(x, 0) = u_0(x),$$

and the corresponding solution of the formal zero diffusion limit with the same initial data

$$(2.2) \quad \begin{aligned} u_t + F(u)_x &= 0, \quad t > 0, \quad x \in R^1, \\ u(x, 0) &= u_0(x). \end{aligned}$$

The structure in (1.6) is assumed so that (2.2) is a nonlinear hyperbolic equation. As explained above (1.2) for the equations of compressible flow, in most problems of physical interest, the diffusion coefficient satisfies  $\nu \ll 1$ . From the physical point of view, one would like to avoid a detailed assessment of the microscopic effects of diffusion since this is very difficult to achieve both theoretically and experimentally. Thus, the basic *physical issue motivating the mathematical problem of the zero diffusion limit* is the following:

(2.3) Given reasonable initial data  $u_0(x)$ , find a suitable solution  $u(x, t)$  of the system of nonlinear hyperbolic equations in (2.2) so that  $u(x, t)$  automatically includes the macroscopic effects of diffusion in  $u^\nu(x, t)$ , i.e., so that  $u^\nu(x, t)$  converges to  $u(x, t)$  as  $\nu \rightarrow 0$  in a suitable sense.

For smooth bounded initial data  $u_0(x)$ , it is not difficult to prove that  $u^\nu(x, t) \rightarrow u(x, t)$  as  $\nu \rightarrow 0$  for any time interval  $[0, T_*]$  such that the solution  $u(x, t)$  stays smooth on  $[0, T_*]$ . S. Klainerman gave an extremely elegant but unpublished proof of this fact several years ago under natural structural assumptions on the diffusion matrix  $D$  in (2.1) (private communication) but the author does not know a detailed published reference. However, subtlety in this problem arises because solutions of (2.2) with smooth

initial data typically become discontinuous in finite time (see (1.13A)), [6], and Chapter 3 of [3]); thus, one is naturally forced to deal with *bounded discontinuous solutions which satisfy (2.2) in the sense of distributions*, i.e.,

$$(2.4) \quad \iint (\phi_t \cdot u + \phi_x F(u)) dx dt = 0$$

for all smooth vector-valued test functions  $\phi = (\phi_1, \dots, \phi_m)$  with  $\phi_j \in C_0^\infty(R^1 \times (0, \infty))$ . With this difficulty, one might as well consider bounded discontinuous initial data,  $u_0$ , at the outset. The simplest initial data of this sort for the system in (2.2) are *Riemann* initial data, i.e.,

$$(2.5) \quad u_0(x) = \begin{cases} u_L, & x < 0, \\ u_R, & x > 0, \end{cases}$$

where  $u_L, u_R \in R^m$  are two prescribed constants; these data are named after Riemann because he was the first to study this problem for the equations of fluid flow. For example, suppose  $u_L, u_R$  are two constant vectors which satisfy the nonlinear algebraic equation (the Rankine-Hugoniot conditions)

$$(2.6) \quad -s(u_L - u_R) + F(u_L) - F(u_R) = 0.$$

Then a weak solution of (2.2) (in the sense of (2.4)) with the initial data in (2.5) is given by

$$(2.7) \quad u(x, t) = \begin{cases} u_L, & x \leq st, \\ u_R, & x > st, \end{cases}$$

with the nonlinear wave speed  $s$  determined from (2.6). I return to this special problem in the next paragraph. In pioneering work, Lax [4] showed how to get solutions of the Riemann problem for general constants  $u_L, u_R$  provided  $|u_L - u_R|$  was suitably small; furthermore, this solution was unique in a suitable class of centered-wave solutions provided that certain geometric shock inequalities introduced by Lax were satisfied (pp. 133–137 of [3] contain a detailed rigorous discussion for the motivation behind Lax's shock inequalities). In a brilliant paper which strongly influenced developments in the field for many years, Glimm [7] used Lax's solutions of the Riemann problem as building blocks and established the existence for all  $t > 0$  of discontinuous solutions of (2.2) provided that the initial data  $u_0$  had both small amplitude and small total variation (see the book by Smoller [8] for the details of this work and other related developments). The additional geometric structure and large time behavior of solutions of (2.2) constructed through Glimm's method has been elucidated by important further work of Glimm-Lax [9], DiPerna, Liu, and Dafermos among others (see the references in [10, 11]; the article by Dafermos is especially recommended by the author as an introduction to the topics discussed in this subsection). Glimm's theoretical method of existence has also led to interesting new numerical methods (see Chorin [12]). This discussion fills in some background on existence of discontinuous solutions of (2.2).

I return to the special discontinuous initial data satisfying (2.6) and ask whether the weak solution in (2.7) arises as the zero diffusion limit of  $u^\nu$  as  $\nu \rightarrow 0$ . For appropriate initial data for (2.1) with  $\nu$ -dependence, this reduces to an interesting question for nonlinear O.D.E.'s which I develop next. It is natural to attempt to build solutions of (2.1) in the form of travelling waves

$$(2.8A) \quad u^\nu(x, t) = U\left(\frac{x - st}{\nu}\right)$$

with

$$(2.8B) \quad \lim_{\xi \rightarrow -\infty} U(\xi) = u_L, \quad \lim_{\xi \rightarrow +\infty} U(\xi) = u_R.$$

Clearly, when such a solution exists,  $u^\nu(x, t)$  converges to the solution  $u(x, t)$  in (2.7) as  $\nu \rightarrow 0$ . The existence of solutions of this sort which solve the zero diffusion limit in (2.3) in a special sense for these special initial data is called the problem of diffusive *shock layers*. This is an extremely important prototype problem because at points of discontinuity, general weak solutions of (2.2) locally look like the solution in (2.7) (see [10]). Of course, the advantage in studying (2.8) is that  $U(\xi)$  satisfies the simpler  $m \times m$  system of nonlinear O.D.E.'s

$$(2.9A) \quad D(U'(\xi)) = F(U(\xi)) - sU(\xi) + C_0$$

with the constant  $C_0$  given by

$$(2.9B) \quad C_0 = -(F(u_L) - su_L) = -(F(u_R) - su_R).$$

Both  $u_L$  and  $u_R$  are critical points of the O.D.E. in (2.9) and the existence of shock layers amounts to proving that there is an orbit defined by  $U(\xi)$  connecting the critical point  $u_L$  with the critical point  $u_R$ . As I mentioned earlier in the discussion below (1.1), the full equations of fluid flow include changes in total energy besides the equations for mass and momentum (see [2, 3]). As in (1.1) there is no diffusion of mass but diffusion of momentum through viscosity and dissipation of energy through both the effect of viscosity and heat conduction. Beginning around 1910, there has been an enormous interest among physicists and engineers in finding solutions of the shock layer equations in (2.9) for fluid flow and various special solutions were discovered for special ratios of viscosity and heat conduction. Mathematically, shock layers for the full equations of fluid flow reduce to questions regarding  $2 \times 2$  autonomous systems of O.D.E.s which are amenable to phase portrait techniques. H. Weyl [13] contributed to this problem and Gilbarg [14] gave the complete solution for generalized ideal gases; in very recent work Pego [15] has generalized Gilbarg's results for general equations of state and has provided counter-examples for nonconvex equations of state where *shock layers do not exist if heat conduction dominates viscosity*. Similar results are extremely easy to obtain for the isentropic gas dynamic equations in (1.1),

(1.2) so I sketch the construction below. The equations in (1.1) in a single space dimension are given by

$$(2.10) \quad \begin{aligned} \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho v) &= 0, \\ \frac{\partial \rho v}{\partial t} + \frac{\partial}{\partial x}(\rho v^2 + p(\rho)) &= \nu \frac{\partial^2 v}{\partial x^2}. \end{aligned}$$

The jump conditions in (2.6) in this special case are given by

$$(2.11) \quad \rho_L(v_L - s) = \rho_R(v_R - s) \equiv m, \quad m^2 = -\frac{p(\rho_L) - p(\rho_R)}{\tau_L - \tau_R},$$

where  $\tau = \frac{1}{\rho}$  is the specific volume and  $m$  is a constant, the mass flux. Because there is no diffusion of mass in (2.10), the shock layer solution in (2.9),  $u(\xi) = {}^t(\rho(\xi), v(\xi))$ , is determined by the solution of the scalar nonlinear O.D.E. for  $\tau(\xi)$ ,

$$(2.12) \quad \tau'(\xi) = -\frac{1}{m}[m^2(\tau - \tau_L) + p(\tau) - p(\tau_L)].$$

If the mass flux  $m$  satisfies  $m > 0$ , then the fluid velocity satisfies  $v_L > s$  and  $v_R > s$  so that the fluid particles which move at speed  $v$  cross the wave front from the left to the right; similar remarks apply for  $m < 0$ . By direct quadrature of (2.12), it is easy to prove the following

**PROPOSITION.** *Consider (isentropic) fluid flow with an equation of state,  $p(\tau)$ , satisfying the convexity condition  $p_{\tau\tau} > 0$  with  $\tau = \frac{1}{\rho}$ . Consider a weak solution of (1.2) defined through (2.6), (2.7), and (2.11). If  $m > 0$  ( $< 0$ ) a shock layer exists if and only if  $\rho_R > \rho_L$  ( $\rho_L > \rho_R$ ). Thus, shock layers exist at a speed  $s$  if and only if the fluid is compressed (the density increases) as fluid particles cross the front.*

The implication of this Proposition for weak solutions of (2.2) is very interesting; there are many weak solutions of (2.2) which are not the zero diffusion limit of solutions of (2.1). For fluid flow, general solutions of (1.2) which satisfy the criterion of being compressive at their discontinuities for convex equations of state are called *entropy solutions* because in the complete equations of fluid flow including entropy, they correspond to weak solutions with increases in entropy in agreement with the second law of thermodynamics. For general equations in (2.1), necessary and sufficient algebraic conditions to guarantee a shock layer in the case that  $|u_L - u_R|$  is small have been developed recently by Pego and the author [16]; the existence of shock layers when  $|u_L - u_R|$  is not small often requires topological methods (see [8]).

The existence of shock layers provides intuition but does not even solve the problem of the zero diffusion limit in (2.1)–(2.3) for the special solution in (2.7) satisfying the entropy condition because the initial data defined by (2.8A) changes with  $\nu$ . To resolve this difficulty, one needs to assess the

nonlinear stability of the shock layer. In recent important work, T. P. Liu has succeeded in doing this both for the equations of fluid flow [17] and in general [18]. Liu has discovered a number of remarkable new diffusion waves in his rigorous analysis of the limit. Liu's work builds on earlier results of Goodman [19] utilizing energy methods under additional assumptions which guaranteed that such diffusion waves are suppressed. Very recently, Liu and Hoff [20] have completed the analysis of the quantitative behavior of the zero diffusion limit for discontinuous shock data while Liu and collaborators (personal communication) have rigorously analyzed the zero diffusion limit for piecewise constant rarefaction data in (2.5) where completely different phenomena occur (see (1.13B) for rarefaction data in a simpler problem). It is quite likely that Liu and coworkers will understand the detailed nature of the zero diffusion approximation for general Riemann data in (2.5) in a rigorous constructive fashion in the next few years.

A completely different attack on the question of the zero diffusion approximation is to take the smooth solutions of (2.1), obtain suitable estimates *independent of  $\nu$* , and pass to the limit to *construct solutions* of (2.2) automatically satisfying (2.3). In 1951, E. Hopf and Cole independently succeeded in doing this for the model scalar equation

$$u_t^\nu + \left(\frac{1}{2}(u^\nu)^2\right)_x = \nu u_{xx}^\nu$$

through explicit solution; a general approach for scalar laws utilizing the maximum principle and total variation estimates was developed by Lax, Oleinik, and Kruskov among others (see the bibliography in [6, 8, 11]). However, for  $m \times m$  systems with  $m > 1$ , it is very difficult to obtain similar a priori estimates which are uniform in viscosity parameters. In 1978, L. Tartar [21] introduced a bold new program involving weak convergence and a tool developed by Tartar and Murat [22], compensated compactness, which relied on only the obvious uniform bound estimates for  $u^\nu$  without the harder a priori estimates for derivatives; furthermore, Tartar [21] succeeded in giving a new proof of convergence for scalar laws by these methods utilizing the weak topology. Through an extremely novel use and generalization of Lax's entropy pairs, DiPerna [23] succeeded in extending Tartar's program to  $2 \times 2$  systems; in this fashion, DiPerna [24] obtained a new existence theorem for the isentropic gas equations in (2.2) in a single space dimension with general large initial data by taking the zero diffusion limit directly.

I end this subsection with two open problems for the mathematical theory of fluid flow in a single space variable. Both problems involve the full  $3 \times 3$  system of equations involving conservation of mass, momentum, and total energy.

**PROBLEM 1.** Can the ideas of Tartar and DiPerna be combined with other estimates to study the zero diffusion limit for the  $3 \times 3$  system of equations for fluid flow?

Another important problem which I have not mentioned is the uniqueness

of weak solutions of (2.2) which satisfy the *entropy* condition (see the above Proposition and [6, 11]). In an interesting paper, DiPerna [25] has proved the uniqueness of solutions of the Riemann problem in (2.5) within all solutions satisfying the entropy condition provided  $m = 2$ . DiPerna's argument makes essential use of the condition  $m = 2$ . A solution of the following problem would be very interesting and seems accessible.

**PROBLEM 2.** For the  $3 \times 3$  system of conservation laws given by the equations of compressible fluid flow, prove the uniqueness of solutions of the Riemann problem in the class of bounded functions with bounded variation which are weak solutions satisfying the entropy condition of P. Lax (in [5, 25]) for this system. Find conditions on the equation of state and size of the initial data which guarantee this uniqueness theorem.

(B) *Structure and stability of wave patterns in several space variables.* We regard the fluid equations in (1.2) as the prototypical example for general  $m \times m$  systems of conservation laws in  $N$  space variables with the form

$$(2.13A) \quad u_t + \sum_{j=1}^N (F_j(u))_{x_j} = 0$$

and initial data

$$(2.13B) \quad u(x, 0) = u_0(x).$$

One striking difference between hyperbolic systems in one and several space variables regards the domain of dependence: the solution at a given point of the linearized equations in (1.5) with  $N = 1$  is determined by the initial data at a finite number of points; in several space variables, this domain of dependence is much larger and includes an entire cone of initial points. Thus, there are the possibilities for more complex nontrivial interaction of nonlinear sound waves coming from many different directions in multi-D. That such complexity occurs in fluid flows is a documented experimental fact. In particular solutions do not obey simple a priori maximum norm estimates because waves can focus; furthermore, no plentiful simple class of exact solutions such as Lax's solution of the Riemann problem in a single space variable have been found for the fluid equations in (1.2) in several space variables. Nevertheless, solutions of these equations are computed with rather high reliability through numerical methods (see the discussion in §2(D) and §4).

With all of the additional theoretical difficulties for systems in  $N$  space variables with  $N \geq 2$ , the mathematical theory of short-time existence and stability of solutions of (2.13) with special discontinuous initial data has been the main focus of research. The goal of this work is to elucidate the role of inherently multi-dimensional wave interactions in the stability of basic nonlinear wave patterns. The author studied the short-time existence and stability of shock fronts in multi-D (see [26, 27], and Chapter 4 of [3] for a leisurely introduction). Metevier [28] has studied the short-time interaction of two multi-D



shock fronts and recently Alinhac [29] has proved the short-time existence and stability of smooth rarefaction fronts (see (1.13B)) emanating from appropriate piecewise discontinuous initial data. In addition, Harabetian [30] has proved the short-time existence of multi-dimensional curved Riemann problems for piecewise analytic initial data through nonlinear Cauchy-Kowaleski theorems; however, these last results provide no insight into the correct geometric conditions which guarantee the multi-dimensional structural stability of these nonlinear wave patterns. This last comment will become more evident to the reader if he remembers that the Cauchy-Kowaleski theorem can be used to solve the initial value problem for the Laplace equation with real analytic initial data; nevertheless, this initial value problem is clearly ill-posed and cannot be solved for any other smooth initial data which is not real analytic. The techniques involved in all these theorems except the work in [30] require sophisticated ideas from the theory of hyperbolic mixed problems, pseudo-differential operators to obtain stability estimates, and complicated nonlinear iteration schemes. Several open problems for systems of conservation laws in several space variables are discussed at the end of Chapters 3 and 4 of [3].

Here I would like to discuss a new approach to the existence of shock fronts due to E. Thomann and the author [31] which provides a rigorous geometric picture of multi-dimensional shock propagation for an interesting class of systems of conservation laws which arise as second-order wave equations. The prototypical example of a second-order wave equation arises in fluid mechanics in the following fashion: For smooth solutions, the compressible fluid equations in (1.2) can be rewritten in the form

$$(2.14) \quad \begin{aligned} \rho_t + \operatorname{div}(\rho \vec{v}) &= 0, \\ \vec{v}_t + (\vec{v} \cdot \nabla) \vec{v} + \nabla h(\rho) &= 0, \end{aligned}$$

where  $h(\rho)$  is determined within a constant by the equation

$$(2.15) \quad h'(\rho) = \frac{p'(\rho)}{\rho} > 0$$

and  $v \cdot \nabla = \sum_{j=1}^N v_j \frac{\partial}{\partial x_j}$ . With the restriction to fluid flows so that there is a potential function  $\phi$  with  $\vec{v} = \nabla \phi$ , the last  $N$  equations can be integrated to obtain Bernoulli's law,

$$(2.16) \quad \phi_t + \frac{1}{2} |\nabla \phi|^2 + h(\rho) = 0.$$

From (2.15) and (2.16), it follows that the density is given by

$$(2.17) \quad \rho = h^{-1}(-(\phi + t + \frac{1}{2} |\nabla \phi|^2)).$$

By substituting (2.17) into the equation for conservation of mass in (2.14), a second-order wave equation for the potential  $\phi$  is obtained with the form

$$(2.18) \quad \frac{\partial}{\partial t} \mathcal{H}_0(d\phi) + \sum_{j=1}^N \frac{\partial}{\partial x_j} \mathcal{H}_j(d\phi) = 0.$$

Here  $d\phi = (\frac{\partial\phi}{\partial t}, \frac{\partial\phi}{\partial x_1}, \dots, \frac{\partial\phi}{\partial x_N})$  and for fluid flow the nonlinear functions  $\mathcal{H}_j$  are given by

$$(2.19) \quad \begin{aligned} \mathcal{H}_0 &= h^{-1}(-(\phi_t + \tfrac{1}{2}|\nabla\phi|^2)), \\ \mathcal{H}_j &= \frac{\partial\phi}{\partial x_j} \mathcal{H}_0, \quad 1 \leq j \leq N. \end{aligned}$$

The steady form of the equations in (2.18), (2.19) is often used in computational aerodynamics. Since  $\vec{v} = \nabla\phi$  satisfies  $\text{curl } \vec{v} = 0$ , the corresponding fluid flow is irrotational and no vorticity waves occur in these solutions (see §1); the only effects that remain in (2.18), (2.19) are from the nonlinear sound waves.

I sketch the simplified picture of multi-dimensional geometric shock propagation which is used as a theoretical tool in [31] to obtain a very simple proof for existence of shock fronts for second-order wave equations. This rigorous formulation of geometric shock propagation should have other interesting pure and applied consequences. The jump conditions for a weak solution of (2.18) with piecewise smooth gradients across a space-time hypersurface,  $S$ , are given by

$$(2.20) \quad \sum_{i=0}^N \nu_i [H_i]_S = 0, \quad [\phi]_S = 0,$$

where brackets,  $[ \ ]$ , denote the jump in a quantity across  $S$  and  $\vec{\nu} = (\nu_0, \nu_1, \dots, \nu_N)$  is the space-time normal to  $S$ . Through the special form of the jump conditions, a partial hodograph transformation is introduced on p. 789 of [31]. The location of the shock front  $S$  as time evolves is given locally by the graph

$$(2.21) \quad x_N = u(t, x_1, \dots, x_{N-1}).$$

With  $y = (y^0, y^1, \dots, y^N)$  and  $x' = (x^1, \dots, x^{N-1})$ , the function  $u(t, x')$  is determined as the restriction of a function  $u(y)$  to the point with  $y^N = 0$  and  $(y^0, y^1, \dots, y^{N-1}) = (t, x')$ . The function  $u(y)$  solves the auxiliary nonlinear Neumann problem given by

$$(2.22) \quad \begin{aligned} \text{(A)} \quad & \sum_{j=0}^N \frac{\partial}{\partial y_j} \tilde{H}_j(du) = 0, \quad \text{for } y^0 > 0 \text{ and } y^N > 0, \\ \text{(B)} \quad & G(du) = 0, \quad \text{for } y^N = 0, \\ \text{(C)} \quad & u = f_0, \quad \frac{\partial u}{\partial y^0} = f_1, \quad \text{for } y^0 = 0, \quad y^N > 0. \end{aligned}$$

The functions  $f_0$  and  $f_1$  are determined by the initial data and the explicit form of the equation in (2.22A) and the boundary conditions in (2.22B) is given on p. 790 of [31]. Thus, for second-order wave equations, the geometric location of a shock front is determined by the restriction to the boundary of

the solution of an explicit nonlinear Neumann problem for a second-order hyperbolic equation determined by the initial data. Approximate theories of geometric shock dynamics can be developed through suitable approximations to the mixed problem in (2.22). While complete hodograph transformations are ubiquitous in fluid dynamics (see [2]), the work in [31] seems to be the first application of partial hodograph transformations to hyperbolic equations for fluid flow.

(C) *Nonlinear geometric optics.* With the mathematical complexity of solutions of the equations of fluid flow already evident from the discussion above in §§2(A), (B), it is not surprising that applied mathematicians have developed formal asymptotic methods to build simplified quantitative asymptotic approximations. The most prominent method of this sort for linear hyperbolic equations is geometric optics; here I discuss geometric optics for the nonlinear system of conservation laws,

$$(2.23) \quad u_t + \sum_{j=\phi}^N (F_j(u))_{x_j} = 0,$$

with a special emphasis on solutions of the compressible fluid equations in (1.2). I summarize some results below and provide references for a more detailed discussion.

The main idea of geometric optics rests in the observation that at high frequencies, i.e., for short wave lengths, solutions of complicated systems can simplify enormously (see [32] for the linear case). Systematic developments of geometric optics for the nonlinear hyperbolic equations in (1.2) or (2.23) involve the basic assumption:

$$(2.24) \quad \text{The initial data are a small amplitude perturbation with amplitude } \varepsilon \text{ and high frequency } O(\varepsilon^{-1}) \text{ with } \varepsilon \ll 1.$$

I illustrate the expansion of geometric optics for (2.23) in the case of initial data that excite only a single mode of propagation. I consider small amplitude perturbations of a constant state  $u_0$  with the initial data given by

$$(2.25A) \quad u_0 + \varepsilon \sigma_0(x, \frac{\phi_0}{\varepsilon}) r_p(\nabla \phi_0),$$

where  $r_p(\vec{n})$  is the right eigenvector defined earlier in (1.6) and  $\nabla \phi_0 \neq 0$ . Of course  $\nabla \phi_0$  is not necessarily a unit vector and  $p$  is fixed with  $1 \leq p \leq m$ . It is assumed that  $\sigma_0(x, \theta)$  is a function with zero mean, i.e.,

$$(2.25B) \quad \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \sigma_0(x, s) ds = 0.$$

Functions  $\sigma_0(s)$  which satisfy these assumptions include sums of almost periodic functions and functions of compact support. With these assumptions, one builds formal asymptotic solutions of (2.23) with the form

$$(2.26) \quad u^\varepsilon = u_0 + \varepsilon \sigma(x, t, \frac{\phi}{\varepsilon}) r_p(\nabla \phi) + \varepsilon^2 u_2(x, t, \frac{\phi}{\varepsilon}).$$

In order to have a uniformly valid asymptotic expansion, the correction  $u_2(x, t, \theta)$  necessarily must grow sublinearly in  $\theta$ , i.e.,

$$(2.27) \quad \lim_{|\theta| \rightarrow \infty} |\theta|^{-1} |u_2(x, t, \theta)| = 0.$$

The ansatz in (2.26) is substituted into (2.23) and successive equations in powers of  $\varepsilon$  are developed. In order to solve these successive equations subject to the constraint in (2.27), much simpler equations than (2.23) emerge for the phase function,  $\phi(x, t)$ , and the amplitude function,  $\sigma(x, t, \theta)$ . In fact,  $\phi$  solves the familiar *eikonal equation of linear geometric optics* for the  $p$ th wave,

$$(2.28) \quad \phi_t + \lambda_p(\nabla \phi) = 0, \quad \phi|_{t=0} = \phi_0,$$

while the amplitude  $\sigma$  solves a nonlinear transport equation

$$(2.29) \quad D_p \sigma + b_p(\tfrac{1}{2}\sigma^2)_\theta = 0, \quad \sigma(x, t, \theta)|_{t=0} = \sigma_0(x, \theta).$$

The operator  $D_p$  is the linear transport operator of geometric optics given by differentiation along the bicharacteristic rays associated with  $\phi$ . As in linear geometric optics [32],  $D_p$  has the form

$$(2.30) \quad D_p = \frac{\partial}{\partial t} + \vec{a}_p \cdot \nabla + c_p$$

and  $\vec{a}_p(x, t)$ ,  $b_p(x, t)$ ,  $c_p(x, t)$  are determined from  $\phi$  by explicit formulas. Thus, geometric optics for a single wave yields enormous simplification in solutions of (2.23). If the coefficient  $b_p$  vanishes identically,  $b_p \equiv 0$ , as occurs for the vorticity waves in fluid flow, then the equations in (2.28), (2.29) reduce to ordinary geometric optics for linear equations [32]. However, if  $b_p \neq 0$ , as occurs for the nonlinear sound waves in fluid flow, we only need to solve the simpler scalar nonlinear equation in (2.29) to construct asymptotic solutions of (2.23). I remark that under the restriction  $b_p \neq 0$ , the equation in (2.29) can be solved explicitly through solution of the much simpler inviscid Burgers equation,

$$\sigma_\tau + (\tfrac{1}{2}\sigma^2)_\theta = 0,$$

which was discussed earlier in (1.12), (1.13) so a complete explicit solution of (2.29) is available. The reduction of (2.29) to (2.30) requires three coordinate changes: introduction of bicharacteristic coordinates for  $\nabla_p$ , rescaling  $\sigma$  by a time dependent factor, and finally renormalizing time suitably (see [33, 34]). Such single wave expansions for nonlinear geometric optics were developed in the 1940s and 1950s by Lighthill [35] and Whitham [36] but a systematic treatment was first presented much later by Choquet-Bruhat [37] and recently generalized by Hunter and J. B. Keller [38]. A leisurely discussion of the derivation of (2.28), (2.29) as well as the remaining topics to be discussed in this subsection is contained in a paper of the author [39].

In geometric optics for linear hyperbolic equations, wave patterns superimpose and geometric optics approximations are given by

$$\sum_{p=1}^m \sigma_p(x, t, \frac{\phi_p}{\varepsilon}) r_p(\nabla \phi_p),$$

where each mode satisfies the eikonal and transport equations separately. Is this true in the nonlinear case? Do new phenomena appear? Recent research of Hunter, Rosales, and the author [40, 41, 42] employing a systematic development of nonlinear geometric optics reveals much more complex phenomena; general almost periodic wave trains do not superimpose but instead interact resonantly. In this situation, the nonlinear eikonal equations in (2.28) remain the same but the different amplitudes  $\{\sigma_p\}_{p=1}^m$  exchange energy through resonant wave interaction. A general theory for this resonant interaction in several space variables is developed in [41] and an application to the equations of compressible fluid flow in (1.2) is presented in §6 of [41]. In particular, the resonant nonlinear interaction of small amplitude sound waves with vorticity waves produces additional sound waves which resonantly interact. For solutions of the eikonal equations in (2.28) defined by plane waves, after some elementary changes of variable, the amplitudes of the two sound waves  $\sigma^\pm(\theta, t)$  satisfy the coupled system of resonant equations,

$$(2.31) \quad \begin{aligned} \sigma_t^+ + \frac{1}{2}(\sigma^+)_\theta^2 + \int_0^1 k(\theta - y) \sigma^-(y, t) dy &= 0, \\ \sigma_t^- + \frac{1}{2}(\sigma^-)_\theta^2 - \int_0^1 k(-\theta - y) \sigma^+(y, t) dy &= 0, \end{aligned}$$

where  $\sigma^+$ ,  $\sigma^-$ , and  $k$  are periodic in  $\theta$  with period one. The kernel  $k$  is a constant multiple of the vorticity of an initial high frequency nonlinear shear layer as described earlier in (1.14); as given there,  $k$  does not change to leading order in time. Clearly the two nonlinear sound waves do not superimpose as in (2.29) but instead resonantly interact through the vorticity wave which defines the convolution kernel in (2.31). Similar asymptotic equations are valid for the complete  $3 \times 3$  system of compressible fluid flow in a single space dimension (see [40]). The matrix convolution kernel in (2.31) is a skew symmetric operator; thus, it conserves the overall energy of the pair  $(\sigma^+, \sigma^-)$ . A recent paper which combines small scale numerical computation and several exact solutions reveals surprising new phenomena in the solutions of (2.31) through resonant wave interaction [43].

The rigorous theory of geometric optics for compressible flow and the general equations in (2.23) is only just beginning. Diperna and the author [71] gave a rigorous justification of geometric optics for appropriate systems including (1.2) in a single space variable for initial data of compact support; for these data of compact support resonances of the type described in (2.31) do not occur. The proofs in [71] indicate that there are more subtle nonlinear

mechanisms which guarantee that geometric optics is even better for nonlinear problems with discontinuous data than could be predicted by the formal theory! Several accessible open problems in the rigorous theory of geometric optics are mentioned in §4 of [39]. The most important and most difficult involves the rigorous justification of the equations in (2.31) describing resonant wave interaction. Another important problem is the rigorous justification of the single wave expansion in (2.25)–(2.30) in several space variables for initial data so that the solutions of (2.29) became discontinuous.

(D) *Nonlinear diffraction at caustics and boundaries.* I illustrate some of the insight which high quality large scale numerical simulations can contribute by discussing the diffraction of nonlinear sound waves. When a shock wave is incident along an inclined ramp, intuition based on linear theory leads one to guess a solution consisting of the incident wave pattern and a reflected curved wave much like that depicted in Figure 1(a). However, recent large scale numerical simulations by Glaz and Colella [44] reveal remarkably complex unexpected wave patterns in the diffraction of shocks by ramps. Besides the expected wave pattern in Figure 1(a), more complex reflected wave patterns such as those in Figure 1(b)–(d) occur for various angles and incident wave strengths. In Figure 1(b), the reflected shock is replaced by three shocks and a vorticity wave—this phenomena is called Mach reflection; in Figure 1(d) two different Mach stem structures of this type occur in the reflected wave. Colella and Glaz have found even more of these complex wave patterns embedded in the reflected wave under different flow conditions and for different equations of state. These complex patterns have practical significance because there are higher pressures associated with the triple shock points. Such kinds of complex reflection have been documented experimentally. Plates 235–238 of Van Dyke's book [1] describe a sequence of shadow graphs which display experimental phenomena corresponding to the four cases depicted in Figure 1.

Why bother with large scale numerical simulations when experiments show the same phenomena? There are several reasons. First, large scale simulations display all components of the solution such as density, velocity, pressure, vorticity, etc., simultaneously while an experimentalist works very hard to measure changes in a single physical quantity—density changes are measured in the plates 235–238 of [1]. A second reason is the ease with which a numerical analyst can change the properties of the material being studied—the differences for example between air, water, and a complex hydrocarbon typically only involve changing a few parameters in a numerical method; for an experimentalist, changing the material under investigation is a major undertaking often requiring a completely different experimental setup.

A basic theoretical issue is the following outstanding problem: why do such complex patterns form and what are the mechanisms generating these patterns? The advantages of numerical codes in studying different parameters

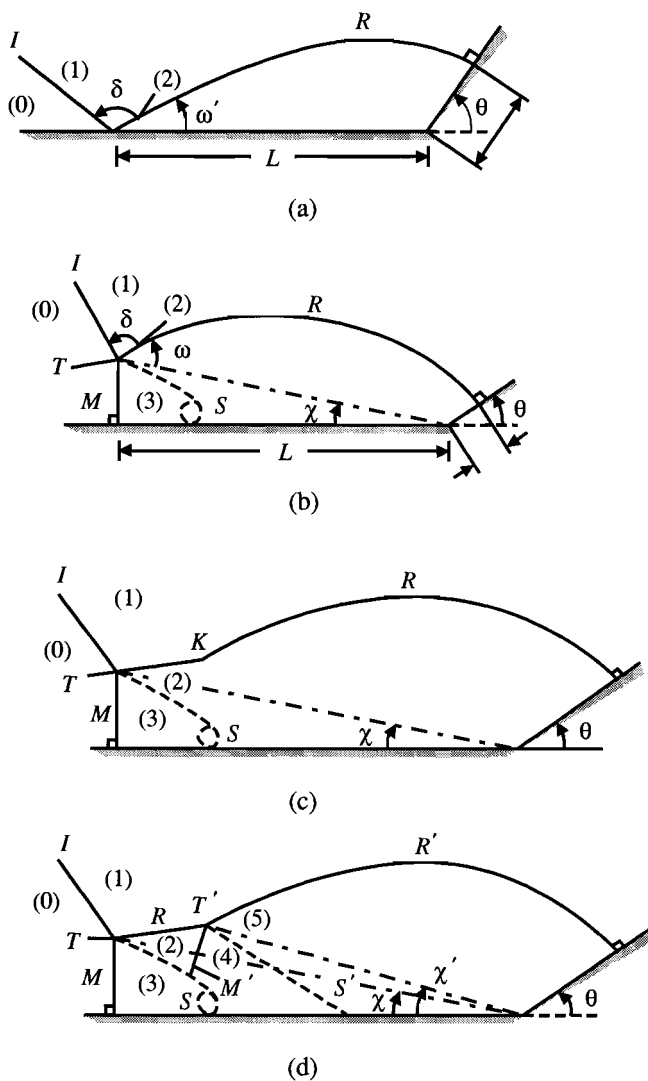


FIGURE 1. Four different types of nonlinear shock diffraction by a ramp. The letter  $I$  denotes the incident wave while  $R$  denotes the reflected wave.

and displaying all fluid variables, and in gaining a theoretical understanding is enormously important—an example is presented in §4(B) where a new phenomenon in fluid flow involving supersonic vortex sheets has been explained through interaction between large scale computation and theory. For the problems described here, the diffraction effects at shock waves are important; von Neumann recognized many of the possible subtleties and paradoxes in diffraction of nonlinear sound waves (see von Neumann's collected works) but the mechanisms are poorly understood.

I end this section with some brief speculation regarding an approach to explaining the transition from Figure 1(c) to 1(d)—the formation of the second Mach stem. In the vicinity of point  $K$  in Figure 1(c), the fluid flow undergoes a transition from subsonic to supersonic, i.e., the local Mach number changes from  $M < 1$  to  $M > 1$ . The steady flow in the vicinity of  $K$  can be described by an equation of mixed type with a free surface defined by the reflected shock front. I suspect that the associated linearized equation becomes ill-posed at a critical value of wave strength in a similar fashion as described by Morawetz in her work on perturbed shockless airfoils [45]. If this occurs, the complex wave pattern in 1(d) could be explained by incorporating small amplitude nonlinear effects through appropriate geometric optics as described in §2(C).

In other very interesting work in large scale computing on diffraction problems Grove and Glimm [46, 47] have discovered new anomalous wave patterns in shock refraction from an interface between two fluids. Another outstanding theoretical problem is to explain the phenomena observed in these recent calculations.

### 3. Vorticity waves and turbulence.

(A) *The equations of compressible and incompressible flow.* I recall that the Mach number is the ratio of the typical fluid velocity to the speed of sound. Since the speed of sound in air at room temperature is about 300 meters/sec, most fluid flows in our everyday experience satisfy  $M \ll 1$ . In the low Mach number limit,  $M \downarrow 0$ , appropriate solutions of the fluid equations in (1.1) converge to solutions of the *Navier-Stokes equations of incompressible fluid flow* given by

$$(3.1) \quad \frac{Dv}{Dt} = -\nabla p + \nu \Delta v, \quad \operatorname{div} v = 0,$$

where  $v = {}^t(v_1, v_2, v_3)$  is the fluid velocity,  $p(x, t)$  is the scalar pressure, and  $\frac{D}{Dt} = \frac{\partial}{\partial t} + \sum_{j=1}^3 v_j \frac{\partial}{\partial x_j}$ . The coefficient  $\nu$  satisfies  $\nu \ll 1$  in turbulent regimes and the equations in (3.1) with  $\nu = 0$  are called the *incompressible Euler equations*.

It is not at all obvious that the solutions of (1.1) converge to those of (3.1) as  $M \rightarrow 0$  because some of the coefficients in (1.1) are becoming infinite as  $M \downarrow 0$ . From (1.8) we see that the sound wave speeds in (1.8A) are becoming



infinite while the vorticity wave speeds in (1.8B) remain finite as  $M \downarrow 0$ . The equations in (3.1) retain only the vorticity waves because the sound waves are moving infinitely fast. The reader can readily verify that the nonlinear vorticity waves in (1.14) are exact solutions of both the equations in (1.2) and the incompressible Euler equations. Rigorous convergence theorems for the incompressible limit due to Ebin [48] and Klainerman and the author [49, 50] are of recent vintage even though the folklore of the limit has been understood for over a century (see Chapter 2 of [3] for a discussion of the formal limit). Ordinary turbulence occurs in solutions of (3.1) with  $\nu \ll 1$ , thus vorticity waves are important in turbulence for  $M \ll 1$  but sound waves are not.

(B) *An outstanding open problem: Breakdown for the 3-D Euler equations.* Given the fact that the equation in (3.1) only involves the propagation of vorticity waves, it is revealing to write the equations in (3.1) in an equivalent fashion in terms of the vorticity,  $\omega = \text{curl } v$ . Taking the curl of the equations in (3.1) leads to the equation,

$$(3.2) \quad \frac{D\omega}{dt} = \mathcal{D}\omega + \nu\Delta\omega,$$

where  $\mathcal{D}$  is the  $3 \times 3$  symmetric matrix, the deformation matrix, given by

$$(3.3) \quad \mathcal{D} = \frac{1}{2}(\nabla v + (\nabla v)^T).$$

The velocity,  $v$ , is determined from the vorticity  $\omega$  via potential theory from

$$\text{curl } v = \omega, \quad \text{div } v = 0$$

and the resulting formula can be differentiated to compute  $\mathcal{D}$  as a strongly singular integral operator applied to  $\omega$ . In this fashion (3.2) becomes an evolution equation for the vorticity alone which is equivalent to (3.1). The terms  $\frac{D\omega}{dt}$  and  $\nu\Delta\omega$  produce convection and diffusion of vorticity respectively; neither of these effects amplifies vorticity. However, the term  $\mathcal{D}\omega$  on the right-hand side of (3.2) is the “tornado mechanism” which can amplify vorticity enormously. Here is the reasoning: the deformation matrix in (3.3) is a symmetric matrix with zero trace and always has positive eigenvalues provided  $\mathcal{D} \neq 0$ . If the vorticity roughly aligns with an eigenvector corresponding to a positive eigenvalue,  $\omega$  increases. It is not difficult to present examples of exact solutions where  $\omega$  increases. I am somewhat terse here because this material has already been discussed with examples in another paper of the author [51].

The outstanding problem regarding breakdown for the 3-D incompressible Euler equations is the following:

$$(3.4) \quad \begin{array}{l} \text{Are there smooth incompressible velocity fields with finite} \\ \text{energy so that the vorticity accumulates so rapidly that the} \\ \text{solution of 3-D Euler becomes singular at a finite time?} \end{array}$$

This is an outstanding unsolved problem which has been attacked through a large number of ingenious numerical approaches (see [51]) which suggest singularity formation but no rigorous examples of smooth solutions with finite energy forming singularities are known. A theorem of Beale, Lato, and the author [52] states that a smooth solution with finite energy becomes singular at a finite time  $T_*$  if and only if

$$(3.5) \quad \int_0^T |\omega(s)|_{L^\infty} ds \rightarrow \infty \quad \text{as } T \nearrow T_*.$$

Here  $|\omega(s)|_{L^\infty} = \max_{x \in \mathbb{R}^3} |\omega(x, s)|$ . This theorem provides a convenient test for checking numerical computations. It is extremely difficult to devise numerical methods to compute solutions with singularities. Is the actual solution blowing up or just the numerical solution? The above theorem suggests that the quantity to monitor in numerical calculations is given on the left-hand side of (3.5). In particular if a numerical method predicts singularity formation but the quantity in (3.5) remains finite, one is guaranteed that the numerical method is creating fictitious singular solutions. Siggia and Shelley/Meiron have independently been using the test in (3.5) in their recent numerical simulations (personal communication). I have mentioned this possibility here as a simple example of rigorous theory providing guidelines for numerical computation. Other comments of the author regarding (3.4) and additional references are found in [51].

Why is (3.4) a fundamental problem in mathematical fluid dynamics? The reason is that most current turbulence theories seem to imply the existence of many singular solutions for 3-D Euler. For example, Kolmogorov's famous theory requires that the dissipation rate  $\varepsilon$  is constant for small values of  $\nu$ ; this dissipation rate is given by

$$\varepsilon = \nu \overline{\omega^2},$$

where the bar denotes averaging over statistical ensembles of velocity fields. If  $\varepsilon$  is constant as  $\nu \rightarrow 0$ , clearly  $\overline{\omega^2}$  becomes infinite and there should exist plentiful families of singular solutions for 3-D Euler. This reasoning is vague and imprecise but it does motivate the importance of the problem posed in (3.4).

(C) *Current turbulence theories and modern mathematical physics.* This subsection is the shortest of this paper but perhaps the most important for future research developments. This is an exciting time in turbulence theory. Chorin has developed a statistical theory of vorticity in turbulence based on analogies in polymer physics and self-avoiding random walks [53]. V. Yakhot and co-workers [54] have developed a theory of renormalization group turbulence based on analogy with Wilson's renormalization group in critical phenomena. Can these connections between modern physics and turbulent fluids be elucidated and made more precise through an appropriate combination of ideas utilizing all four facets of (0.1)?

**4. Examples of the interaction between large scale computing and modern mathematical theory: Vortex sheets in two distinct regimes of fluid motion.** Vortex sheets are piecewise smooth exact solutions of the fluid equations in (1.2) characterized by continuity of the pressure and normal velocity but with a discontinuity in the tangential velocity. The vorticity in a vortex sheet,  $\text{curl } \vec{v}$ , is a distribution given by a Dirac delta measure supported by the curve. A simple example of a vortex sheet is the steady exact solution of (1.2) defined by

$$(4.1) \quad \begin{pmatrix} \rho \\ v_1 \\ v_2 \end{pmatrix} = \begin{cases} {}^t(\rho_0, 0, V), & x_1 > 0, \\ {}^t(\rho_0, 0, -V), & x_1 < 0, \end{cases}$$

with  $V > 0$ , a constant velocity. This is a weak solution of the fluid equations in (1.2) with the vortex sheet given by the line  $x_1 = 0$  and the vorticity,  $\text{curl } \vec{v}$ , given by

$$(4.2A) \quad \text{curl } \vec{v} = 2V\delta(x_1)$$

with  $\delta(x_1)$  the Dirac mass at  $x_1 = 0$ . The Mach number associated with the vortex subset in (4.1) is given by

$$(4.2B) \quad M = \frac{V}{c(\rho_0)},$$

where  $c^2(\rho_0) = p'(\rho_0)$  defines the speed of sound. In this section, we give examples of the interaction between large scale computing and mathematical theory in studying the evolution of vortex sheets. The mathematical ideas and physical phenomena will be very different depending on the two regimes of motion to be considered as I explain next.

In §4(A), we discuss the evolution of vortex sheets like the one in (4.1) in the regime of low Mach numbers,  $M \ll 1$ . For  $M \ll 1$ , as I discussed in §3, it is appropriate to use the equations of incompressible flow without viscosity. Furthermore, like the other vorticity waves in (1.14), the simple exact solutions in (4.1) are also weak solutions of the equations for incompressible flow. Such thin layers of vorticity occur naturally in fluid flows for instance in the trailing wake of airwings and in the piston motion of an internal combustion engine. Plates 76, 145, and 146 from Van Dyke's book [1] are experimental photographs of vortex sheets with  $M \ll 1$ —the vorticity is concentrated in such thin layers with  $\nu \ll 1$  that idealizations with  $\nu = 0$  and infinitely thin layers are commonly used. Such problems are extremely difficult from both the computational and mathematical point of view for the following reason: for the incompressible limit with  $M \ll 1$ , the exact solutions in (4.1) are violently unstable—the linearized initial value problem about these exact solutions behaves like the initial value problem for the Laplace equation and is strongly ill-posed [55]. The plates 145, 146 from [1] give an experimental confirmation of this behavior which is known as nonlinear Kelvin-Helmholtz instability. The numerical computations and mathematical theory in §4(A)

address the incredibly complex phenomena which occur for longer times and for vortex sheet initial data; these data are not typically real analytic in many practical applications.

In §4(B), we discuss the evolution of vortex sheets like the one in (4.1) in a completely different regime where the flow speed  $V$  is extremely fast, i.e., the Mach number satisfies  $M > 1$  so that these are *supersonic* vortex sheets. There has been increasing applied interest in the structure of perturbed supersonic vortex sheets motivated by both the structure and motion of galactic jets in astrophysics and the possibility of designing space planes which fly at extremely high Mach numbers—the mixing environment for fuel and oxidant involves fluid flows like the one in (4.1) with  $M > 1$ . It is extremely difficult and expensive to do experiments at such high Mach numbers so a combination of large scale computing and mathematical theory is potentially very useful in understanding fluid phenomena in this regime. In fact, in §4(B) we present a completely new mechanism for the nonlinear instability of supersonic vortex sheets at high Mach numbers—these phenomena occur for Mach numbers satisfying  $M > \sqrt{2}$ . For  $M > \sqrt{2}$  vortex sheets develop nonlinear instability in a completely different fashion than through nonlinear Kelvin-Helmholtz instability. In this regime  $M > \sqrt{2}$ , small amplitude nonlinear sound waves interact through shock and rarefaction patterns and generate increasing vorticity through resonant wave interaction. In contrast, for vortex sheets with  $M \ll 1$ , no interaction of sound waves and vorticity can occur in the incompressible limit because the sound waves have been removed (see §3). These new mechanisms of instability have been discovered and understood very recently through a sophisticated combination of numerical experiments and mathematical theory which we summarize briefly in §4(B).

(A) *The nonlinear evolution of vortex sheets for 2-D incompressible flow.* A two-dimensional incompressible velocity field,  $v_0(x)$ , defines *vortex sheet initial data* provided that there is a piece of a smooth curve  $C$  in the plane so that the tangential velocity of  $v_0(x)$  jumps across  $C$  while the normal velocity for  $v_0(x)$  remains continuous. Here, for simplicity, we assume that  $v_0(x)$  is a potential flow outside  $C$ . Thus, the vorticity  $\omega_0 = \text{curl } v_0$  has the following structure:

$$(4.3A) \quad \text{the vorticity } \omega_0 \text{ is a surface Dirac delta measure supported on } C.$$

Another obvious physical requirement for vortex sheet initial data is that  $v_0$  has locally finite kinetic energy, i.e.,

$$(4.3B) \quad \int_{|x| \leq R_0} |v_0|^2 dx \leq C_{R_0} \quad \text{for any } R_0 > 0.$$

In the remainder of this subsection, when we refer to vortex sheet initial data, we tacitly assume the conditions in (4.3). There is an enormous engineering

and applied mathematics literature on these topics involving both formal asymptotic methods and numerical simulation (see [56]).

First I describe some of the remarkable computational results achieved by R. Krasny [57] recently using computational vortex methods with a regularization parameter  $\delta$ . The actual motion of the vortex sheet is recovered in the limit  $\delta \downarrow 0$ . Krasny presents calculations of roll-up vortex sheets without a distinguished sign and the subsequent evolution can be incredibly complex. In the first calculations discussed in [57], Krasny studies the numerical solution for vortex sheet initial data corresponding to an elliptically loaded wing. Figures 7–15 of [57] demonstrate the convergence as  $\delta \downarrow 0$  of the regularized algorithm in the vicinity of the tips which roll-up. As further evidence for the validity of his numerical procedure he also compares the numerical solution with Kaden's self-similar spiral; asymptotic arguments predict this solution controls the behavior of the roll-up. The second calculations reported by Krasny have initial data with a vortex sheet strength that changes sign three times. The vortex sheet rolls-up at six different locations and the different pieces of the sheet globally interact like large scale vortices with various signs. Figure 2 on the next page. (Figure 19 from [57]) shows the middle stages of development while Figure 3 see page 379 (Figure 23 from [57]) shows the incredible small scale complexity in the vortex sheet generated by the large scale coherent structures on the sheet which drive their development. Figures 4, 5, and 6 see page 380 (Figure 24 from [57]) give closeup views of the incredibly complex portions of the vortex sheet that develop as time evolves. The "fat" portions from the vortex sheet in Figures 5 and 6 are artifacts of the graphics printing; in fact the vortex sheet has folded in several closely packed strips. These last calculations use the crude value of  $\delta = 0.1$ —it is difficult to imagine the structure of the solution as  $\delta \downarrow 0$ . A mathematical framework designed to address such observed complexity is discussed later in this section.

Next I describe another very interesting numerical computation of a perturbed periodic vortex sheet by Shelley and Baker [58]. They approximate the slightly perturbed initial vortex sheet by a layer with mean finite thickness  $h$  and uniform vorticity inside with strength  $C_0 h^{-1}$ , where  $C_0$  is a given constant. These authors consider a sequence of calculations with these initial data as  $h \rightarrow 0$ , i.e., as the uniform sheet gets thinner—this is another regularization procedure for the singular vortex sheet. They resolve the behavior of the inviscid fluid flow with fixed  $h$  by a sophisticated interface algorithm which tracks the boundaries. The results of their computations with their smallest value of  $h = 0.025$  are presented in Figures 7 and 8 see page 381. The successive times depicted are  $t = 0, 1.5, 2.0, 2.2$ , and  $2.4$ . One interesting facet of these calculations is the approximate Kirchhoff ellipse which forms by time 2.4 in the roll-up process. An ellipse with the same ratio of major and minor axes was present in all the resolved runs from [58] and has

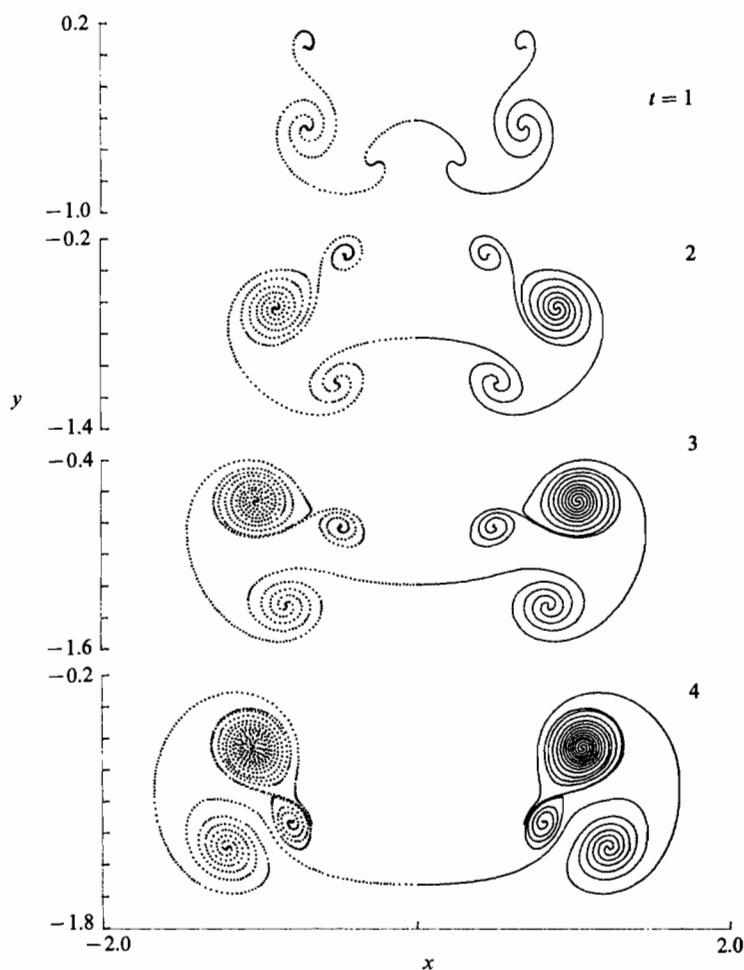


FIGURE 2. The solution plotted over the time interval  $1 \leq t \leq 4$  using  $\delta = 0.1$  (R. Krasny [57]).\*

\* Reprinted with the permission of Cambridge University Press. This figure originally appeared in *Computation of vortex sheet roll-up in the Trefftz plane*, by Robert Krasny, *Journal of Fluid Mechanics*, Cambridge University Press, Cambridge, 1987.

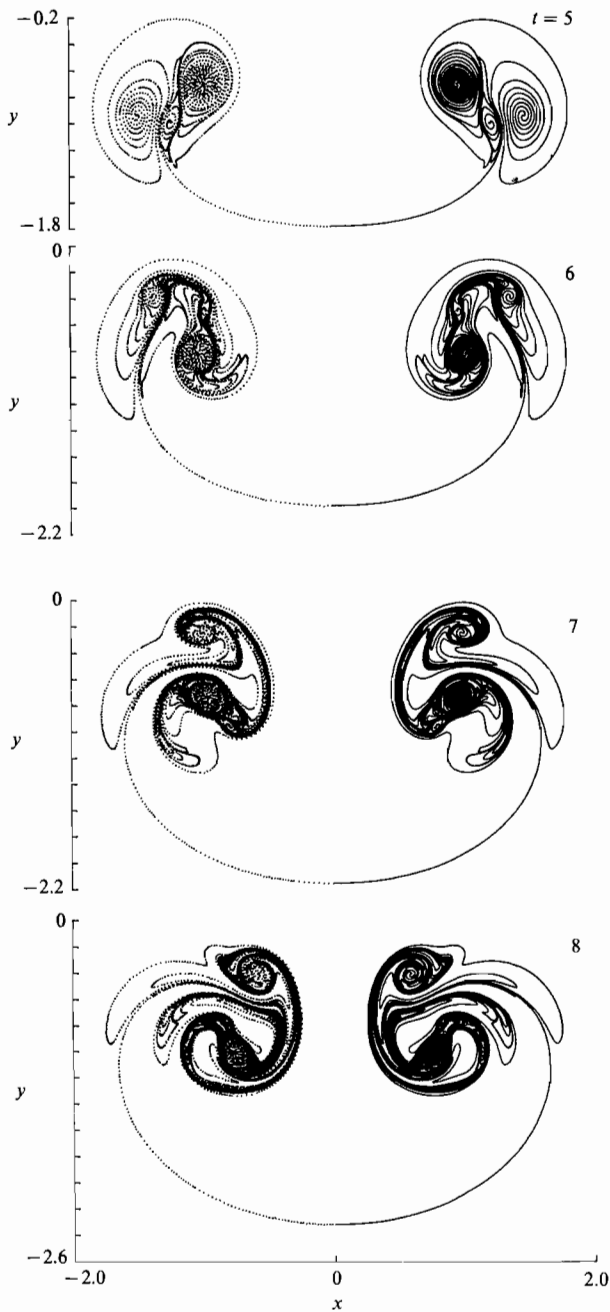


FIGURE 3. Evolution over the time interval  $5 \leq t \leq 8$  for  $\delta = 0.1$  (R. Krasny [57]). \*\*

\*\* Reprinted with the permission of Cambridge University Press. This figure originally appeared in *Computation of vortex sheet roll-up in the Trefftz plane*, by Robert Krasny, *Journal of Fluid Mechanics*, Cambridge University Press, Cambridge, 1987.

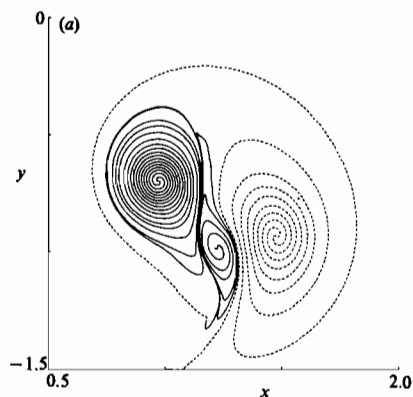


FIGURE 4. Closeup view of the solution at  $t = 5$  (R. Krasny [57]).<sup>†</sup>

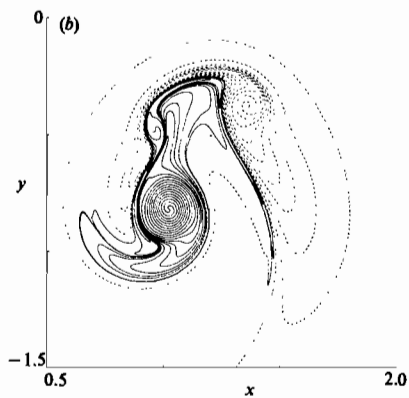


FIGURE 5. Closeup view of the solution at  $t = 6$  (R. Krasny [57]).<sup>†</sup>

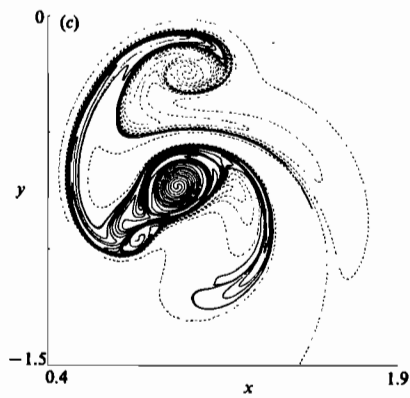


FIGURE 6. Closeup view of the solution at  $t = 7$  (R. Krasny [57]).<sup>†</sup>

<sup>†</sup> Reprinted with the permission of Cambridge University Press. This figure originally appeared in *Computation of vortex sheet roll-up in the Trefftz plane*, by Robert Krasny, Journal of Fluid Mechanics, Cambridge University Press, Cambridge, 1987.



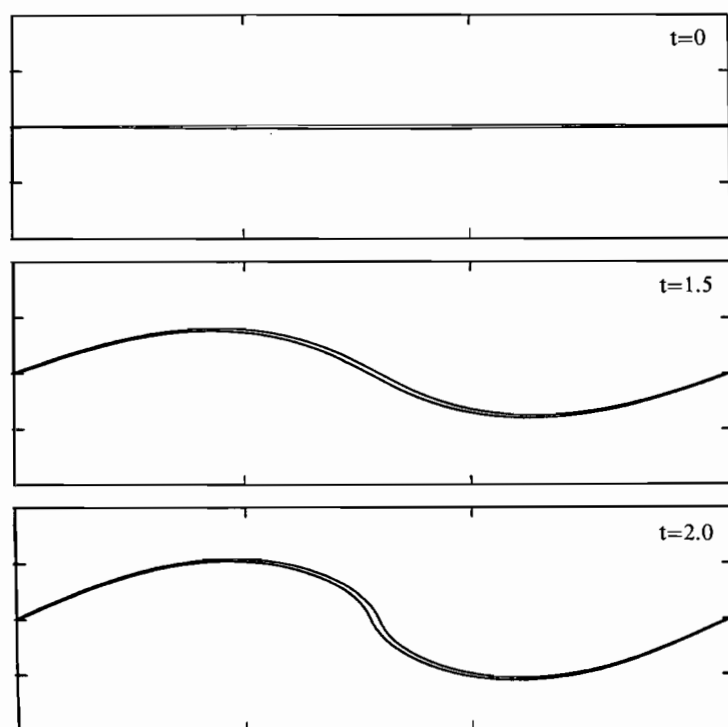


FIGURE 7. The solution at time  $t = 0, 1.5$ , and  $2.0$  with  $h = 0.025$ .<sup>††</sup>

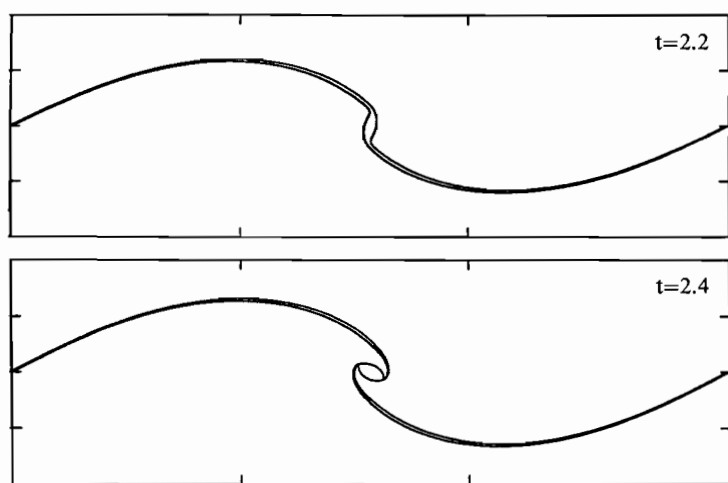


FIGURE 8. The solution at time  $t = 2.2, 2.4$  with  $h = 0.025$ .<sup>††</sup>

<sup>††</sup> Provided by Michael Shelley. More complete results appear in [58].

an area which scales with  $h$  as  $O(h^{1.6})$  as  $h \downarrow 0$ . This fact together with the fact that that vorticity has a single sign is an important aspect of these calculations for the mathematical theory which I sketch below.

I briefly describe a program developed in joint work by DiPerna and the author [59, 60, 61] specifically designed to address the new phenomena that arise in taking limits of suitable approximate solution sequences for 2-D and 3-D Euler. I mostly discuss this theory and its implications for approximate solution sequences for 2-D Euler with vortex sheet initial data and indicate how the above theory provides insight into the trends observed in the calculations of Krasny and Shelley and Baker. A leisurely but more detailed discussion than presented here of the work in [59]–[61] is presented in another recent paper of the author [62].

APPROXIMATE SOLUTION SEQUENCES FOR 2-D EULER WITH VORTEX SHEET INITIAL DATA. I consider an initial incompressible velocity field  $v_0$  in two space dimensions for the evolution of a vortex sheet. Thus I assume that  $v_0$  has the structure (4.3), i.e., the vorticity  $\omega_0$  is a surface Dirac delta measure and the velocity  $v_0$  has locally finite kinetic energy. The basic physical problem involved in the evolution of vortex sheets in the high Reynolds number limit is the following:

- (4.4) If  $v^\nu(x, t)$  is the solution of the Navier-Stokes equations with vortex sheet initial data  $v_0$ , does  $v^\nu(x, t)$  converge in the high Reynolds number limit as  $\nu \rightarrow 0$  to a solution of the inviscid 2-D Euler equations? Do new phenomena occur in the limiting process? Do solutions of inviscid 2-D Euler with vortex sheet initial data exist for all time?

I have just described two other ways to generate approximate solution sequences for 2-D Euler with vortex sheet initial data via computational algorithms. The same basic questions described in the first part of (4.4) apply to the two computational regularizations for 2-D Euler just described above. Another important question is the following:

- (4.5) Do different regularizations for 2-D Euler with vortex sheet initial data converge to the same answer?

In §1 of [60], DiPerna and the author introduce the concept of approximate solution sequence for 2-D Euler. Loosely speaking, we say that  $v^\varepsilon$  is an approximate solution sequence for 2-D Euler provided that the following three conditions are satisfied:

(1)  $v^\varepsilon$  is an incompressible velocity field with *local kinetic energy* uniformly bounded independent of  $\varepsilon$ , i.e.,

$$\max_{0 \leq t \leq T} \int_{|x| \leq R} |v^\varepsilon|^2 dx \leq C(R, T)$$

for any  $R, T > 0$ .

(2) The vorticity  $\omega^\varepsilon = \text{curl } v^\varepsilon$  satisfies

$$(4.6) \quad \max_{0 \leq t \leq T} \int |\omega^\varepsilon| dx \leq C.$$

(3) The *sequence of velocity fields*  $v^\varepsilon$  is *velocity consistent with 2-D Euler*, i.e., for all smooth vector test functions  $\phi(x, t) = (\phi_1, \phi_2)$ , with bounded support and  $\text{div } \phi = 0$ , as  $\varepsilon \rightarrow 0$

$$\iint (\phi_t \cdot v^\varepsilon + \nabla \phi : v^\varepsilon \otimes v^\varepsilon) dx dt \rightarrow 0.$$

Here  $v \otimes v = (v_i v_j)$ ,  $\nabla \phi = (\partial \phi_i / \partial x_j)$ , and  $A : B$  denotes the matrix product  $\sum_{i,j} a_{ij} b_{ij}$ . The estimate in (2) of (4.6) is the natural one for 2-D vortex sheet data (see [60]).

I note that for vortex sheet initial data  $v_0$ , if 2-D Euler has a solution  $v(x, t)$  then this solution is not smooth and would only be a solution in the “weak” distributional sense, i.e., for all smooth  $\phi$  with  $\text{div } \phi = 0$

$$(4.7) \quad \iint (\phi_t \cdot v + \nabla \phi : v \otimes v) dx dt = 0,$$

$\text{div } v = 0$  in the sense of distributions.

The equation in the first part of (4.7) arises from writing the 2-D Euler equations in conservation form, multiplying the  $\phi$ , and integrating by parts in a fashion familiar to the reader perhaps from equation (2.4) in hyperbolic shock wave theory. With the natural definition in (4.7), I observe that the condition in (3) of (4.6) is the minimum requirement needed for a sequence of approximate solutions to have any change of converging to a solution of 2-D Euler as  $\varepsilon \rightarrow 0$ . A portion of [60] is devoted to a proof of the following important results:

**THEOREM.** *All three of the regularization processes described above with vortex sheet initial data generate approximate solution sequences for 2-D Euler satisfying the conditions in (4.6). For the high Reynolds number limit of the Navier-Stokes equations,  $\nu = \varepsilon$ ; for the regularization of Shelley and Baker,  $h = \varepsilon$ ; while for the class of computational vortex methods described in [60],  $\varepsilon = \delta$  with  $\delta$  and  $h$  suitably linked.*

The complexity observed in the calculations of Krasny indicates that the limits as  $\varepsilon \rightarrow 0$  of approximate solution sequences can be incredibly complex. Do new phenomena occur in the limiting process for approximate

solution sequences? Examples indicate that the answer is yes. The simplest way to generate examples of approximate solution sequences is to take exact solutions for 2-D Euler satisfying the conditions in (4.6)—of course, (4.6)(3) is trivially satisfied for these sequences. Two examples discussed in [59, 60] are generated by utilizing swirling flows given by

$$(4.8) \quad v = \begin{pmatrix} -x_2 \\ x_1 \end{pmatrix} |x|^{-2} \int_0^{|x|} s\omega(s) ds,$$

where  $\omega(r)$  is a radial function. These are well-known exact solutions of 2-D Euler with  $\text{curl } v = \omega(r)$ .

EXAMPLE 1. Pick a *positive radial vorticity* distribution  $\omega \geq 0$  with bounded support and define  $v^\varepsilon$  to be the scaled exact swirling flow

$$v^\varepsilon = \left( \log \frac{1}{\varepsilon} \right)^{-1/2} \frac{1}{\varepsilon} v \left( \frac{x}{\varepsilon} \right),$$

with  $v$  given from  $\omega$  by (4.8). Then all the assumptions in (4.6) are satisfied.

EXAMPLE 2. Pick a *radial vorticity distribution* with bounded support but *zero total circulation*, i.e.,  $\int_0^\infty s\omega(s) ds = 0$ , and define  $v^\varepsilon$  by

$$v^\varepsilon = \varepsilon^{-1} v \left( \frac{x}{\varepsilon} \right),$$

with  $v$  given from  $\omega$  by (4.8). Then all of the assumptions in (4.6) are satisfied. These exact solutions are called “phantom” vortices in [60] because these swirling flows vanish identically outside a circle of radius  $O(\varepsilon)$  as  $\varepsilon \rightarrow 0$  (see [59, 60] and the recent paper by Greengard-Thomann [63] for more sophisticated examples).

What happens to the limit as  $\varepsilon \rightarrow 0$  of these exact swirling flows? First, it is easy to see in both examples that

$$(4.9) \quad v^\varepsilon \rightharpoonup 0$$

although the convergence is weak and certainly not uniform. On the other hand, if we multiply the nonlinear terms  $v_i^\varepsilon v_j^\varepsilon$  by a smooth function  $\phi(x_1, x_2)$  with bounded support and average,

$$(4.10) \quad \lim_{\varepsilon \rightarrow 0} \int_{\mathbb{R}^2} \phi v_i^\varepsilon v_j^\varepsilon = C \phi(0) \delta_{ij}, \quad C \neq 0,$$

where

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases}$$

The constant  $C$  differs in Example 1 and Example 2 and depends on different averages of the vortex core structure in each of the two different cases. In the language of distributions, (4.10) means that

$$(4.11) \quad v^\varepsilon \otimes v^\varepsilon \rightharpoonup C \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \delta(x),$$

with  $\delta(x)$  the Dirac delta function at the origin. Naively, one might have expected from (4.9) that  $v^\varepsilon \otimes v^\varepsilon \rightarrow 0$ . Instead new phenomena of *concentration* have occurred in the limit. A *finite amount of local kinetic energy* (exactly  $2C$ ) *has been lost in the limit* in these examples and *concentrates on a small set of measure zero*, the origin in  $R^2$ . Thus, new phenomena of concentration occur in limits of approximate solution sequences. The concept of measure-valued solution for 2-D Euler is introduced in [59, 60] to allow for such potential complexity in the limiting process. This concept of measure-valued solution with concentrations and oscillations was strongly motivated by Tartar's [21] earlier use of the Young measure to study oscillations. One important fact proved in [59, 60] is the following

**THEOREM.** *Every approximate solution sequence for 2-D Euler with vortex sheet initial data converges for all time to a measure-valued solution of 2-D Euler. This solution has concentrations but no oscillations.*

I remark that the concept of measure-valued solution guarantees only that the 2-D Euler equations are satisfied in a very weak sense involving expected values of certain probability measures. Nevertheless, this is an extremely flexible concept. Thus, not every measure-valued solution for 2-D Euler is a weak solution as defined in (4.7) although the converse is true. The exact solution sequences from Example 1 and Example 2 generate examples of nontrivial measure-valued solutions (see [60]). In fact, the author conjectures that if one takes a limit as  $\delta \downarrow 0$  for a sequence of calculations like the one of Krasny depicted in Figures 2–6 with the crude value of  $\delta = 0.1$ , the following scenario is possible: there is a critical time  $t_c$  so that for  $t < t_c$  the limit is a weak solution of 2-D Euler in the standard sense of (4.7) while for  $t > t_c$  the weak solution bursts into a much more complex measure-valued solution for 2-D Euler. The guess that a measure-valued solution occurs for  $t > t_c$  is based not only on the enormous complexity of the evolving vortex sheet for crude  $\delta = 0.1$  but also because this complexity occurs as a consequence of the fact that *vorticity of different signs concentrates* in a *small region of space* and attempts to cancel in an inviscid flow. The simple Example 2 involving phantom vortices necessarily has vorticity with changing signs; as explained in detail in §1 of [60], much less singular local behavior occurs in the concentrations from Example 2 than in those from Example 1 with a fixed sign of the vorticity. Since the behavior of concentration is less singular when the vorticity changes sign, the changes of developing concentrations are much greater.

Next, through a combination of theory from [60] and observed trends in the numerical data, I present strong evidence that in the limit as the thickness  $h \rightarrow 0$  for the calculations of Shelley and Baker depicted in Figures 7 and 8 with  $h = 0.025$ , the limit is expected to be an ordinary weak solution of 2-D Euler. The vorticity in this calculation has a distinguished positive sign. Theorem 3.1 of [60] contains a criterion to check whether a given

approximate solution sequence converges to a classical weak solution of 2-D Euler; this criterion is especially useful when the vorticity has one sign. For the regularization considered by Shelley and Baker, the far-field condition in (3.5) of [60] is readily verified. Thus, for the specific regularization used by Shelley and Baker, Theorem 3.1 from [60] has the following special form:

**THEOREM.** *Assume that the approximate vorticity  $\omega^\varepsilon$  has a distinguished sign,  $\omega^\varepsilon(x, t) \geq 0$ . Also assume that*

$$(4.12) \quad \max_{\substack{x_0 \in \mathbb{R}^2 \\ 0 \leq t \leq T}} \int_{|x-x_0| \leq R} \omega^\varepsilon dx \leq C \log \left( \frac{1}{R} \right)^\beta \quad \text{for all } R \leq R_0$$

*for some fixed constant  $C$  and some  $\beta$  with  $\beta > 1$ . Then as  $\varepsilon \rightarrow 0$  the corresponding velocity fields  $v^\varepsilon$  of the approximate solution sequence converge strongly to an ordinary weak solution of 2-D Euler satisfying (4.6) for  $0 < t < T$ .*

Next, I check the criterion in the above theorem according to the computational trends depicted in Figures 7 and 8. We recall that the approximate Kirchhoff ellipse that occurs at time  $t = 2.4$  was present in a sequence of three resolved runs with roughly the same ratio of major to minor axes in the ellipse as  $h \downarrow 0$  while the area of this ellipse was  $O(h^{1.6})$ . Since the vorticity has constant strength  $O(h^{-1})$  inside the bounding curves, with the information just presented, the maximum value of the ratio of the left- and right-hand sides of (4.12) occurs at  $t = 2.4$  at the center of the ellipse for  $R \approx h^{0.8}$ . Since

$$h^{0.6} \ll (\log(h^{-0.8}))^{-\beta} \quad \text{as } h \downarrow 0,$$

for any  $\beta > 1$ , the criterion in (4.12) is satisfied and the limit is expected to be a classical weak solution for 2-D Euler.

I mention here that the criterion in (4.12) is almost sharp; the sequence of swirling flows with vorticity of positive sign from Example 1 satisfies an estimate like (4.12) with the value  $\beta = 1/2$  but develops concentrations and does not converge strongly. Thus, a criterion such as  $\beta > 1$  is needed and almost sharp. When the vorticity locally has a distinguished sign, formal asymptotic methods [56] often predict that the left-hand side behaves like  $O(h^\alpha)$  with  $\alpha > 0$ , thus a classical weak solution is predicted by the Theorem in these instances.

I also remark that it is still possible for the limit of an approximate solution sequence to be an ordinary weak solution of 2-D Euler even though concentrations develop and there is a loss of kinetic energy—this possibility is explored in detail in [61] with several positive results and is called concentration-cancellation. A more leisurely mathematical discussion is presented in [62]. Also, while the author knows of no explicit rigorous examples where a weak solution for 2-D Euler with vortex sheet initial data bursts at a certain time into a measure-valued solution as expected in the  $\delta \downarrow 0$  limit

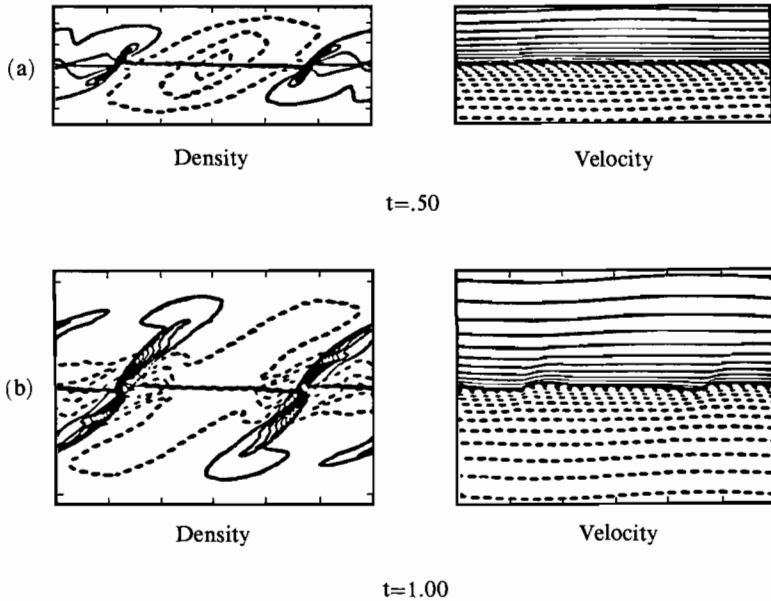


FIGURE 9. The density and velocity profiles at (a)  $t = 0.50$  and (b)  $t = 1.00$ .

of Krasny's calculations, there are explicit examples where a weak solution of the 1-D Vlasov-Poisson equations bursts at a critical time into a measure-valued solution with concentrations in the charge density. In a very precise sense certain weak solutions of the 1-D Vlasov-Poisson equations have analogous but much simpler behavior than 2-D Euler. This is being developed in a forthcoming paper of the author [64].

(B) *The nonlinear development of instability for supersonic vortex sheets.* I begin this section by giving some graphs from numerical calculations of Paul Woodward [65, 66, 67] involving initial data consisting of small amplitude perturbations of the supersonic vortex sheet in (4.1) at Mach number  $M = 1.5$ . Figures 9 (above) and 10–12 (pages 388–390) give the density contours and velocity contours of the solution that evolves at the successive times  $0.50 \leq t \leq 500$ . The density contours are on the left in each figure.

It is clear from this sequence of pictures that the small amplitude perturbation grows into a large amplitude instability in the basic vortex sheet from (4.1) by the finite time,  $t = 5.00$ . Furthermore, one can see the nature of the instability process. By the time  $t = 1.50$ , kinks have formed in the surface; these kinks are accompanied by a rarefaction wave (depicted by the fan in the density contour on one side of the vortex sheet) and a shock wave (depicted by the abrupt black lines indicating discontinuities in the density contour) and occur in pairs due to symmetry in the initial data. The reader might recall the solutions in (1.13B) at this point. These “kink modes” travel

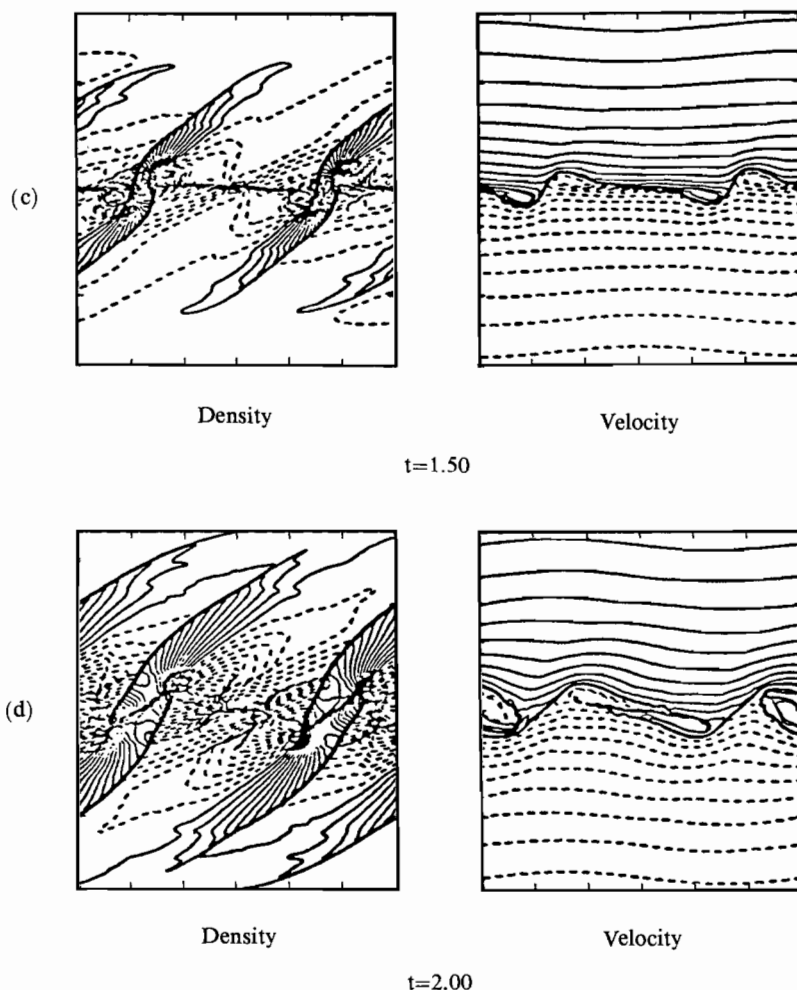


FIGURE 10. The density and velocity profiles at (c)  $t = 1.50$  and (d)  $t = 2.00$ .

along the vortex sheet at different speeds and collide and generate more kink modes with a similar structure by the time  $t = 5.00$  as depicted in Figures 10–12. The velocity contours show that an ever increasing amount of vorticity has been generated from the nonlinear interaction of the kink modes. When the calculations are continued beyond time  $t = 5.00$ , enough vorticity is generated through the interaction of the kink modes that the vortex sheet eventually develops rolls much like the situation described earlier with  $M \ll 1$ . In numerical computations of Woodward with  $M < \sqrt{2}$ , the kink modes never appear at small amplitudes and the vortex sheet develops the expected nonlinear Kelvin-Helmholtz instability.



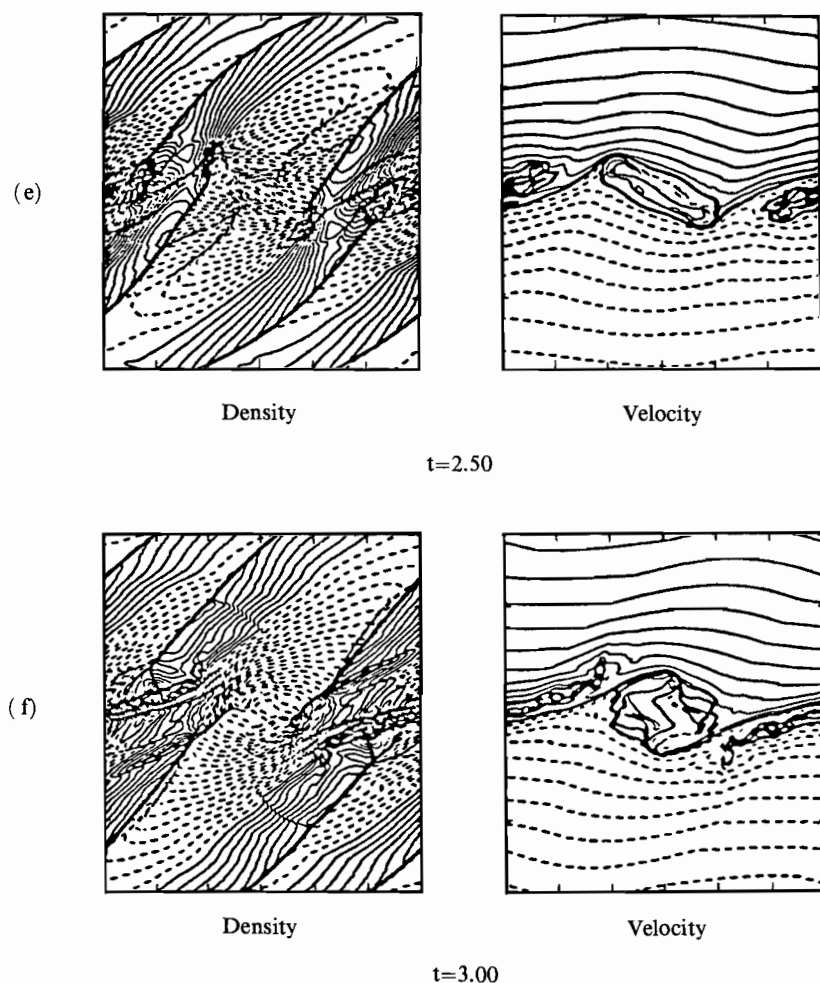


FIGURE 11. The density and velocity profiles at (e)  $t = 2.50$  and (f)  $t = 3.00$ .

The calculations of Woodward suggest the following theoretical questions:

- (4.13) Why do “kink” modes form on supersonic vortex sheets?  
 Once such modes form, how do they interact to generate increasing vorticity? What are the quantitative conditions for the formation of kink modes?

Since kink modes are very complicated nonlinear structures simultaneously involving kinks in the vortex sheet bracketed by shock and rarefaction waves, the answers to these questions are not apparent. Furthermore, there is an interesting

PARADOX. The classical engineering and physics linearized stability analysis predicts no growing modes and stability for Mach numbers  $M > \sqrt{2}$ .

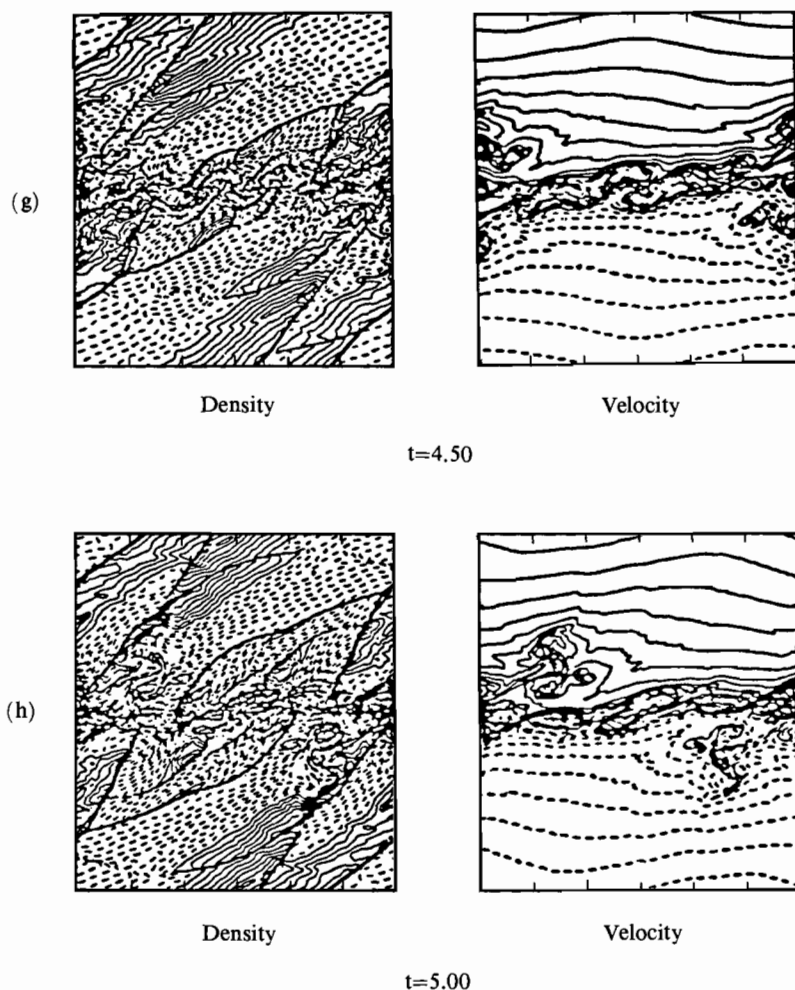


FIGURE 12. The density and velocity profiles at (g)  $t = 4.50$  and (h)  $t = 5.00$ .

The author and his recent Ph.D. student M. Artola have developed a program which answers all of the questions posed in (4.13) and resolves the paradox just mentioned [68, 69, 70]. This program has three parts:

(1) A quantitative explanation for the appearance of three different families of travelling kink modes at small amplitudes for  $M > \sqrt{2}$  together with quantitative prediction of the complete small amplitude shock and rarefaction struction [68].

(2) A quantitative theory for the resonant nonlinear interaction of kink modes at the boundary including the increase in vorticity from interaction of the kink modes [69].

(3) Exact solutions of the fluid equations given by kink structures at both

large and small amplitudes via bifurcation analysis [70].

I do not have the space to develop any of the parts of this program in detail here; however, I make a few comments regarding the methods and results in each part.

In Part (1), the methods of nonlinear geometric optics are generalized to apply to the complex free surface problem defined by the perturbed vortex sheet. Through geometric optics, nonlinear sound waves are generated which strike the vortex sheet and a parameter, the angle of incidence, is varied. For most angles of incidence the response is a transmitted nonlinear sound wave and a reflected nonlinear sound wave with the same magnitude as the incident wave. However, *for  $M > \sqrt{2}$ , there are three angles of incidence where nonlinear resonance occurs and the transmitted and reflected sound waves are an order of magnitude larger.* For these three critical directions, simplified asymptotic equations are obtained for the resonant response in a more sophisticated but similar fashion as described in §2(C). The simplified asymptotic equations consist of an appropriate Hamilton-Jacobi equation for the vortex sheet perturbation coupled with boundary value problems for two nonlinear transport equations as in (2.29) for the reflected and transmitted sound waves. These equations are readily solved exactly and automatically predict the time dependent development of three distinct types of kink modes for  $M > \sqrt{2}$ . For  $1 < M < \sqrt{2}$ , there is only a single type of kink mode and for  $M < 1$ , no such kink modes exist. The details are presented in [68].

In Part (2) a theory for resonant interaction of kink modes is developed. Since for  $M > \sqrt{2}$  there are three kink modes, it is possible for these kink modes to resonantly interact and generate more vorticity. The arguments in [69] are extremely technical since there are double resonances involved—the kink modes are generated by boundary resonances and simultaneously resonantly interact with each other. The result of this technical analysis is an interesting set of simplified asymptotic equations for the resonant interaction of kink modes. These equations for the resonant interaction on the vortex sheet consist of Hamilton-Jacobi equations coupled through resonant integro-differential convolution operators in a fashion already described in (2.31)—one important difference is that the convolution operators in this case are symmetric and allow rapid local amplification. This local amplification is coupled to the growth of vorticity automatically through the asymptotics. Exact solutions of these asymptotic equations display the rapid production of vorticity through these resonant effects as observed in Woodward's calculations.

Finally, I discuss Part (3) of the program [70]. Here classical oblique wave theory from [2] is utilized to solve a bifurcation problem for exact solutions of the fluid equations consisting of kink modes. Unlike the arguments in Parts (1) and (2), time does not enter explicitly in this steady state analysis. These exact solutions are used at small amplitudes to provide a completely independent confirmation of the time dependent arguments in Part (1). The

exact solutions are also used at large amplitudes to explain the computational results of Woodward for Mach numbers satisfying  $1 < M < \sqrt{2}$ . In this regime kink modes are not observed at small amplitudes but only at large amplitudes in a very specific way in Woodward's computations and the special nature of the bifurcation diagrams in this regime explains this phenomena.

The graphs from Figures 9–12 are from the early unpublished calculations of Woodward. Recently [67] he has confirmed the quantitative predictions of Part (1) in detailed calculations and has generated a color movie with this and other interesting results.

**Acknowledgment.** The author thanks P. Colella, R. Krasny, M. Shelley, and P. Woodward for the use of unpublished graphs from their numerical simulations.

### BIBLIOGRAPHY

1. M. Van Dyke, *An album of fluid motion*, Parabolic Press, Stanford, CA, 1982.
2. R. Courant and K. Friedrichs, *Supersonic flow and shock waves*, Springer-Verlag, New York, 1949.
3. A. Majda, *Compressible fluid flow and systems of conservation laws in several space variables*, Appl. Math. Sciences, vol. 53, Springer-Verlag, New York, 1984.
4. P. Lax, *Hyperbolic systems of conservation laws*. II, Comm. Pure Appl. Math. **10** (1957), 537–567.
5. —, *Shock waves and entropy*, Contributions to Nonlinear Functional Analysis (E. A. Zarantonello, ed.), Academic Press, New York, 1971.
6. —, *Hyperbolic systems of conservation laws and the mathematical theory of shock waves*, Regional Conf. Ser. Appl. Math., No. 13, SIAM, Philadelphia, 1973.
7. J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math. **18** (1965), 697–715.
8. J. Smoller, *Shock waves and reaction diffusion equations*, Springer-Verlag, New York, 1983.
9. J. Glimm and P. Lax, *Decay of solutions of systems of nonlinear hyperbolic conservation laws*, Mem. Amer. Math. Soc., No. 101, Amer. Math. Soc., Providence, RI, 1970.
10. T. P. Liu, *Admissible solutions to systems of conservation laws*, Mem. Amer. Math. Soc., No. 240, Amer. Math. Soc., Providence, RI, 1982.
11. C. Dafermos, *Hyperbolic systems of conservation laws*, Systems of Nonlinear Partial Differential Equations (J. Ball, ed.), Reidel, Boston, 1983, pp. 25–70.
12. A. Chorin, *Random choice solutions of hyperbolic systems*, J. Comput. Phys. **22** (1976), 517–533.
13. H. Weyl, *Shock waves in arbitrary fluids*, Comm. Pure Appl. Math. **2** (1949), 103–122.
14. D. Gilbarg, *The existence and limit behavior of the one dimensional shock layer*, Amer. J. Math. **7** (1951), 256–274.
15. R. Pego, *Nonexistence of a shock layer in gas dynamics with a nonconvex equation of state*, Arch. Rational Mech. Anal. **94** (1986), 165–178.
16. A. Majda and R. Pego, *Stable viscosity matrices for systems of conservation laws*, J. Differential Equations **56** (1985), 229–262.
17. T. P. Liu, *Shock waves for compressible Navier-Stokes equations are stable*, Comm. Pure Appl. Math. **39** (1986), 565–594.
18. —, *Nonlinear stability of shock waves for viscous conservation laws*, Mem. Amer. Math. Soc., no. 328, Amer. Math. Soc., Providence, RI, 1985.
19. J. Goodman, *Nonlinear asymptotic stability of viscous shock profiles for conservation laws*, Arch. Rational Mech. Anal. **95** (1986), 325–344.
20. T. P. Liu and D. Hoff, *The inviscid limit for the Navier-Stokes equations of compressible isentropic flow with shock data*, preprint, 1988.

21. L. Tartar, *Compensated compactness and applications to partial differential equations*, Research Notes in Mathematics, Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, vol. 4 (R. Knops, ed.), Pitman, London, 1979.
22. F. Murat, *Compacite par compensation*, Ann. Scuola Norm. Sup Pisa Cl. Sci. (4) **5** (1978), 69–102.
23. R. DiPerna, *Convergence of approximate solutions to conservation laws*, Arch. Rational Mech. Anal. **82** (1983), 27–70.
24. —, *Convergence of the viscosity method for isentropic gas dynamics*, Comm. Math. Phys. **91** (1983), 1–30.
25. —, *Uniqueness of solutions to hyperbolic conservation laws*, Indiana Univ. Math. J. **28** (1979), 137–188.
26. A. Majda, *The stability of multi-dimensional shock fronts*, Mem. Amer. Math. Soc., No. 275, Amer. Math. Soc., Providence, RI, 1983.
27. —, *The existence of multi-dimensional shock fronts*, Mem. Amer. Math. Soc., No. 281, Amer. Math. Soc., Providence, RI, 1983.
28. G. Metevier, *Interaction de deux chocs pour un systeme de deux lois de conservation en dimension deux d'espace*, Trans. Amer. Math. Soc. **296** (1986), 431–479.
29. S. Alinhac, *Existence d'ondes de rarefaction pour des systemes quasilineaires hyperboliques multidimensionnels*, preprint, 1988.
30. E. Harabetian, Ph.D. thesis, U.C.L.A., 1985.
31. A. Majda and E. Thomann, *Multi-dimensional shock fronts for second order wave equations*, Comm. Partial Differential Equations **12** (1987), 777–828.
32. R. Courant and D. Hilbert, *Methods of mathematical physics, Vol. II*, Wiley, New York, 1962.
33. Y. Choi and A. Majda, *Amplification of small amplitude high frequency waves in a reactive mixture*, SIAM Rev. **31** (1989), 401–442.
34. A. Majda, *Lectures on linear and nonlinear hyperbolic waves*, Graduate course, Princeton University, 1986.
35. M. J. Lighthill, *A method for rendering approximate solutions to physical problems uniformly valid*, Philos. Mag. **40** (1949), 1179–1201.
36. G. Whitham, *The flow pattern of a supersonic projectile*, Comm. Pure Appl. Math. **5** (1952), 301–348.
37. Y. Choquet-Bruhat, *Ondes asymptotiques et approches pour des systemes d'equations aux derivees partielles non lineaires*, J. Math. Pures Appl. **48** (1969), 117–158.
38. J. K. Hunter and J. B. Keller, *Weakly nonlinear high frequency waves*, Comm. Pure Appl. Math. **36** (1983), 547–569.
39. A. Majda, *Nonlinear geometrical optics for hyperbolic systems of conservation laws*, The I.M.A. Volumes in Mathematics and Its Applications, vol. 2, Springer-Verlag, New York, 1986, pp. 116–166.
40. A. Majda and R. Rosales, *Resonantly interacting weakly nonlinear hyperbolic waves, I: A single space variable*, Stud. Appl. Math. **71** (1984), 149–179.
41. J. K. Hunter, A. Majda, and R. Rosales, *Resonantly interacting, weakly nonlinear, hyperbolic waves II: Several space variables*, Stud. Appl. Math. **75** (1986), 187–226.
42. P. Cehelsky and R. Rosales, *Resonantly interacting weakly nonlinear hyperbolic waves in the presence of shocks: a single space variable in a homogeneous time independent medium*, Stud. Appl. Math. **74** (1986), 117–138.
43. A. Majda, R. Rosales, and M. Schonbek, *A canonical system of integro-differential equations arising in resonant nonlinear acoustics*, Stud. Appl. Math. **79** (1988), 205–262.
44. H. Glaz, P. Colella, I. I. Glass, and R. Deschambault, *A detailed numerical, graphical, and experimental study of oblique shock wave reflections*, Lawrence Berkeley Report, April 1985.
45. C. S. Morawetz, *On the non-existence of continuous transonic flows past profiles*, I, II, III, Comm. Pure Appl. Math. **9** (1956), 45–68; **10** (1957), 107–132; **11** (1958), 129–144.
46. J. Grove, *The interaction of shock waves with fluid interfaces*, preprint, 1988.
47. J. Grove and J. Glimm, private communication.
48. D. Ebin, *The motion of slightly compressible fluids viewed as motion with a strong constraining force*, Ann. of Math. (2) **150** (1977), 102–163.

49. S. Klainerman and A. Majda, *Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids*, Comm. Pure Appl. Math. **34** (1981), 481–524.
50. —, *Compressible and incompressible fluids*, Comm. Pure Appl. Math. **35** (1982), 629–653.
51. A. Majda, *Vorticity and the mathematical theory of incompressible fluid flow*, Comm. Pure Appl. Math. **39** (1986), S 187–220.
52. J. T. Beale, T. Kato, and A. Majda, *Remarks on the breakdown of smooth solutions for the 3-D Euler equations*, Comm. Math. Phys. **94**, (1984), 61–66.
53. A. J. Chorin, *Spectrum, dimension and polymer analogies in fluid turbulence*, Phys. Rev. Lett. **60** (1988), 1947–1949.
54. V. Yakhot and S. Orszag, *Renormalization group analysis of turbulence I. Basic theory*, J. Sci. Comput. **1** (1986), 3–51.
55. C. Sulem, P. Sulem, C. Bardos, and U. Frisch, *Finite time analyticity for the two and three dimensional Kelvin-Helmholtz instability*, Comm. Math. Phys. **80** (1981), 485–516.
56. P. Saffman and G. Baker, Ann. Rev. Fluid Mech. **11** (1979), 95–122.
57. R. Krasny, *Computation of vortex sheet roll-up in the Trefftz plane*, J. Fluid Mech. (1987).
58. M. Shelley and G. Baker, *On the connection between thin vortex layers and vortex sheets*, J. Fluid Mech. (1990), 161–194, Cambridge University Press.
59. R. DiPerna and A. Majda, Comm. Math. Phys. **108** (1987), 667–689.
60. —, *Concentrations in regularizations for 2-D incompressible flow*, Comm. Pure Appl. Math. **60** (1987), 301–345.
61. —, *Reduced Hausdorff dimension and concentration-cancellation for 2-D incompressible flow*, J. Amer. Math. Soc. **1** (1988), 59–95.
62. A. Majda, *Vortex sheets, potential theory and concentration-cancellation for 2-D incompressible flow*, Princeton lecture notes, 1988.
63. C. Greengard and E. Thomann, *On DiPerna-Majda concentration sets for two-dimensional incompressible flow*, Comm. Pure Appl. Math. **61** (1988).
64. A. Majda, *Concentrations in electron sheets for the 1-D Vlasov-Poisson equations*, in preparation.
65. P. Woodward, in *Numerical methods for the Euler equations of fluid dynamics* (Angrand, Dewieux, Desideri, and Glowinski, eds.), SIAM, Philadelphia, PA, 1985.
66. —, in *Astrophysical radiation hydrodynamics* (K. H. Winkler and M. Norman, eds.), Reidel, 1986.
67. P. Woodward and K. H. Winkler, *Simulation and visualization of fluid flow in a numerical laboratory*, preprint, October 1988.
68. M. Artola and A. Majda, *Nonlinear development of instabilities in supersonic vortex sheets I: the basic kink modes*, Physica D **28** (1988), 253–281.
69. —, *Nonlinear development of instabilities in supersonic vortex sheets II: resonant interaction among kink modes*, SIAM J. Appl. Math. **49** (1989), 1310–1349.
70. —, *Nonlinear kink modes for supersonic vortex sheets*, Phys. Fluids A **1** (1989), 583–596.
71. R. DiPerna and A. Majda, *The validity of nonlinear geometric optics for weak solutions of conservation laws*, Comm. Math. Phys. **98** (1985), 313–347.

DEPARTMENT OF MATHEMATICS AND PROGRAM IN APPLIED AND COMPUTATIONAL MATHEMATICS, PRINCETON UNIVERSITY, PRINCETON, NEW JERSEY 08544

## Two Examples of Mathematics and Computing in the Biological Sciences: Blood Flow in the Heart and Molecular Dynamics

CHARLES S. PESKIN

**Introduction.** Mathematics plays a unifying role in the sciences, since diverse natural phenomena can often be described with the help of similar mathematical methods. This paper discusses two quite different biomathematical problems in which essentially the same issue of numerical stability can be resolved in essentially the same way: through the use of an implicit method that requires the solution of an optimization problem at each time step. In the case of blood flow in the heart, this leads to a computational method that can be used to study the normal and pathological function of the heart and also to predict the performance of proposed designs for prosthetic cardiac valves. In the case of molecular dynamics, the method described in this paper has the unexpected ability to simulate certain quantum-mechanical effects in an otherwise classical context. Moreover, this method should make it possible to use large time steps and thus to simulate slow molecular processes which have hitherto been beyond the reach of molecular dynamics.

**The equations of cardiac fluid dynamics.** Our purpose in this section is to give a new derivation of the equations that describe the mechanical function of the heart. These equations encompass the fluid mechanics of the blood, the passive elasticity of the heart valve leaflets, and the active (time-dependent) elasticity of the muscular heart walls. The entire system is incompressible, and we use the incompressible-elasticity approach of Ebin and Saxton [1, 2], which produces considerable simplification in the equations of nonlinear elasticity. This simplification is achieved through the strategic use of a mixed Lagrangian-Eulerian description of the motion.

One way to express the connection between these two descriptions is to use integral transformations in which the Dirac  $\delta$ -function appears as a kernel.

---

1980 *Mathematics Subject Classification* (1985 Revision). Primary 92–08; Secondary 65L20, 65M12, 70H05, 70L05, 76Z05, 81–08, 81V55, 82B10, 82C31, 92C35, 92C40.

©1992 American Mathematical Society  
0-8218-0167-8 \$1.00 + \$.25 per page

We adopt this approach here, as it illuminates the path that we follow in the construction of a computational method for the solution of the equations derived in this section.

We begin with a Lagrangian description of the motion:

$$(1) \quad \mathbf{x} = \mathbf{X}(a, t),$$

where  $a = (q, r, s)$  is a list of Lagrangian parameters, i.e., fixed  $a$  marks a material point. Let  $m(a)da = m(q, r, s)dq dr ds$  be the mass of the material element  $da$ . Note that  $m$  is independent of time, since mass is conserved.

Let the elastic potential energy of the system be given by a functional  $E[\cdot, t]$  which takes as input the function  $\mathbf{X}(\cdot, t)$ . The explicit time-dependence in  $E$  is what makes it possible for the heart to contract and relax. We shall denote by  $-\mathbf{f}(a, t)$  the Fréchet derivative of  $E$  with respect to  $\mathbf{X}$ . This means that

$$(2) \quad \lim_{\varepsilon \rightarrow 0} \frac{d}{d\varepsilon} E[\mathbf{X} + \varepsilon \mathbf{Y}, t] = - \int \mathbf{f}(a, t) \cdot \mathbf{Y}(a) da$$

for all  $\mathbf{Y}(a)$ .

With the notation developed above, we may write down the Lagrangian of the system as

$$(3) \quad L(t) = \frac{1}{2} \int m(a) \left| \frac{\partial \mathbf{X}}{\partial t}(a, t) \right|^2 da - E[\mathbf{X}(\cdot, t), t].$$

According to Hamilton's principle of least action, the actual motion  $\mathbf{X}(a, t)$  is a stationary point of the action integral

$$(4) \quad S = \int_0^T L(t) dt.$$

This means that if the actual motion  $\mathbf{X}(a, t)$  is embedded in a family of hypothetical motions  $\mathbf{X}(a, t, \varepsilon)$ , all of which are consistent with the constraints (see below) and in such a manner that  $\mathbf{X}(a, t) = \mathbf{X}(a, t, 0)$ , and if  $S(\varepsilon)$  is the action corresponding to the hypothetical motion  $\mathbf{X}(a, t, \varepsilon)$ , then  $dS/d\varepsilon = 0$  at  $\varepsilon = 0$ .

The constraints in question are as follows. First, we assume that the initial and final configurations of the material particles of the system are specified:

$$(5) \quad \mathbf{X}(a, 0, \varepsilon) = \mathbf{X}_0(a) \quad (\text{given}),$$

$$(6) \quad \mathbf{X}(a, T, \varepsilon) = \mathbf{X}_T(a) \quad (\text{given}).$$

Next, we assume that the motions under consideration are all volume-preserving. This means that for any region  $A$  in parameter space and for all  $(t, \varepsilon)$ , we have

$$(7) \quad \text{volume}(\mathbf{X}(A, t, \varepsilon)) = \text{volume}(\mathbf{X}_0(A)).$$



This is the same as saying that

$$(8) \quad \det \left( \frac{\partial \mathbf{X}}{\partial a}(a, t, \varepsilon) \right) = J_0(a)$$

independent of  $t$  and  $\varepsilon$ . This completes the statement of the constraints.

Evaluating  $S(\varepsilon)$ , integrating by parts with respect to time in the kinetic energy term, and making use of the initial and final conditions (5) and (6), we find

$$(9) \quad \begin{aligned} \frac{dS}{d\varepsilon} = & - \int_0^T \int m(a) \frac{\partial^2 \mathbf{X}}{\partial t^2}(a, t, \varepsilon) \cdot \frac{\partial \mathbf{X}}{\partial \varepsilon}(a, t, \varepsilon) da dt \\ & - \int_0^T \frac{d}{d\varepsilon} E[\mathbf{X}(\cdot, t, \varepsilon), t] dt. \end{aligned}$$

Now apply Hamilton's principle and the definition of  $\mathbf{f}$  (eq. (2)) to obtain

$$(10) \quad 0 = \frac{dS}{d\varepsilon}(0) = \int_0^T \int \left[ -m(a) \frac{\partial^2 \mathbf{X}}{\partial t^2}(a, t) + \mathbf{f}(a, t) \right] \cdot \frac{\partial \mathbf{X}}{\partial \varepsilon}(a, t, 0) da dt.$$

Equation (10) holds for all  $\partial \mathbf{X} / \partial \varepsilon$  consistent with the constraints.

Before doing any further analysis of (10), we switch to Eulerian variables. Let  $\mathbf{u}(\mathbf{x}, t)$  and  $\mathbf{v}(\mathbf{x}, t)$  be implicitly defined by

$$(11) \quad \frac{\partial \mathbf{X}}{\partial t}(a, t) = \mathbf{u}(\mathbf{X}(a, t), t),$$

$$(12) \quad \frac{\partial \mathbf{X}}{\partial \varepsilon}(a, t, 0) = \mathbf{v}(\mathbf{X}(a, t), t).$$

In terms of  $\mathbf{u}$  and  $\mathbf{v}$ , the constraint of volume conservation reduces to

$$(13) \quad \nabla \cdot \mathbf{u} = 0,$$

$$(14) \quad \nabla \cdot \mathbf{v} = 0.$$

These results can be derived by differentiating (8) with respect to  $t$  (to obtain  $\nabla \cdot \mathbf{u} = 0$ ) or with respect to  $\varepsilon$  (to obtain  $\nabla \cdot \mathbf{v} = 0$ ).

Differentiating (11) with respect to  $t$ , we obtain

$$(15) \quad \frac{\partial^2 \mathbf{X}}{\partial t^2}(a, t) = \left( \mathbf{u} \cdot \nabla \mathbf{u} + \frac{\partial \mathbf{u}}{\partial t} \right) (\mathbf{X}(a, t), t) = \frac{D\mathbf{u}}{Dt}(\mathbf{X}(a, t), t),$$

where we have introduced the shorthand  $D\mathbf{u}/Dt$  for  $\mathbf{u} \cdot \nabla \mathbf{u} + \partial \mathbf{u} / \partial t$ . Despite its interpretation as the "material derivative" of the velocity, note that  $D\mathbf{u}/Dt$  is a function of  $(\mathbf{x}, t)$ .

Additional functions of  $(\mathbf{x}, t)$  that we shall need are mass density  $\rho(\mathbf{x}, t)$  and the force density  $\mathbf{F}(\mathbf{x}, t)$ . These are defined as follows:

$$(16) \quad \rho(\mathbf{x}, t) = \int m(a) \delta(\mathbf{x} - \mathbf{X}(a, t)) da,$$

$$(17) \quad \mathbf{F}(\mathbf{x}, t) = \int \mathbf{f}(a, t) \delta(\mathbf{x} - \mathbf{X}(a, t)) da.$$

Making use of these definitions, we transform (10) to Eulerian form:

(18)

$$\begin{aligned} 0 &= \int_0^T \int \left[ -m(a) \frac{\partial^2 \mathbf{X}}{\partial t^2}(a, t) + \mathbf{f}(a, t) \right] \cdot \frac{\partial \mathbf{X}}{\partial \mathbf{e}}(a, t, 0) da dt \\ &= \int_0^T \int \int \left[ -m(a) \frac{D\mathbf{u}}{Dt}(\mathbf{x}, t) + \mathbf{f}(a, t) \right] \cdot \mathbf{v}(\mathbf{x}, t) \delta(\mathbf{x} - \mathbf{X}(a, t)) d\mathbf{x} da dt \\ &= \int_0^T \int \left[ -\rho(\mathbf{x}, t) \frac{D\mathbf{u}}{Dt}(\mathbf{x}, t) + \mathbf{F}(\mathbf{x}, t) \right] \cdot \mathbf{v}(\mathbf{x}, t) d\mathbf{x} dt. \end{aligned}$$

To see that the second line of (18) is equivalent to the first, just apply the defining property of the Dirac  $\delta$ -function,  $\int \delta(\mathbf{x} - \mathbf{y}) \phi(\mathbf{x}) d\mathbf{x} = \phi(\mathbf{y})$ . The third line follows by interchanging the order of integration and applying the definitions of  $\rho$  and  $\mathbf{F}$ , (16)–(17).

Now (18) holds for all  $\mathbf{v}$  such that  $\nabla \cdot \mathbf{v} = 0$ . It follows that the coefficient of  $\mathbf{v}$  must be the gradient of some quantity (which is conventionally called the pressure). This gives

$$(19) \quad \rho(\mathbf{x}, t) \left( \frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} \right) + \nabla p = \mathbf{F}(\mathbf{x}, t).$$

Equation (19) may be augmented by the addition of a term that takes into account the presence of viscosity. (Frictional forces are not easily incorporated into the Lagrangian formalism used above.) For simplicity, we use a viscous force of the form  $\mu \nabla^2 \mathbf{u}$  with  $\mu = \text{constant}$ . In the future one might want to modify this to allow for a different form or magnitude of the viscous force in the muscular heart walls or to model the non-Newtonian character of the blood. (The latter effect is a small one in vessels as large as the heart.)

We now summarize the equations of motion (with viscosity included). They are

$$(20) \quad \rho(\mathbf{x}, t) \left( \frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} \right) + \nabla p = \mu \nabla^2 \mathbf{u} + \mathbf{F}(\mathbf{x}, t),$$

$$(21) \quad \nabla \cdot \mathbf{u} = 0,$$

$$(22) \quad \rho(\mathbf{x}, t) = \int m(a) \delta(\mathbf{x} - \mathbf{X}(a, t)) da,$$

$$(23) \quad \mathbf{F}(\mathbf{x}, t) = \int \mathbf{f}(a, t) \delta(\mathbf{x} - \mathbf{X}(a, t)) da,$$

$$(24) \quad \frac{\partial \mathbf{X}}{\partial t}(a, t) = \mathbf{u}(\mathbf{X}(a, t), t) = \int \mathbf{u}(\mathbf{x}, t) \delta(\mathbf{x} - \mathbf{X}(a, t)) d\mathbf{x},$$

$$(25) \quad \mathbf{f}(\mathbf{x}, t) = -E_{\mathbf{x}}[\mathbf{X}(\mathbf{x}, t), t],$$

where  $E_{\mathbf{x}}$  denotes the Fréchet derivative of  $E$  with respect to  $\mathbf{X}$  (see (2)).

Note that these equations may be partitioned naturally into three groups. The Eulerian group (20)–(21) may be recognized as the Navier-Stokes equations of a viscous incompressible fluid with variable mass density  $\rho(\mathbf{x}, t)$  and

an applied or external force density  $\mathbf{F}(\mathbf{x}, t)$ . In the present context, these equations are used not only for the blood in the cardiac chambers (where  $\mathbf{F} \equiv 0$ ) but also for the muscular heart walls and the valves, where  $\mathbf{F}$  is an expression of the elasticity of the material. The Lagrangian group contains only (25), which defines the elastic properties of the valve leaflets and the muscular heart walls. A more specific example will be given below. Finally, (22)–(24) may be described as interaction equations. They define certain Eulerian quantities in terms of Lagrangian quantities and vice versa. In each case, the interaction may be expressed as an integral transformation with a  $\delta$ -function kernel.

It is easy to show from (22) and from volume conservation that  $\rho(\mathbf{x}, t)$  is constant along particle trajectories:  $\rho(\mathbf{X}(a, t), t) = m(a)/J_0(a)$ . An important special case is where  $m(a)/J_0(a) = \rho_0$ , independent of  $a$ , for it then follows that  $\rho(\mathbf{x}, t) = \rho_0$ , independent of  $\mathbf{x}$  and  $t$ . This special case is quite realistic for the heart, since cardiac muscle has nearly the same density as blood, and it is the only case for which we have actually implemented a scheme for the numerical solution of (20)–(25). The general case opens up an intriguing prospect of additional applications, however, in which a fluid interacts with an (active or passive) elastic medium of a different density. Bird flight is one example that comes to mind.

We conclude this section by discussing a special case of an elastic energy functional  $E$  that plays an important role in modeling the heart. A striking characteristic of cardiac muscle is that the heart is made of fibers and that there is a definite local fiber orientation that changes smoothly from point to point [3–6]. Let the Lagrangian parameters  $(q, r, s)$  be chosen in such a way that  $q, r = \text{constant}$  along a fiber. We assume that  $E$  is of the form

$$(26) \quad E = \int \mathcal{E} \left( \left| \frac{\partial \mathbf{X}}{\partial s} \right|; q, r, s, t \right) dq dr ds;$$

that is, the local energy density depends only on the local fiber strain, which is determined by  $|\partial \mathbf{X}/\partial s|$ , and not on the strain in the cross-fiber directions. The Fréchet derivative of  $E$  is evaluated as follows:

$$(27) \quad \begin{aligned} \lim_{\varepsilon \rightarrow 0} \frac{d}{d\varepsilon} E[\mathbf{X} + \varepsilon \mathbf{Y}, t] &= \int \mathcal{E}' \left( \left| \frac{\partial \mathbf{X}}{\partial s} \right|; q, r, s, t \right) \frac{\partial \mathbf{X}/\partial s}{|\partial \mathbf{X}/\partial s|} \cdot \frac{\partial \mathbf{Y}}{\partial s} dq dr ds \\ &= - \int \frac{\partial}{\partial s} \left[ \mathcal{E}' \left( \left| \frac{\partial \mathbf{X}}{\partial s} \right|; q, r, s, t \right) \frac{\partial \mathbf{X}/\partial s}{|\partial \mathbf{X}/\partial s|} \right] \cdot \mathbf{Y} dq dr ds. \end{aligned}$$

Here  $\mathcal{E}'$  denotes the derivative of  $\mathcal{E}$  with respect to its first argument. It follows (see (2)) that

$$(28) \quad \mathbf{f}(q, r, s, t) = \frac{\partial}{\partial s}(T\boldsymbol{\tau}),$$

where

$$(29) \quad T(q, r, s, t) = \mathcal{E}' \left( \left| \frac{\partial \mathbf{X}}{\partial s} \right| ; q, r, s, t \right),$$

$$(30) \quad \boldsymbol{\tau}(q, r, s, t) = \frac{\partial \mathbf{X} / \partial s}{|\partial \mathbf{X} / \partial s|}.$$

The vector  $\boldsymbol{\tau}$  is the unit tangent to the fibers, and the scalar  $T$  is the fiber tension in the sense that  $T \boldsymbol{\tau} dq dr$  is the force transmitted by the bundle of fibers  $dq dr$ . Equations (28)–(30) may also be derived from more elementary force-balance arguments (see [7]).

**The issue of numerical stability and the computation of the elastic force.** The numerical method that we use to solve (20)–(25) in the special case that  $\rho$  is constant and that  $E$  is given by (26) is described in detail in [7–9]. Here we discuss only one aspect of this method (see also [10]): the influence of the particular way that  $\mathbf{f}$  is computed on the stability of the computation as a whole. This issue arises because of the finite size of the time step  $\Delta t$ . Consider the step from  $t = n\Delta t$  to  $t = (n+1)\Delta t$ . (We shall refer to this as time step  $n$ .) If the force used for this time step is computed according to the straightforward recipe  $\mathbf{f} = -E_{\mathbf{x}}[\mathbf{X}(a, n\Delta t), n\Delta t]$ , then the computed solution exhibits violent instability unless  $\Delta t$  is very small. The cause of this instability is that the fiber configurations may drastically overshoot equilibrium on a single time step. This results in a reversal of sign and a large increase in the magnitude of  $\mathbf{f}$ , a situation which compounds itself from one time step to the next.

One possible cure for this well-known “stiffness” difficulty is to compute the force that is used during time step  $n$  from the unknown fiber configuration  $\mathbf{X}(a, (n+1)\Delta t)$ . This “backward-Euler” or “implicit” approach has recently been tried (on a two-dimensional model problem in which the fluid is described by the Stokes equations) in the Ph.D. thesis of C. Tu [11a,b]. This work confirms the unconditional stability of the backward-Euler method, but it also makes clear the complexity of the dense nonlinear system that must then be solved for  $\mathbf{f}$  at each time step.

Here, we describe an intermediate approach which enhances the stability of the method (without, unfortunately, making it unconditionally stable) at a more modest cost. (See [11a,b] for a comparison of all three methods.) The basic idea is to compute the force for time step  $n$  from a configuration that *approximates*  $\mathbf{X}(a, (n+1)\Delta t)$ . For this to work, however, it is crucial that the approximation include, at least crudely, the effect of  $\mathbf{f}$  at time step  $n$  on  $\mathbf{X}(a, (n+1)\Delta t)$ . The approximation that we use is obtained by considering a model problem in which the Lagrangian is again given by (3) but the constraint of incompressibility (and the fluid viscosity) are ignored. The equations of motion for this model problem are simply

$$(31) \quad -m(a) \frac{\partial^2 \mathbf{X}}{\partial t^2}(a, t) + \mathbf{f}(a, t) = 0,$$

where

$$(32) \quad \mathbf{f}(\cdot, t) = -E_{\mathbf{X}}[\mathbf{X}(\cdot, t), t]$$

as before. Equation (31) may be put in first-order form by introducing  $\mathbf{U}(a, t) = (\partial \mathbf{X} / \partial t)(a, t) = \mathbf{u}(\mathbf{X}(a, t), t)$ . Then

$$(33) \quad -m(a) \frac{\partial \mathbf{U}}{\partial t}(a, t) + \mathbf{f}(a, t) = 0,$$

$$(34) \quad \frac{\partial \mathbf{X}}{\partial t}(a, t) = \mathbf{U}(a, t).$$

Now the backward-Euler method for the system comprised of (32)–(34) is as follows:

$$(35) \quad -m(a) \frac{\mathbf{U}^{n+1}(a) - \mathbf{U}^n(a)}{\Delta t} + \mathbf{f}^{n+1}(a) = 0,$$

$$(36) \quad \frac{\mathbf{X}^{n+1}(a) - \mathbf{X}^n(a)}{\Delta t} = \mathbf{U}^{n+1}(a),$$

$$(37) \quad \mathbf{f}^{n+1}(\cdot) = -E_{\mathbf{X}}^{n+1}[\mathbf{X}^{n+1}(\cdot)],$$

where the superscripts denote the time step index, as in  $\mathbf{X}^n(a) = \mathbf{X}(a, n\Delta t)$ .

We now derive an energy inequality for the difference equations (35)–(37) which gives some insight into the good stability properties of the backward-Euler method. The inequality is derived under the hypothesis that  $E^n$  is a convex functional for every  $n$ . This means that

$$(38) \quad E^n[\mathbf{Y}] \geq E^n[\mathbf{X}] + (\mathbf{Y} - \mathbf{X}, E_{\mathbf{X}}^n[\mathbf{X}]),$$

where  $(\cdot, \cdot)$  denotes the inner product

$$(39) \quad (\mathbf{X}, \mathbf{Y}) = \int \mathbf{X}(a) \cdot \mathbf{Y}(a) da.$$

(We leave it as an exercise for the reader to show that the particular energy functional given by (26) is convex provided that  $\mathcal{E}' \geq 0$ ,  $\mathcal{E}'' \geq 0$ ,  $\mathcal{E}'(0; q, r, s, t) = \mathcal{E}''(0; q, r, s, t) = 0$ .) We apply the foregoing definition of convexity to (35)–(37) in the following way:

$$\begin{aligned} E^{n+1}[\mathbf{X}^n] &\geq E^{n+1}[\mathbf{X}^{n+1}] + (\mathbf{X}^n - \mathbf{X}^{n+1}, E_{\mathbf{X}}^{n+1}[\mathbf{X}^{n+1}]) \\ &= E^{n+1}[\mathbf{X}^{n+1}] + \Delta t(\mathbf{U}^{n+1}, \mathbf{f}^{n+1}) \\ (40) \quad &= E^{n+1}[\mathbf{X}^{n+1}] + (\mathbf{U}^{n+1}, m(\mathbf{U}^{n+1} - \mathbf{U}^n)) \\ &= E^{n+1}[\mathbf{X}^{n+1}] + \frac{1}{2}(\mathbf{U}^{n+1}, m\mathbf{U}^{n+1}) - \frac{1}{2}(\mathbf{U}^n, m\mathbf{U}^n) \\ &\quad + \frac{1}{2}((\mathbf{U}^{n+1} - \mathbf{U}^n), m(\mathbf{U}^{n+1} - \mathbf{U}^n)). \end{aligned}$$

Because the last term is positive, we may throw it away without disturbing the inequality. Rearranging the result and writing  $E^{n+1}[\mathbf{X}^n] = E^n[\mathbf{X}^n] + (E^{n+1} - E^n)[\mathbf{X}^n]$ , we get

$$(41) \quad E^{n+1}[\mathbf{X}^{n+1}] + \frac{1}{2}(\mathbf{U}^{n+1}, m\mathbf{U}^{n+1}) \leq E^n[\mathbf{X}^n] + \frac{1}{2}(\mathbf{U}^n, m\mathbf{U}^n) + (E^{n+1} - E^n)[\mathbf{X}^n].$$

For a time-independent elastic energy functional ( $E^n$  independent of  $n$ ), the inequality that we have just derived states that the total energy at time  $t = (n+1)\Delta t$  is bounded by the total energy at  $t = n\Delta t$  and hence by the total energy at  $t = 0$ . In the general case, the inequality states that the increase in the total energy at time step  $n$  is at most the increase in elastic energy that would have occurred if the system had been artificially held at  $\mathbf{X}^n$ . These reasonable bounds prevent the explosive growth in energy that may occur when  $\mathbf{f}^n$  is used in place of  $\mathbf{f}^{n+1}$  in (35).

Let us now consider the practical question of how to solve (35)–(37) for the unknowns  $\mathbf{X}^{n+1}$ ,  $\mathbf{U}^{n+1}$ ,  $\mathbf{f}^{n+1}$ . Eliminating  $\mathbf{U}^{n+1}$  and  $\mathbf{f}^{n+1}$ , we find

$$(42) \quad m(\mathbf{X}^{n+1} - \mathbf{X}^{n+1,0}) + (\Delta t)^2 E_{\mathbf{X}}^{n+1}[\mathbf{X}^{n+1}] = 0,$$

where

$$(43) \quad \mathbf{X}^{n+1,0} = \mathbf{X}^n + \Delta t \mathbf{U}^n.$$

Note that (42) is of the form

$$(44) \quad \phi_{\mathbf{X}}^{n+1}[\mathbf{X}^{n+1}] = 0,$$

where  $\phi_{\mathbf{X}}^{n+1}$  is the Fréchet derivative of the functional  $\phi^{n+1}$  defined by

$$(45) \quad \phi^{n+1}[\mathbf{X}] = \frac{1}{2}(\mathbf{X} - \mathbf{X}^{n+1,0}, m(\mathbf{X} - \mathbf{X}^{n+1,0})) + (\Delta t)^2 E^{n+1}[\mathbf{X}].$$

Thus, the solutions of (42) are the stationary points of  $\phi^{n+1}$ . Among the different types of stationary points that one might consider, it is intuitively clear that minima are best from the standpoint of stability. Moreover,  $\phi^{n+1}$  is guaranteed to have at least one minimum provided that  $m(a)$  is bounded from below by some positive constant  $m_0$ , that  $E^{n+1}$  is bounded from below (say) by zero, and that  $E^{n+1}$  is continuous. Thus, one may use any algorithm that searches for a local minimum of  $\phi^{n+1}$  to find a solution  $\mathbf{X}^{n+1}$  of (42). We use algorithms based on Newton's method for this task.

Of course, the scheme given by (35)–(37) solves only the model problem in which incompressibility and viscosity are ignored. How can we use it in the solution of the full system (20)–(25)? What we actually do is described as follows. At each time step, we begin by solving difference equations of the same form as (35)–(37) but with the unknowns  $\mathbf{X}^{n+1}$ ,  $\mathbf{U}^{n+1}$ ,  $\mathbf{f}^{n+1}$  replaced by  $\mathbf{X}^{n+1,*}$ ,  $\mathbf{U}^{n+1,*}$ ,  $\mathbf{f}^{n+1,*}$ , the notation being changed because these quantities no longer have the status of being the final results of the time step. No further use is made of  $\mathbf{X}^{n+1,*}$  or  $\mathbf{U}^{n+1,*}$ , but the force  $\mathbf{f}^{n+1,*}$  is then used as input to the algorithm that updates the solution of (20)–(24). For details of this algorithm, see [7].

The method of computing  $\mathbf{f}$  that we have just described differs from our previously published work by the inclusion of the nonconstant factor  $m(a)$ . In certain problems  $m(a)$  may be given, but in others it may have to be estimated from the given mass density of the material and the spatial distribution of the Lagrangian marker points.

To discuss the latter situation, we must say a little about the spatial discretization of the equations. The Eulerian variable  $\mathbf{x}$  is represented by a regular cubic lattice of points  $\mathbf{x}_j = \mathbf{j}h$ , where  $h$  is the lattice spacing. The Lagrangian parameter  $a$  is represented by some finite collection of values  $a_k$  each with an associated "volume" in parameter space  $\Delta a$ . We let  $m_k = m(a_k)$  and  $\mathbf{X}_k^n = \mathbf{X}(a_k, n\Delta t)$ . The product  $m_k \Delta a$  is the mass associated with material element  $k$ . Note that the points  $\mathbf{X}_k^n$  need not coincide with any of the lattice points  $\mathbf{x}_j$ . The connection between the discrete Lagrangian and Eulerian representations is made with the help of a spread-out version of the  $\delta$ -function, which we denote  $\delta_h$ . For the specific definition of  $\delta_h$  and for a discussion of its properties, see [12, 13]. This function is used in the discretization of the interaction equations (22)–(24).

We now turn to the task of estimating  $m_k$  in cases where  $m(a)$  is not given. Suppose, for example, that  $\rho(\mathbf{x}, t) = \rho_0$  as in the heart problem. Evaluating this at  $\mathbf{x} = \mathbf{X}(a, t)$  we get

$$\begin{aligned} \rho_0 &= \rho(\mathbf{X}(a, t), t) = \int \rho(\mathbf{x}, t) \delta(\mathbf{x} - \mathbf{X}(a, t)) d\mathbf{x} \\ (46) \quad &= \iint m(a') \delta(\mathbf{x} - \mathbf{X}(a', t)) \delta(\mathbf{x} - \mathbf{X}(a, t)) d\mathbf{x} da', \end{aligned}$$

where we have used the definition of the  $\delta$ -function and also (16), which defines  $\rho$  in terms of  $m$ . Clearly we may replace  $m(a')$  by  $m(a)$  in (46). [To see this, do the  $\mathbf{x}$  integral first to obtain  $\rho_0 = \int m(a') \delta(\mathbf{X}(a', t) - \mathbf{X}(a, t)) da'$  and use the invertibility of  $\mathbf{X}(\cdot, t)$ .] Thus we may write

$$(47) \quad m(a) = \frac{\rho_0}{\iint \delta(\mathbf{x} - \mathbf{X}(a', t)) \delta(\mathbf{x} - \mathbf{X}(a, t)) d\mathbf{x} da'}.$$

This is equivalent to  $m(a) = \rho_0 J_0(a)$ , but (47) is more useful for our purpose, which is to estimate  $m(a)$  solely from the given distribution of Lagrangian markers at any particular time. To do this, we discretize (47) as follows:

$$(48) \quad m_k = \frac{\rho_0}{\sum_l \sum_j \delta_h(\mathbf{x}_j - \mathbf{X}_l) \delta_h(\mathbf{x}_j - \mathbf{X}_k) h^3 \Delta a}$$

or

$$(49) \quad m_k \Delta a = \frac{\rho_0}{\sum_l A_{kl}},$$

where

$$(50) \quad A_{kl} = \sum_j \delta_h(\mathbf{x}_j - \mathbf{X}_l) \delta_h(\mathbf{x}_j - \mathbf{X}_k) h^3.$$

The formula for  $m_k$  that we have just derived is new. In previous work [7, 8, 10–13, 17–19] we used the approximation to (49) given by

$$(51) \quad m_k \Delta a \approx \frac{\rho_0}{A_{kk}} = \left(\frac{8}{3}\right)^3 \rho_0 h^3,$$

which seemed reasonable because our particular choice of  $\delta_h$  is such that  $0 \leq A_{kl} \leq A_{kk} = (\frac{3}{8h})^3$  and because most of the off-diagonal terms of  $A$  are zero. In fact, (49) reduces to (51) when the Lagrangian markers are so far apart that the support of their  $\delta_h$ -functions do not overlap. We were forced to reconsider the use of this approximation, however, when we constructed a heart model in which large numbers of Lagrangian markers were crowded close together in a particular plane, which is the plane of valve rings. A violent instability then appeared which was confined to the plane of the rings and which gave a strong visual hint that the high local density of Lagrangian markers was the culprit. This instability disappeared when we replaced (51) by (49). Intuitively, this is because (49) takes account of the high local number density of Lagrangian markers and assigns an appropriately small mass to each of them.

**Applications (cardiac fluid dynamics).** A two-dimensional version of the above-described computational method has been used in a series of studies concerning the mitral valve, which is the inflow valve to the left ventricle of the heart. Some of these studies have involved the natural mitral valve, while others have been aimed at the improved design of mechanical valves for the replacement of diseased mitral valves.

In the case of the natural mitral valve, the computational method has been used to confirm and refine Leonardo da Vinci's theory [14] concerning the role of the valve-generated vorticity in efficient valve closure [12, 15]. It has also been used to determine the optimal time delay between the contraction of the atrium and that of the ventricle [16]. Simulation of a disease process (prolapse of the mitral valve) has been achieved by adjusting the parameters of the model heart [17].

Parametric studies aimed at optimizing the design of prosthetic mitral valves have also been performed. In particular, the optimal pivot point and curvature have been determined for pivoting single-disc valves [18] and for butterfly leaflet valves [19]. The design criteria used in these studies are intended to create valves with good pressure-flow characteristics and with a low propensity for stagnation and thrombosis.

Our recent efforts have been dedicated to the three-dimensional implementation of the computational method [7-9] and to the construction of a three-dimensional fiber-based computer model of the heart for use with that method. The method is operational, and the model is near completion. Some preliminary results are shown in Figure 1.

**The equations of Newtonian molecular dynamics.** The goal of molecular dynamics is to describe the motion of a single molecule (or of a collection of molecules) at the *atomic* level of description. Thus, the dynamical variables are the coordinates and momenta of the atoms; there is no attempt to look inside the atoms at their electronic (or nuclear) structure. Note that



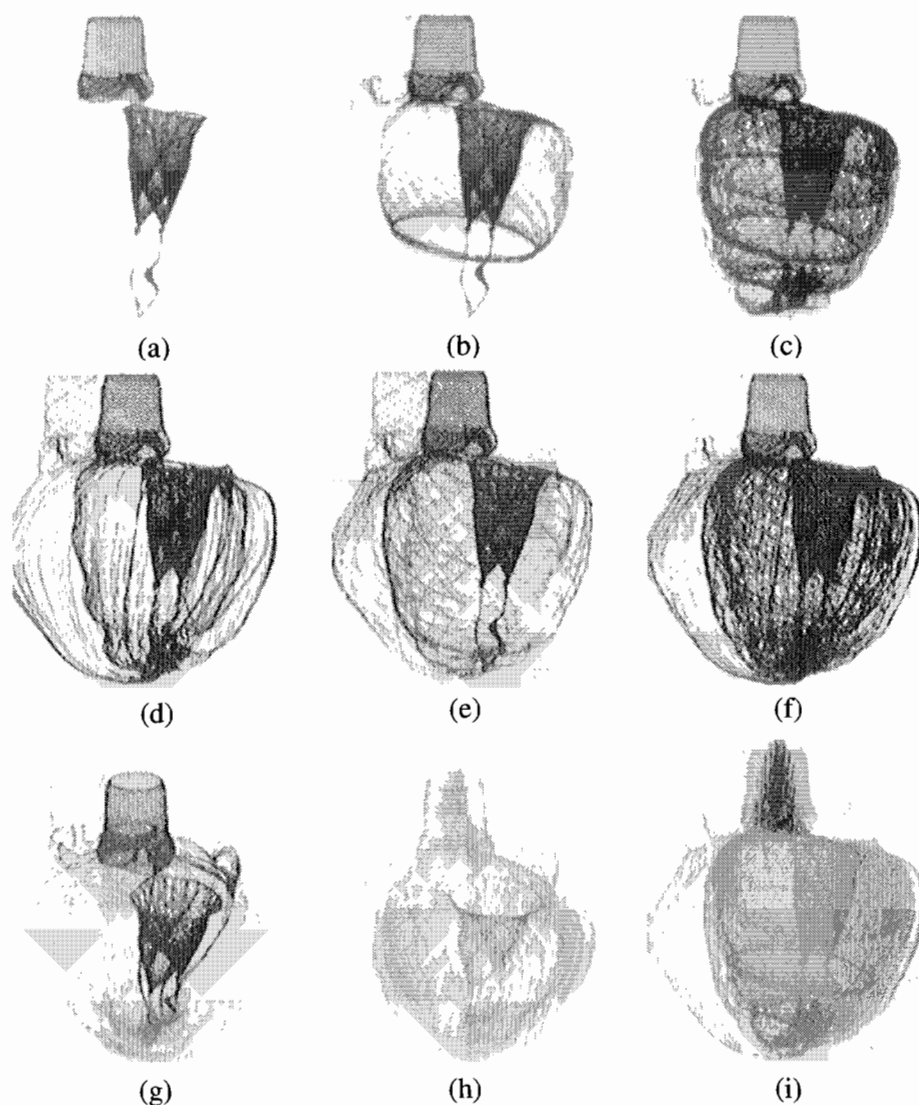


FIGURE 1. A computer model of the heart at a single instant during ventricular ejection.

- (a) Detail of the mitral and aortic valves.
- (b) The pulmonic valve and a (double-sheeted) layer of the left ventricular wall have been added.
- (c) Additional layers of the left ventricular wall. Each layer is double-sheeted with the fibers making a smooth transition from one sheet to the other at their common edge. The different layers are nested.
- (d) Vertical fibers that surround the heart as a whole, penetrate at the apex, and form the inner lining of the left ventricle. Note the right ventricle (left in figure). The tricuspid valve is obscured by the mitral valve in this view.
- (e) A layer of fibers that wraps (figure-eight fashion) around both ventricles.
- (f) Complete model of the left and right ventricles connected to the aorta and the pulmonary artery.
- (g) A perspective view of the complete model with all four valves visible.
- (h) Velocity vectors emanating from the plane of the aortic valve ring (viewpoint from slightly above the plane).
- (i) The same velocity vectors viewed from within the plane of the aortic valve ring.

the atomic level is perhaps the smallest at which it makes sense to think in classical (as opposed to quantum-mechanical) terms. Even here, however, there are certain quantum effects that should not be ignored. An important theme of the work that we shall describe is how to include such effects in an essentially classical description.

The classical equations of motion of a molecular system may be written in the form

$$(52) \quad M \frac{d^2 X}{dt^2} + E'(X) = 0,$$

where  $X$  is a vector with  $N = 3n_a$  components ( $n_a = \#$  of atoms) that lists the Cartesian coordinates of the atoms,  $M$  is a diagonal matrix of order  $3n_a$  with diagonal entries equal to the masses of the atoms (each mass repeated three times),  $E(X)$  is the potential energy of the molecular system as a function of its configuration  $X$ , and  $E'$  denotes the gradient of  $E$ .

The mass matrix  $M$  and the potential energy function  $E$  define the system. In principle (according to the Born-Oppenheimer approximation [20]), the value of  $E$  at each configuration  $X$  may be found by solving the Schrödinger equation of the electrons with the atomic nuclei held fixed in the configuration  $X$ . In practice, the function  $E(X)$  is obtained by a much less systematic procedure that combines guesswork, empirical data, and theoretical considerations. In this "empirical" approach, there are two types of terms that appear in  $E(X)$ : bonded and nonbonded. In the first category are energies associated with bond stretching, with distortions in bond angle, and with rotations about individual bonds. In the second category are the Coulomb and Van der Waals energies; these involve sums over all atom pairs.

A typical energy function is constructed as follows. Let the atoms be numbered  $i = 1, \dots, n_a$ , and let  $B_{ij} = 1$  if  $i \neq j$  and there is a chemical bond connecting atom  $i$  to atom  $j$ , with  $B_{ij} = 0$  otherwise. Then let

$$(53) \quad S_2 = \{(i, j) : B_{ij} = 1\},$$

$$(54) \quad S_3 = \{(i, j, k) : (i, j) \in S_2, (j, k) \in S_2, i \neq k\},$$

$$(55) \quad S_4 = \{(i, j, k, l) : (i, j, k) \in S_3, (j, k, l) \in S_3\}.$$

In the foregoing definitions,  $(i, j)$ ,  $(i, j, k)$ , and  $(i, j, k, l)$  denote *ordered* pairs, triples, and quadruples of integers selected from  $\{1, \dots, n_a\}$ . Also let

$$(56) \quad \mathbf{X}_{ij} = \mathbf{X}_j - \mathbf{X}_i,$$

$$(57) \quad \mathbf{X}_{ijk} = \mathbf{X}_{ij} \times \mathbf{X}_{jk},$$

and introduce the notation  $\theta(\mathbf{a}, \mathbf{b})$  for the angle between the two vectors  $\mathbf{a}$  and  $\mathbf{b}$ . (That is,  $\cos \theta(\mathbf{a}, \mathbf{b}) = (\mathbf{a} \cdot \mathbf{b}) / (|\mathbf{a}| |\mathbf{b}|)$  and  $0 \leq \theta(\mathbf{a}, \mathbf{b}) \leq \pi$ .)

For  $(i, j) \in S_2$ ,  $|\mathbf{X}_{ij}|$  is the length of the bond connecting atom  $i$  and atom  $j$ . For  $(i, j, k) \in S_3$ ,  $\theta(\mathbf{X}_{ji}, \mathbf{X}_{jk})$  is a bond angle with vertex at atom

$j$ . For  $(i, j, k, l) \in S_4$ ,  $\theta(\mathbf{X}_{ijk}, \mathbf{X}_{jkl})$  is the dihedral angle about the bond connecting atom  $j$  and atom  $k$ .

With the notation introduced above, the potential energy  $E$  of the molecular system may be defined as follows:

$$\begin{aligned}
 2E(\mathbf{X}_1 \cdot s\mathbf{X}_n) = & \sum_{(i,j) \in S_2} f_{ij}(|\mathbf{X}_{ij}|) + \sum_{(i,j,k) \in S_3} f_{ijk}(\theta(\mathbf{X}_{ji}, \mathbf{X}_{jk})) \\
 (58) \quad & + \sum_{(i,j,k,l) \in S_4} f_{ijkl}(\theta(\mathbf{X}_{ijk}, \mathbf{X}_{jkl})) \\
 & + \sum_{(i,j): i \neq j} \left\{ \frac{q_i q_j}{4\pi\epsilon |\mathbf{X}_{ij}|} + \left( \frac{a_{ij}}{|\mathbf{X}_{ij}|^{12}} - \frac{b_{ij}}{|\mathbf{X}_{ij}|^6} \right) \right\}.
 \end{aligned}$$

The factor 2 on the left-hand side of (58) reflects the fact that each term on the right appears exactly twice. If  $(i, j, k) \in S_3$ , for example, then  $(k, j, i) \in S_3$ , the functions  $f_{ijk}$  and  $f_{kji}$  are the same, and  $\theta(\mathbf{X}_{ji}, \mathbf{X}_{jk}) = \theta(\mathbf{X}_{jk}, \mathbf{X}_{ji})$ . The functions  $f_{ij}$ ,  $f_{ijk}$ , and  $f_{ijkl}$  which appear in (58) specify respectively the energies associated with bond length, bond angle, and dihedral angle. One might, for example, set  $f_{ij}(r) = \frac{1}{2}K_{ij}(r - (r_0)_{ij})^2$ , where the parameters  $K_{ij}$  and  $(r_0)_{ij}$  would be determined by the atom types (C, O, N, H,  $\cdot s$ ) of the two atoms  $i$  and  $j$  and by the type of bond (single, double, aromatic, ...) that connects these two atoms. The last summation in (58) describes the nonbonded interactions. In these terms the  $q_i$  are the partial charges,  $\epsilon$  is the dielectric constant (see below), and the coefficients  $a_{ij}$ ,  $b_{ij}$  describe the Van der Waals interaction between atoms  $i$  and  $j$ . (These coefficients are determined by the atom types of the two atoms in question.) The appropriate dielectric constant is that of free space if all atoms in the system are modeled explicitly. Otherwise it may be necessary to adjust the dielectric constant to account for the screening effect of charges or dipoles that are not explicitly modeled.

**The backward-Euler Langevin method.** For systems with large numbers of atoms, the numerical solution of (52) is plagued by the same "stiffness" difficulty that arises in the case of cardiac fluid dynamics, as described above. Here we see the problem in its pure form without the complications introduced by the fluid. In fact, the molecular dynamics problem is of precisely the same form as the model problem that was used for the computation of the elastic force in the cardiac case (see (31)–(32)).

In the molecular case, however, the wide disparity of time scales that is the fundamental source of the stiffness difficulty mentioned above has additional physical significance. The fastest time scales in the classical problem are not only awkward to compute, they are actually filtered out by quantum-mechanical effects. This is because a mode with natural frequency  $\omega$  has a high probability of being found in its ground state when  $\hbar\omega > kT$ , where  $\hbar$  is Planck's constant divided by  $2\pi$ ,  $k$  is Boltzmann's constant, and  $T$  is

the temperature. This has the important consequence of limiting the number of modes that contribute to the specific heat of a substance at any given temperature [21]. Thus, a purely classical computation will overestimate the specific heat, sometimes drastically.

These considerations suggest the use of the backward-Euler method for the numerical solution of (52), both because that method makes possible stable computations with large time steps and also because it filters out high-frequency modes in a way that is reminiscent of the filtering provided by quantum mechanics. A fundamental difficulty that immediately comes to mind, however, is the dissipative character of the backward-Euler method as expressed, for example, by the energy inequality (41). A computational method in which the energy continually runs down does not seem to be a reasonable model of a molecular system at a given temperature. The way out of this difficulty is to excite the system continually with a random force that simulates the interaction with a thermal reservoir. This makes it possible to establish and maintain a prescribed temperature  $T$  despite the dissipation introduced by the backward-Euler method. We call this combination the "backward-Euler/Langevin method," since it amounts to a backward-Euler discretization of the Langevin equations of motion.

An important feature of the backward-Euler/Langevin method is the choice of parameters. The particular choice that we shall describe results in a cutoff frequency  $\omega_c$  that is independent of the time step  $\Delta t$  (provided that  $\omega_c \Delta t$  is small). By setting  $\omega_c = kT/\hbar$ , one can simulate (in an otherwise classical context) the quantum-mechanical suppression of high-frequency molecular motions.

The classical Langevin equations of motion are a modification of (52) that is obtained by imposing both a random force and a frictional force as follows:

$$(59) \quad M \left( \frac{d^2 X}{dt^2} + \gamma \frac{dX}{dt} \right) + E'(X) = R(t).$$

Here  $\gamma$  is a parameter with units of reciprocal time that is called the "collision frequency," and  $R(t)$  is a vector-valued stationary stochastic process with the following first and second moments:

$$(60) \quad \langle R(t) \rangle = 0,$$

$$(61) \quad \langle R(t)(R(t'))^T \rangle = 2kT\gamma\delta(t-t')M.$$

Here  $\langle \rangle$  denotes the ensemble average and the superscript  $T$  denotes the transpose of a matrix. (We regard  $R$  as a column vector, so  $R^T$  is a row vector.) Note the appearance of the collision frequency  $\gamma$  not only as coefficient of the friction term in (59) but also in the covariance of the random force, (62).

Discretizing the foregoing equations according to the backward-Euler method, we obtain the discrete-time stationary stochastic process that is im-

plicitly defined as

$$(62) \quad (1 + \gamma\Delta t)M(X^{n+1} - X^{n+1,0}) + (\Delta t)^2 E'(X^{n+1}) = 0,$$

where

$$(63) \quad X^{n+1,0} = X^n + \frac{(X^n - X^{n-1}) + (\Delta t)^2 M^{-1} R^{n+1}}{1 + \gamma\Delta t},$$

$$(64) \quad \langle R^n \rangle = 0,$$

$$(65) \quad \langle R^n (R^{n'})^T \rangle = 2kT\gamma \frac{\delta_{nn'}}{\Delta t} M.$$

As in the cardiac case, one can compute  $X^{n+1}$  by minimizing the function

$$(66) \quad \phi^{n+1}(X) = \frac{1}{2}(X - X^{n+1,0})^T M(X - X^{n+1,0}) + (\Delta t)^2 E(X).$$

A reasonable initial guess for this minimization is  $X^{n+1,0}$ . Equations (62)–(65) define the backward-Euler/Langevin method [22–24].

**The choice of parameters.** The computational method that we have just described has three adjustable parameters:  $T$ ,  $\gamma$ , and  $\Delta t$ . It turns out that the ratio  $\gamma/\Delta t$  is an important quantity which we shall call  $\omega_c^2$ . We have studied the continuum limit of the backward-Euler/Langevin method as  $\Delta t \rightarrow 0$  with  $\omega_c$  fixed [22, 24]. Note that this limit is *not* described by the continuous Langevin equations of motion, since  $\gamma \rightarrow 0$  along with  $\Delta t$ . In the special case of a system of coupled harmonic oscillators with frequencies  $\omega_j$ , we have shown that the mean energy of the system is given (in the above-described limit) by

$$(67) \quad \langle E \rangle = \sum_j \frac{kT}{1 + (\omega_j/\omega_c)^2}.$$

According to (67), the modes for which  $\omega_j \ll \omega_c$  each contribute nearly  $kT$  to the total mean energy. Such modes, then, behave as in the classical statistical mechanics of a system at temperature  $T$ ; this substantiates the interpretation of the parameter  $T$  as the temperature of the system, at least for the low-frequency modes. The modes for which  $\omega_j \gg \omega_c$ , on the other hand, each contribute  $\ll kT$  to the total mean energy. Since the transition between the two regimes occurs near  $\omega_c$ , we refer to  $\omega_c$  as the *cutoff* frequency. We call these modes for which  $\omega_j < \omega_c$  the *active* modes and those for which  $\omega_j > \omega_c$  the *inactive* modes.

In practice one cannot set  $\Delta t = 0$ , but to resolve all of the active modes it is only necessary to pick  $\Delta t$  so that  $\omega_c \Delta t \ll 1$ . There is no point in resolving the inactive modes, since they are repressed in any case. Depending on the choice of  $\omega_c$ , the restriction  $\omega_c \Delta t \ll 1$  may be much less severe than the restriction  $(\max\{\omega_j\})\Delta t \ll 1$ , which would have to be imposed for stability reasons if an explicit scheme were used.

In summary, the parameters should be selected as follows. First, set  $T$  equal to the (absolute) temperature of interest. Next, choose a cutoff frequency  $\omega_c$  that marks the upper limit of the frequency range of interest. Finally select  $\Delta t$  so that  $\Delta t \omega_c \ll 1$ , and set  $\gamma = \omega_c^2 \Delta t$ .

We conclude this section with a qualitative discussion of what happens as the parameter  $\Delta t$  is varied within the limits implied by the restriction  $\Delta t \omega_c \ll 1$ . If  $\Delta t$  is made smaller,  $\gamma$  also becomes smaller, since  $\gamma = \omega_c^2 \Delta t$ . The effect of this is to reduce the influence of both the friction term and the random force on the backward-Euler/Langevin process. This is evident for the friction term but less so for the random force, since the covariance of the random force has amplitude  $\gamma/\Delta t = \omega_c^2$ , independent of  $\Delta t$ . Nevertheless, the continuum limit of a discrete-time random process with first and second moments given by  $\langle r^n \rangle = 0$  and  $\langle r^n r^m \rangle = \delta_{nm}$  is the zero process. The reason is that there is so much opportunity for cancellation as  $\Delta t \rightarrow 0$ . In summary, a reduction in  $\Delta t$  weakens the Langevin terms that provide coupling between the molecular system and the thermal reservoir which establishes and maintains the target temperature  $T$ . This has no effect on the equilibrium properties of the system, but it does reduce the rate at which such equilibrium is achieved, and it also increases the duration of a system trajectory that one must examine in order to obtain reliable estimates of equilibrium quantities. If  $\Delta t$  is reduced by a factor of 2, twice as much time (and hence a record containing *four* times as many time steps) is required to obtain averages with the same reliability as before.

**A connection with quantum statistical mechanics.** The suppression of fast modes that occurs in the backward-Euler/Langevin method is reminiscent of a similar suppression that occurs in quantum statistical mechanics. This physical effect plays an important role in the specific heat of diatomic gases, for example. At high temperature (e.g., 3000°K) such gases have a mean internal energy of  $\frac{7}{2}kT$  per molecule. This is as predicted by classical statistical mechanics (provided that we regard the molecule as two point masses connected by a spring and ignore the electronic degrees of freedom!) The factor 7 is arrived at by counting three translational and two rotational degrees of freedom together with one vibrational degree of freedom (which counts twice because of the kinetic and potential vibrational energies). As the temperature is lowered, however, a new phenomenon appears which is completely inexplicable from a classical point of view [21]. First, the vibrational energy drops out of the picture and the molecule acts like two point masses connected by a rigid rod: its mean internal energy in this regime is  $\frac{5}{2}kT$ . This is the typical situation near room temperature. At still lower temperature (e.g., 30°K) the rotation also drops out and the molecule acts like a monatomic gas: its mean internal energy is then only  $\frac{3}{2}kT$ .

These effects are important not only because they influence the mean energy that a molecule will carry at any given temperature, but also because they

imply that the real motions of the atoms in a complex molecule at ordinary temperatures are much more highly *correlated* than one would expect on the basis of classical statistical mechanics. This simplifying feature should be exploited in molecular dynamics computations.

The analogy between the suppression of fast modes in quantum statistical mechanics and the corresponding suppression in the backward-Euler/Langevin method can be made somewhat quantitative by making a particular choice of the cutoff frequency:  $\omega_c = kT/\hbar$ . We call this the quantum-mechanical cutoff frequency. With this choice the formula for the mean energy of a collection of coupled harmonic oscillators modeled by backward-Euler/Langevin method (equation (67)) becomes

$$(68) \quad \langle E \rangle_{\text{BEL}} = \sum_j \frac{kT}{1 + (\hbar\omega_j/kT)^2}$$

(in which the subscript BEL denotes backward-Euler/Langevin). This should be compared with the quantum-mechanical formula derived by Planck [25]

$$(69) \quad \langle E \rangle_{\text{QM}} = \sum_j = \frac{kT(\hbar\omega_j/kT)}{\exp(\hbar\omega_j/kT) - 1}$$

(where QM denotes quantum-mechanical). Note that (69) does not include the zero-point energy  $\sum_j \frac{1}{2}(\hbar\omega_j)$  which would be present even at absolute zero temperature. Thus, (69) gives the energy required to raise the temperature of the system of coupled harmonic oscillators from absolute zero to the temperature  $T$ . It is also of interest to compare  $\langle E \rangle_{\text{BEL}}$  with the classical expression for the mean energy of a system of coupled harmonic oscillators. The classical result is

$$(70) \quad \langle E \rangle_{\text{CM}} = \sum_j (kT)$$

(in which CM denotes classical mechanics). This formula is easily derived from (69) (or for that matter from (68)) by letting  $\hbar \rightarrow 0$ .

The three formulas given by (68)–(70) are all of the form

$$(71) \quad \langle E \rangle = \sum_j kT f\left(\frac{\hbar\omega_j}{kT}\right),$$

but the function  $f$  is different in each case:

$$(72) \quad f_{\text{BEL}}(\theta) = \frac{1}{1 + \theta^2},$$

$$(73) \quad f_{\text{QM}}(\theta) = \frac{\theta}{\exp(\theta) - 1},$$

$$(74) \quad f_{\text{CM}}(\theta) = 1.$$

These three functions are plotted in Figure 2 (on the next page). Note the close resemblance of  $f_{\text{BEL}}$  and  $f_{\text{QM}}$  and also strong disagreement between

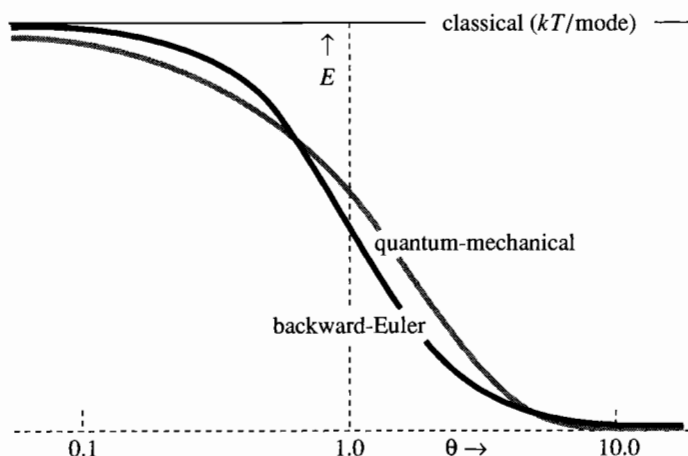


FIGURE 2. Comparison of the backward-Euler/Langevin method with classical mechanics and with quantum mechanics. All three graphs refer to a coupled system of harmonic oscillators at temperature  $T$ . The ordinate  $E$  is the mean energy per vibrational mode, and the abscissa  $\theta$  is the scaled natural frequency of the mode in question ( $\theta = \hbar\omega/kT$ ) plotted on a logarithmic scale. In the classical case the mean energy is the same for all modes, while in both the quantum case and in the backward-Euler/Langevin case the mean energy per mode decreases dramatically as  $\omega$  crosses the frequency given by  $kT/\hbar$ . For further details, see [22].

$f_{\text{BEL}}$  and  $f_{\text{CM}}$ . This is quite astonishing when one considers the classical character of the backward-Euler/Langevin method.

Of course, the harmonic-oscillator problem is very special, and so one should consider other (nonlinear) examples. This has been done computationally for a diatomic molecule with bond energy given by a Morse potential [23]. As in quantum statistical mechanics, the specific heat computed by the backward-Euler/Langevin method is roughly  $\frac{5}{2}k$  at moderate temperature and rises to  $\frac{7}{2}k$  at high temperatures. The transition occurs as the vibrational motion comes into play. (Computations were not performed at low enough temperatures to see the  $\frac{3}{2}k$  specific heat that should be observed when both vibration and rotation are inactive, but this was remedied in the analytic study of a rigid rotator, described below.)

In addition to this computational study of a diatomic molecule, we have also worked out an analytic framework for studying the equilibrium properties of a general Hamiltonian system modeled by the backward-Euler/



Langevin method [24]. This framework has been applied to the case of a rigid rotator, for which we have again obtained results which are qualitatively consistent with quantum statistical mechanics (though the agreement is less good than in the harmonic-oscillator case) and qualitatively *inconsistent* with classical statistical mechanics.

**Applications (molecular dynamics).** The principal application of the backward-Euler/Langevin method will be the computational prediction of the three-dimensional structure of large biological molecules such as proteins and nucleic acids. (For an overview of this field, see [26].) A dynamic method is needed for this apparently static problem for the following reason. Large biological molecules have energy functions with multiple local minima. For a complete description of molecular structure it is not sufficient to find the global minimum (even if one knew how to do so). Rather, one needs to construct a catalog of low-energy minima, to determine the fraction of time that the molecule spends near each of them, and to evaluate the rate constants for the spontaneous transitions from one local minimum to another. The most straightforward way to collect this information is to simulate the physical process by which the molecule changes its configuration over time and to compile statistics concerning the molecular trajectory in configuration space. For such statistics to be meaningful, it is necessary that the simulated trajectory be of long enough duration to be representative. It is our hope that the backward-Euler/Langevin method will make this possible. For preliminary results, see [22, 27].

**Conclusions.** Hearts and biological macromolecules have more in common than might at first appear. Each can be viewed as an elastic structure, the dynamics of which is governed by an internal energy function. Each has many degrees of freedom and is capable of motions on widely different temporal (and spatial) scales. These multiple time scales require the use of implicit numerical methods, which can be formulated as optimization problems to be solved at each time step.

These similarities notwithstanding, there are also important differences between the two problems. In the cardiac case, the elastic energy function is time-dependent, and the elastic structure interacts with a viscous, incompressible fluid. In the molecular case, thermal fluctuations are important and quantum effects cannot be ignored.

As the present paper has shown, one can exploit the similarities between the two problems without overlooking the differences. This is done by using a common computational framework based on the backward-Euler method. In the cardiac case, however, this framework is coupled to the numerical solution of the Navier-Stokes equations, while in the molecular case, random forces are used to simulate thermal collisions, and parameters are chosen so that the backward-Euler dissipation simulates the quantum-mechanical suppression of fast molecular motions.

**Acknowledgment.** The cardiac research described in this paper is joint work with David M. McQueen; it is supported by the National Institutes of Health under research grant HL17859. The molecular-dynamics research is joint work with Tamar Schlick; it is supported by the National Science Foundation under research grant ASC8705589. The author is indebted to the mathematical community for providing a warm and supportive home for this interdisciplinary research.

## REFERENCES

1. D. G. Ebin and R. A. Saxton, *The initial-value problem for elastodynamics of incompressible bodies*, Arch. Rational Mech. Anal. **94** (1986), 15–38.
2. —, *The equations of incompressible elasticity*, Contemp. Math. **60** (1987), 25–34.
3. D. D. Streeter, Jr., H. M. Spotnitz, D. P. Patel, J. Ross, Jr., and E. H. Sonnenblick, *Fiber orientation in the canine left ventricle during diastole and systole*, Circ. Res. **24** (1969), 339–347.
4. D. D. Streeter, Jr., W. E. Powers, M. A. Ross, and F. Torrent-Guasp, *Three-dimensional fiber orientation in the mammalian left ventricular wall*, Cardiovascular System Dynamics (J. Baan, A. Noordergraaf, and J. Raines, eds.), M.I.T. Press, Cambridge, MA, 1978, pp. 73–84.
5. C. E. Thomas, *The muscular fiber architecture of the ventricles of hog and dog hearts*, Amer. J. Anatomy **101** (1957), 17–57.
6. C. S. Peskin, *Fiber-architecture of the left ventricular wall: an asymptotic analysis*, Comm. Pure Appl. Math. **42** (1989), 79–113.
7. C. S. Peskin and D. M. McQueen, *A three-dimensional computational method for blood flow in the heart. (I) Immersed elastic fibers in a viscous incompressible fluid*, J. Comput. Phys. **81** (1989), 372–405.
8. D. M. McQueen and C. S. Peskin, *A three-dimensional computational method for blood flow in the heart. (II) Contractile fibers*, J. Comput. Phys. **82** (1989), 289–297.
9. S. Greenberg, D. M. McQueen, and C. S. Peskin, *Three-dimensional fluid dynamics in a two-dimensional amount of central memory*, Wave Motion: Theory, Modeling, and Computation (Proc. Conf. in Honor of the 60th Birthday of Peter D. Lax) (A. J. Chorin, ed.), Springer-Verlag, New York, 1987, pp. 85–146.
10. C. S. Peskin and D. M. McQueen, *Modeling prosthetic heart valves for numerical analysis of blood flow in the heart*, J. Comput. Phys. **37** (1980), 113–132.
- 11a. C. Tu, *A study of stability in the computation of flows with moving immersed boundaries: a comparison of three methods*, thesis, New York University, 1989.
- 11b. C. Tu and C. S. Peskin, *Stability and instability in the computation of flows with moving immersed boundaries*, SISSC (to appear).
12. C. S. Peskin, *Flow patterns around heart valves: A digital computer method for solving the equations of motion*, thesis, Albert Einstein College of Medicine, July, 1972. (University Microfilms #72–30, 378. 211pp.)
13. —, *Numerical analysis of blood flow in the heart*, J. Comput. Phys. **25** (1977), 220–252.
14. Leonardo da Vinci, in *Leonardo da Vinci on the human body* (C. C. O'Malley and J. B. de C. M. Saunders, eds.), Henry Schuman, New York, 1952, pp. 258–275.
15. C. S. Peskin, *The fluid dynamics of heart valves: experimental, theoretical, and computational methods*, Ann. Rev. Fluid Mech. **14** (1982), 235–259.
16. J. S. Meisner, D. M. McQueen, Y. Ishida, H. O. Vetter, U. Bortolotti, J. A. Strom, R. W. M. Frater, C. S. Peskin, and E. L. Yellin, *Effects of timing of atrial systole on LV filling and mitral valve closure: computer and dog studies*, Amer. J. Physiol. **249** (1985), H604–H619.
17. D. M. McQueen, C. S. Peskin, and E. L. Yellin, *Fluid dynamics of the mitral valve: physiological aspects of a mathematical model*, Amer. J. Physiol. **242** (1982), H1095–H1110.

18. D. M. McQueen and C. S. Peskin, *Computer-assisted design of pivoting disc prosthetic mitral valves*, J. Thorac. Cardiovasc. Surg. **86** (1983), 126–135.
19. ———, *Computer-assisted design of butterfly bileaflet valves for the mitral position*, Scand. J. Thor. Cardiovasc. Surg. **19** (1985), 139–148.
20. M. Born and J. R. Oppenheimer, Ann. D. Phys. **84** (1927), 457–484.
21. R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman lectures on physics*, Addison-Wesley, Reading, MA, 1963, pp. 40.7–40.10.
22. C. S. Peskin and T. Schlick, *Molecular dynamics by the backward-Euler method*, Comm. Pure Appl. Math. **42** (1989), 1001–1031.
23. T. Schlick and C. S. Peskin, *Can classical equations simulate quantum-mechanical behavior? A molecular dynamics investigation of a diatomic molecule with a Morse potential*, Comm. Pure Appl. Math. **42** (1989), 1141–1163.
24. C. S. Peskin, *Analysis of the backward-Euler/Langevin method for molecular dynamics*, Comm. Pure. Appl. Math. **43** (1990), 599–645.
25. R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman lectures on physics*, Addison-Wesley, Reading, MA, 1963, pp. 41.6–41.7.
26. J. A. McCammon and S. C. Harvey, *Dynamics of proteins and nucleic acids*, Cambridge Univ. Press, 1987.
27. T. Schlick, B. E. Hingerty, C. S. Peskin, M. L. Overton, and S. Broyde, *Search strategies, minimization algorithms, and molecular dynamics simulations for exploring conformational spaces of nucleic acids*, Theoretical Biochemistry and Molecular Biophysics (D. L. Beveridge and R. Lavery, eds.), Adenine Press, 1990, pp. 39–58.

COURANT INSTITUTE OF MATHEMATICAL SCIENCES, NEW YORK UNIVERSITY, 251 MERCER STREET, NEW YORK, NEW YORK 10012

## Bounds, Quadratic Differentials, and Renormalization Conjectures

DENNIS SULLIVAN

**Introduction.** Consider mixing a deck of  $n$  cards by shuffling as usual after turning over one of the stacks. The resulting permutations are building blocks of the rich dynamics of mappings which fold the line. More specifically, given such a shuffle permutation  $\sigma$  which is irreducible there is a folding mapping  $f$  of an interval  $I$  (Figure 1(b)) and a smaller interval  $I_1 \subset I$  about the turning point of  $f$  whose inverse orbit under  $f$  contains a *disjoint finite* collection of intervals permuted according to  $\sigma$  under the iteration of  $f$ . In fact it is known that each such shuffle permutation of intervals happens for some member of any complete family of mappings as indicated in Figure 1(b) (Milnor and Thurston [MT], Sharkovski [Sh]).

We make use of the operator  $f \rightarrow f_1$ , where  $f_1$  is the first return mapping of  $f$  to  $I_1$ . This  $f_1 = (f^n|_{I_1})$  is again a folding mapping of an interval and  $f \xrightarrow{R} f_1$  is called *renormalization*. The operator  $R$  is partially defined on folding mappings and is specified precisely by taking  $n$  minimal and  $I_1$  minimal. If  $R$  is defined and continues to be defined we define  $f_1 = Rf$ ,  $f_2 = Rf_1$ ,  $\dots$ ,  $f_n = Rf_{n-1}$ ,  $\dots$ . We say  $f$  is *infinitely renormalizable of type*  $(\sigma_0, \sigma_1, \sigma_2, \dots)$ , where the  $\sigma_i$  are the shuffle permutations that arise inductively. Infinitely renormalizable mappings of every type  $(\sigma_0, \sigma_1, \sigma_2, \dots)$  occur in every complete family (Figure 1(b)).

The description renormalization comes from statistical physics. An analogy was discovered between phenomena there like critical opalescence and one of the examples here, type  $(\tau, \tau, \tau, \dots)$ , where  $\tau$  is the permutation of order 2 (see Figure 4 from [CT2] on page 419). This type appears at the end of a cascade of period doubling bifurcations in any complete family. The physicists Feigenbaum [F1] and Coulet and Tresser [CT1] working numerically, and independently, in the U. S. and France found universal numerical characteristics about this cascade and about the limit geometry of

1991 *Mathematics Subject Classification*. Primary 58F13, 58F14; Secondary 30D05, 30F60, 39B12.

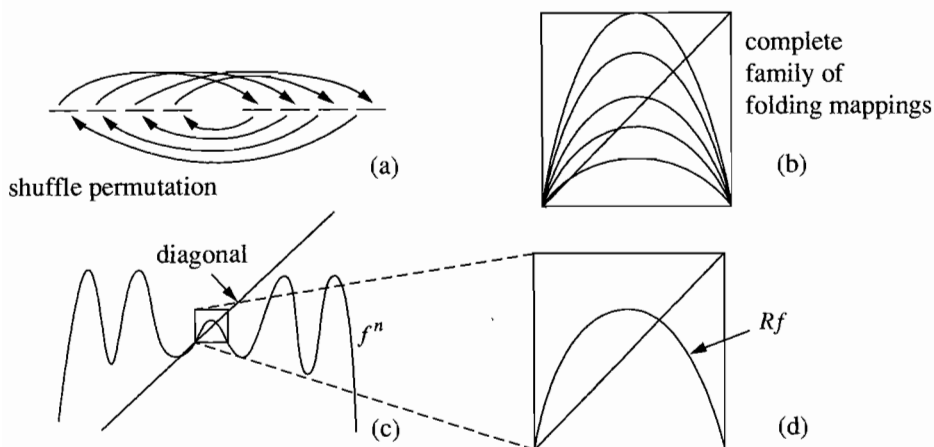


FIGURE 1

the type  $(\tau, \tau, \dots)$  orbital Cantor set. For example, in terms of any smooth control parameter the bifurcations occurred faster and faster at a relative rate  $(4.6692\dots)^n$ ,  $n$  large. The boxes within boxes for the renormalization of type  $(\tau, \tau, \tau, \dots)$  had the limiting successive ratio  $0.3995\dots$ . Finally, the graph inside the tiny box became canonical. These characteristics were universal in the sense that they were computed numerically to be independent of the choice of complete family of mappings with quadratic turnings.

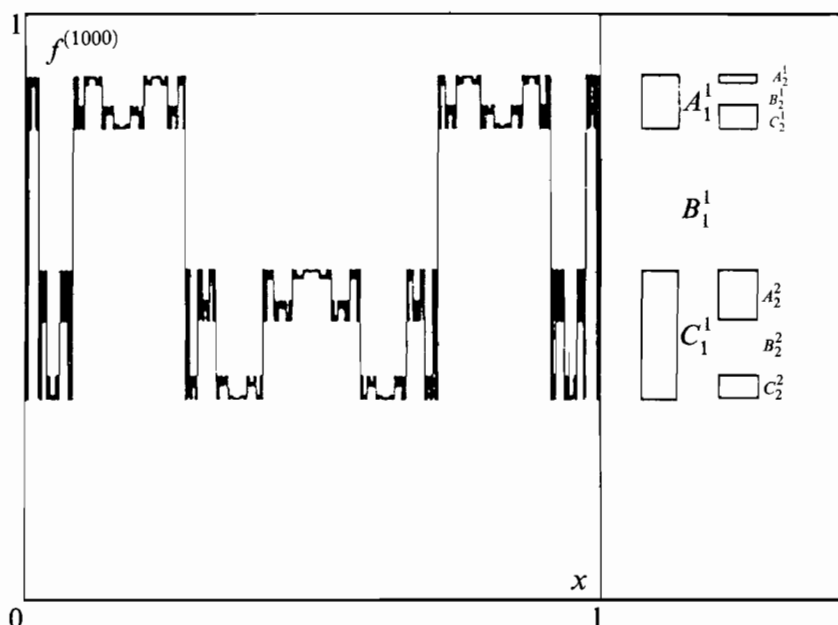
Some of us have been wondering for a long time what the domain of validity of these discoveries is and what techniques from dynamics need to be employed or invented to make a proof. We describe results obtained over the last few years in the following theorems.

Let us consider continuous mappings  $f: I \rightarrow I$ , where  $I = [a, b]$ ,  $f(a) = f(b) = a$ , and  $f$  is a local homeomorphism except at one turning point  $c$ . To express the smoothness we require, write  $f = Qh$ , where  $Q: I \rightarrow I$  is a quadratic polynomial and  $h: I \rightarrow I$  is a homeomorphism. This decomposition is unique. We say  $f$  is a smooth quadratic-like mapping bounded by  $B$  if  $h$  is a diffeomorphism and  $\varphi = \log|h'|$  satisfies the Zygmund and  $\frac{1}{2}$ -Hölder inequalities,

$$(*) \quad \left| \varphi \left( \frac{x+y}{2} \right) - \frac{\varphi x + \varphi y}{2} \right| \leq B|x-y| \quad \text{and} \quad |\varphi x - \varphi y|^2 \leq B|x-y|.$$

Thus, all  $f = Qh$ , where  $h$  is a  $C^2$  diffeomorphism, are included but our technique does not prove anything for smoothness lower than  $(*)$  which implies  $\log h'$  has a  $t \log t$  modulus of continuity (see §1). We note that the  $h$  satisfying  $(*)$  are precompact in  $C^1$  diffeomorphisms and any limit satisfies  $(*)$ .

The first theorem uses combinatorics, real analysis (§1), and the de Melo-Van Strien [MV1], Swiatek [Sw] Koebe distortion technique in real dynamics



[CT2, Figure 4]  $f_R^{(1000)}(X)$  for  $R = R_c$  is plotted on the left part. One sees the Cantor-like structure of the asymptotic orbit. On the right we sketch the first steps of the usual geometrical construction of this Cantor set.

extended to (\*) in §2. Theorem 1 is also valid for  $|x|^r$  singularities,  $r > 1$ .

**THEOREM 1** ("beau" property of renormalization). *If  $f$  is a smooth quadratic-like mapping bounded by  $B$ , then all the renormalizations  $f_1, f_2, \dots$  are smooth quadratic-like mappings with a bound depending only on  $B$ . After a number of renormalizations only depending on  $B$  further renormalizations are bounded universally.*

Using Theorem 1, the note, and the fact the critical value stays away from  $a$  (§4), we can form renormalization limits.

If  $f_0 = \lim_{i \rightarrow \infty} R^{n_i} f$ , set  $f_1 = \lim'_{i \rightarrow \infty} R^{n_i-1} f$ ,  $f_2 = \lim'_{i \rightarrow \infty} R^{n_i-2} f$ , etc., where  $\lim'$  means we take limits over subsequences. Whenever  $f_1, f_2, \dots$  are in the domain of renormalization (which is assured if we assume the individual renormalization return times are uniformly bounded), then we obtain an *inverse chain of limits related by renormalization*,

$$(**) \quad \dots \rightarrow f_{n+1} \xrightarrow{R} f_n \rightarrow \dots \rightarrow f_2 \xrightarrow{R} f_1 \xrightarrow{R} f_0,$$

where each  $f_n$  is a smooth quadratic-like mapping bounded by some universal  $B$ .

Denote the type of  $f_n$  by  $\tilde{\sigma}_n = (\sigma_0^n, \sigma_1^n, \dots)$  so that  $\tilde{\sigma}_n = (\tilde{\sigma}_{n+1}$  shifted by one). Let  $\tilde{\sigma} = \varprojlim \tilde{\sigma}_n$  be the inverse limit 2-sided infinite sequence

$\tilde{\sigma} = (\dots, \sigma_{-2}, \sigma_{-1}, \sigma_0, \sigma_1, \sigma_2, \dots)$ .

The second theorem, whose explicit formulation was motivated by a lecture of Curt McMullen at IHES (May 1990), uses holomorphic dynamics and quasiconformal mappings. It is valid also for analytic singularities  $x^{2k}$ ,  $k = 1, 2, \dots$ , but remains mysterious for  $|x|^r$  singularities,  $r$  real and greater than one. W. Paluba [Pa] has recently made some progress for  $r$  real (see also [E2] and [CEL]).

**THEOREM 2** (Generalized Feigenbaum functions). *For each bi-infinite sequence  $\tilde{\sigma} = (\dots, \sigma_{-1}, \sigma_0, \sigma_2, \sigma_2, \dots)$  of uniformly bounded shuffles there is one and only one analytic function  $F(\tilde{\sigma})$  which is a smooth quadratic-like mapping, infinitely renormalizable of type  $(\sigma_0, \sigma_1, \sigma_2, \dots)$ , and which lies at the beginning of an infinite inverse chain of uniformly bounded smooth quadratic-like mappings related by renormalization of combinatorics  $(\dots, \sigma_{-2}, \sigma_{-1})$ ,*

$$\cdots \rightarrow f_n \xrightarrow{R} f_{n-1} \rightarrow \cdots \rightarrow f_2 \rightarrow f_1 \rightarrow F(\tilde{\sigma}).$$

**COROLLARY 1.** *For each shuffle permutation  $\sigma$  and any smooth quadratic-like  $f$  (i.e.,  $f = Qh$ ,  $\log h'$  Zygmund) of type  $(\sigma, \sigma, \dots)$  we have*

$$\lim_{n \rightarrow \infty} R_\sigma^n f = F(\dots, \sigma, \sigma, \sigma, \dots).$$

Here  $R_\sigma$  means the renormalization associated to the shuffle  $\sigma$ . For example, the stable manifold of the Feigenbaum renormalization operator  $R_\tau$  at the fixed point  $F(\dots, \tau, \tau, \tau, \dots)$  consists of all the smooth quadratic-like mappings of type  $(\tau, \tau, \tau, \dots)$ .

If  $f: I \rightarrow I$  is infinitely renormalizable of type  $(\sigma_0, \sigma_1, \dots)$ , then  $I$  contains  $n_0$  disjoint intervals on each of which a conjugate of  $f_1 = R(\sigma_0)f$  is defined. Each of these  $n_0$  intervals contains  $n_1$  intervals on which a conjugate of  $f_2 = R(\sigma_1)R(\sigma_0)f$  is defined, etc. These interval collections nest down to a closed set which is the closure of the critical point orbit of  $f$ . We see in §3 that the total length of these intervals tends to zero exponentially quickly in the depth of renormalization, in the case of smooth quadratic-like mappings. Also, in the bounded type case (degree  $\sigma_i$  bounded), these Cantor sets have *bounded geometry* in the sense that the ratio  $r_{\alpha\beta}$  between the length of an interval  $I_\alpha$  at one level  $n$  and the length of an interval or gap  $\beta$  in  $I_\alpha$  at the next level  $n+1$  has a bounded logarithm. Actually more is true, for bounded combinatorics.

**MAIN THEOREM** (Coullet-Tresser geometric rigidity of Cantor sets). *All smooth quadratic-like mappings of type  $(\sigma_0, \sigma_1, \dots)$  have critical orbit Cantor sets with the same universal ratio asymptotics. Namely, if  $f$  and  $g$  have type  $(\sigma_0, \sigma_1, \dots)$ , then uniformly in the depth  $n$*

$$\lim_{n \rightarrow \infty} (r_{\alpha_n \beta_n}(f) - r_{\alpha_n \beta_n}(g)) = 0.$$

For example, the computed self-similarity ratio  $0.3995\dots$  of successive intervals nesting down to the critical point in the period doubling examples appears in the Cantor set of any smooth quadratic-like mapping of type  $(\tau, \tau, \dots)$ . Also, the Hausdorff dimension of the Cantor set of any such  $f$  is  $0.538045\dots$ .

This theorem, stating that quasiperiodic orbital Cantor sets of bounded type  $(n_0, n_1, \dots)$  have a rigid geometric structure at asymptotically fine scales, was our main objective (cf. [DGP]). In Rand [R2] and Sullivan [S3] independent proofs were given that the rigidity of the  $(\tau, \tau, \tau, \dots)$  Cantor set follows from and implies convergence of  $R_\tau^n f$  to a universal limit. The study here of renormalization naturally leads to a stronger theorem about the structure of renormalization.

Let  $\dots \rightarrow f_n \xrightarrow{R} f_{n-1} \rightarrow \dots \rightarrow f_2 \rightarrow f_1 \rightarrow f_0$  be any inverse chain of bounded smooth quadratic-like mappings related by renormalizations,  $R(\sigma_n)[f_n] = f_{n-1}$ , where degree  $\sigma_n$  is bounded.

**THEOREM 2'.** (a) *The mappings  $f_n$  are canonical analytic functions determined by the combinatorics  $\sigma_i$  and a real number  $c \in [-2, 1/4]$ . The real number  $c$  is determined by the complex analytic extension of  $f_0$  which is complex quadratic-like in the sense of Douady-Hubbard. The real number  $c$  is the unique element of  $[-2, 1/4]$  so that  $f_0$  is qc conjugate to  $z \rightarrow z^2 + c$  on a neighborhood of the invariant set by a conjugacy which is a.e. conformal there [DH1].*

(b) *For any bounded combinatorics and any  $c \in [-2, 1/4]$  there is an inverse chain (unique by (a)) with these invariants.*

Theorems 1, 2, and 2' describe the dynamics of bounded time renormalization on smooth quadratic-like mappings.

(i) *A folding mapping is either finitely renormalizable or it is infinitely renormalizable and under repeated renormalization it becomes universally bounded.*

(ii) *Any smooth quadratic-like mapping is in the image of each renormalization  $R(\sigma)$ , but only the canonical mappings of Theorem 2' are in the infinite image of renormalization restricted to a bounded part of the space of folding mappings.*

(iii) *We see topologically a hyperbolic set for renormalization with points labeled by the bi-infinite combinatorics  $(\dots, \sigma_{-2}, \sigma_{-1}, \sigma_0, \sigma_1, \sigma_2, \dots)$ , unstable manifolds labeled by backwards combinatorics  $(\dots, \sigma_{-2}, \sigma_{-1}, \dots)$  and canonically parametrized by the Douady-Hubbard internal class  $c$ , and finally stable manifolds labelled by the forward combinatorics  $(\sigma_0, \sigma_1, \dots)$  and consisting of all the infinitely renormalizable mappings of type  $(\sigma_0, \sigma_1, \dots)$  (see Figure 2 on next page).*

Along the way to Theorem 2' we derive information about the complex analytic structure of renormalization limits. Say a folding mapping  $f$  has the Epstein form if  $f = hQ$  and  $h^{-1}$  has a complex analytic injective (in



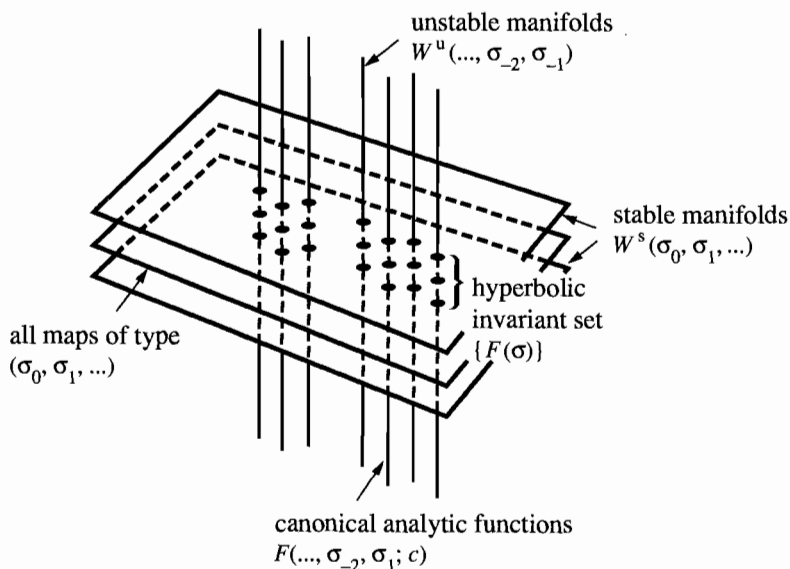


FIGURE 2

short “schlicht”) extension to  $\mathbb{C} - \{x \text{ real outside an open neighborhood of } I\}$ . We write  $f \in E(\lambda)$  if the open neighborhood of  $I$  is  $\{x \text{ so that distance } (x, I) \leq \lambda \text{ length } I\}$ .

**THEOREM 3.** *There is a universal  $\lambda > 1$  so that if  $f$  is a smooth quadratic-like mapping bounded by  $B$ , then any limit of renormalization,  $\lim R^{k_n} f = F$ ,  $k_n \rightarrow \infty$ , has the Epstein form and belongs to  $E(\lambda)$ . Unbounded combinatorics are allowed in this statement.*

Theorem 3 is based on the analytic estimates of the disjoint interval collections at the  $n$ th renormalization level. As we mentioned, the sum of lengths goes to zero exponentially in  $n$ . Also, the sum of the squares of the integrals of  $|dx/x|$  over all the intervals except the one containing the critical point stays uniformly bounded in  $n$  (§3). This means that in the exponentially long composition defining the renormalization,

$$hQ \cdots hQ \cdots hQhQ,$$

the  $h$  factors become linear on these tiny intervals and the  $Q$  factors have a bounded effect there. The inverse is basically a long composition of square roots—which yields Theorem 3. Now the real analysis is over and we must begin to work in the complex plane. We analyze compositions of square roots and schlicht mappings in §§5–7.

Theorem 3 provides preliminary information about holomorphic dynamics. (See Figure 3.) Notice the inverse branches of  $f = hQ$  are of the form  $S_{\pm} \cdot h^{-1}$ , where  $S_{\pm}$  are branches of  $z \rightarrow \sqrt{z}$  composed with appropriate linear maps. These two branches have interior disjoint images. They fit along the boundary so we can form a forward mapping with domain of

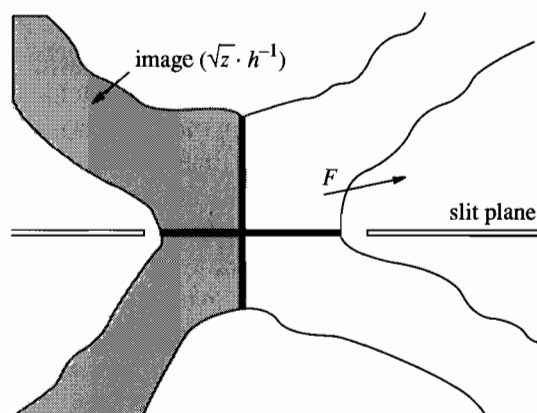


FIGURE 3

after sufficient  
renormalization

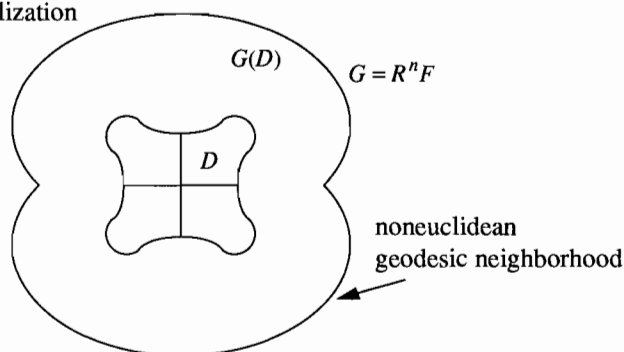


FIGURE 4

definition the union of these two images. We obtain a holomorphic mapping  $F$  extending  $f$  defined on a four-fold symmetric simply connected domain containing a definite neighborhood of the dynamic interval and mapping onto (the domain of  $h^{-1}$ )  $= \mathbb{C} - \{x \text{ real not in } I(\lambda)\}$ .

In §8, using §§3–7, we show how to cut down such a mapping after enough bounded time renormalization so it has the form of Figure 4.

**THEOREM 4.** *Assume bounded combinatorics ( $\leq T$ ) and  $F$  belongs to the Epstein class  $E(\lambda)$ . After  $n \geq n(T)$  renormalizations the inverse branches of  $G = R^n F$  map the geodesic neighborhoods of Figure 3 well inside and the annulus  $G(D) - D$  has a conformal modulus  $\geq m(T)$ .*

**NOTE.** The formulation of Theorem 4 and one of the key steps in its proof was motivated by papers of Epstein [E1, E2] and Epstein and Lascoux [EL]. A new point in the proof is a systematic use of information of Theorem 3 moving carefully up through the renormalization hierarchy.

The map depicted in Figure 4 is one of the *complex quadratic-like*

mappings of Douady and Hubbard [DH1]. In the *conjugacy pullback argument* we make use of their insight that the annulus  $G(D) - D$  with the  $\partial$ -relation given by  $G$  is a *fundamental domain* for the holomorphic dynamics of  $G$ . We go one step further (§9) and construct a Riemann surface lamination of orbits which in effect removes the branching locus singularity in the Douady-Hubbard fundamental domain modulo side identifications. In §11 one constructs a quasiconformal conjugacy whose distortion only depends on the universal bound on the conformal modulus of Theorem 4. The argument is reminiscent of the one in Michael Shub's thesis about expanding mappings. There are new twists—complex analyticity and quasiconformality replace the expanding property, branched coverings replace coverings so Thurston's insight about the role of the forward critical orbit is needed, and there is McMullen's remedy of the shrinking domain  $D \supset G^{-1}D \supset G^{-2}D \supset \dots$  using the Douady-Hubbard insight mentioned above.

Given quasiconformal conjugacies we can describe all the dynamical systems of one topological form by a Teichmüller space of conformal structures. The relevant Riemann surface lamination is studied in the appendix. Let  $Q(R, T)$  be the complex quadratic-like mappings which are symmetric about a real axis and are infinitely renormalizable of bounded type  $\leq T$ . Let  $d$  be the Teichmüller distance (appendix) on the manifolds of quasiconformally conjugate systems. Then we can use the *almost geodesic lemma* (appendix) to prove Theorem 5 (§13).

**THEOREM 5.** *There is  $\lambda(T) < 1$  so that for any two points  $x, y$  in  $Q(R, T)$  there is a power of  $R$  that reduces  $d$  by a factor of  $\lambda$ ,*

$$d(R^n(x), R^n(y)) \leq \lambda d(x, y).$$

*The power depends only on the moduli of representatives of  $x$  and  $y$  (§10).*

The renormalization  $R$  is defined canonically (§12) on  $Q(R, T)$  respecting the operation  $R$  considered previously. It is clearly distance nonincreasing,  $d(Rx, Ry) \leq d(x, y)$ . More importantly the new  $R$  is defined on the level of germs of invariant conformal structures near the Julia set so *Teichmüller theory* applies.

With Theorems 3, 4, and 5 in place, the proof of Theorem 2 is a limit argument using the idea that  $f_0$  is deeply embedded inside  $f_n$  for  $n$  large. In the type  $(\tau, \tau, \tau, \dots)$  case Curt McMullen, motivated by the rigidity theory of Kleinian groups, has a different proof of *Theorem 4 implies Theorem 2* using a geometric limit of this embedding idea (see [Mc]). Theorem 2' is really what the proof of Theorem 2 yields.

Very recently Edson de Faria [deF] found the replacement of complex quadratic-like mappings and the pullback construction in the context of critical circle mappings. Thus, one can expect versions of Theorems 2, 3, 4, and 5 in that context as well. Theorem 1 is known in that context by combining Swiatek's original argument [Sw] with the techniques of §3. Unbounded type

can also be treated better for circle mappings because of work of Yoccoz [Y1] and Lanford [L2].

Returning to folding mappings let us compare the results here with previous ones. First there is Lanford's computer assisted proof of the Feigenbaum calculations [L1]. It shows there is an analytic fixed point  $g$  of  $R_\tau$  and the linearized version of  $R_\tau$ , acting on a space of analytic functions at  $g$ , is hyperbolic with one eigenvalue  $4.6692\dots$  outside the unit circle and the rest inside. The small spectrum increases out to the unit circle as the regularity is decreased. According to Lanford on a sufficiently small real analytic neighborhood of  $g$  one has the expected hyperbolic picture for  $R_\tau$  with stable manifold all the folding mappings in the neighborhood of type  $(\tau, \tau, \dots)$ .

By the above we have a global version of Lanford's results. This fixed point  $g$ , or that of Epstein [E1] or of Campanino-Epstein-Ruelle [CER], etc. must agree with  $F(\dots, \tau, \tau, \tau, \dots)$  by the corollary to Theorem 2. At  $F(\dots, \tau, \tau, \tau, \dots)$  we have the "stable manifold"  $W^s(\tau\tau\tau, \dots)$  consisting of all smooth quadratic-like mappings of type  $(\tau\tau\tau, \dots)$ . Its intersection with Lanford's real analytic neighborhood of  $g$  is Lanford's stable manifold. At  $F(\dots, \tau, \tau, \tau, \dots)$  we have the "unstable manifold"  $W^u(\dots, \tau, \tau, \tau; c)$ . By the topological picture (Figure 2) this curve of canonical mappings constructed synthetically must extend Lanford's unstable manifold defined near  $g$  by the computer assisted hyperbolicity results.

The critical smoothness in Lanford's linear problem is certainly  $1 +$  Hausdorff dimension of the critical orbit Cantor set—although only something weaker may be rigorously proven [L3]. By the main theorem this dimension is a universal value. In the case of  $(\tau, \tau, \tau, \dots)$  it is computed [CCR] to be  $0.538045143580549911671415567\dots$ , but for the other types  $(\sigma, \sigma, \sigma, \dots)$  these dimensions vary in  $(0, 1)$ . So this fits with our smoothness class  $1 +$  Zygmund which is contained in every  $C^{1+\alpha}$  for  $\alpha < 1$  but is bigger than first derivatives Lipschitz.

Theorem 1 has a precursor in Guckenheimer [G1] in the important special case of negative schwarzian  $f'$  of period doubling combinatorics (see also [VSK]). In the circle mapping case the first part was obtained by Świątek [Sw] and Herman [H]. There have been important generalizations of [G1] in Guckenheimer and Johnson [GJ] and of our method in Martens [M]. Other important papers involving such real bounds are Lyubich [Ly2], de Melo and van Strien [MV2], Blohk and Lyubich [BL1, BL2], and Jakobson and Świątek [JS1]. Many of these results and an outline of our proof can be found in [MV].

I am grateful for the steadfast interest and guidance provided by my colleagues in Europe—Collet, Cvitanovic, Douady, Eckmann, Epstein, Herman, Lanford, Rand, van Strien, and Yoccoz—and in America—de Melo, Feigenbaum, Guckenheimer, Hubbard, McMullen, Świątek, Tangerman, and Tresser. I would also like to express appreciation to the younger mathematicians Jiang, de Faria, and Paluba who listened to and critiqued arguments

for untold hours over the last few years. Finally, I thank Wellington de Melo who has studied the proof relentlessly and provided crucial assists at various difficult points.

## CONTENTS

1. Poincaré length distortion and smoothness class one plus Zygmund
  2. The Koebe distortion argument of Denjoy, de Melo-Van Strien, Świątek, Yoccoz, et al. and Zygmund smoothness
  3. The a priori real bounds (proof of Theorem 1)
  4. Renormalization limits and schlicht mappings—the Epstein class
  5. Composition of roots and the sector theorem
  6. The factoring of the sector theorem
  7. The sector inequality
  8. The complex quadratic-like mapping produced by renormalization
  9. Douady-Hubbard theory and Riemann surface laminations
  10. The modulus function on external classes
  11. Thurston equivalences and the pull back conjugacy
  12. Renormalization of complex quadratic-like mappings
  13. Teichmüller contraction of renormalization for symmetric complex quadratic-like mappings
  14. Proof of Theorem 2'
  15. Proof of Theorem 2
- Appendix. Riemann surface laminations and their Teichmüller theory

**1. Poincaré length distortion and smoothness class one plus Zygmund.** In this section (in two stages) we show that the distortion of the cross ratio for standard 4-triples being  $O$  (scale of 4-triple) is equivalent to smoothness one plus Zygmund. Also, the smoothness implies control on cross ratio distortions for sufficiently many nonstandard 4-triples to yield the dynamical Koebe distortion argument of §2.

We want to study the smoothness required for a diffeomorphism  $h$  to only distort cross ratios of small standard 4-tuples by an amount commensurable to the size of the 4-tuple.

One cross ratio  $[a, b, c, d]$  can be computed by

$$-\log[a, b, c, d] = \iint_S \frac{dx dy}{(x-y)^2}, \quad a < b < c < d,$$

where  $S$  is  $\{(x, y) | a \leq x \leq b, c \leq y \leq d\}$ .

Thus the distortion by  $h$ , given by

$$\log \frac{[ha, hb, hc, hd]}{[a, b, c, d]},$$

equals  $\int_S \mu - (h \times h)^* \mu$ , where  $\mu$  is the measure,  $dx dy / (x-y)^2$ .

Calculating the integrand we get

$$\left( \frac{1}{(x-y)^2} - \frac{h'xh'y}{(hx-hy)^2} \right) \quad \text{or} \quad \frac{1}{(x-y)^2} \left[ 1 - \frac{h'xh'y}{[h']_{xy}^2} \right],$$

where  $[h']_{xy} = \text{average}_{[x,y]} h'$  is the average of  $h'$  over the interval  $[x, y]$ .

Because we are assuming  $b-a=c-b=d-c$ , for every point  $(x, y)$  in the square  $S$  the factor  $1/(x-y)^2$  is commensurable to  $1/\text{area } S$ . Thus, a small bound  $\varepsilon$  on  $\log(h'xh'y/[h']_{xy}^2)$  yields the bound  $\varepsilon$  on the distortion of the cross ratio,  $\log[ha, hb, hc, hd]/[a, b, c, d]$ .

Let us say  $h$  satisfies the *local Koebe condition* if for  $|x-y|$  sufficiently small one of the following equivalent conditions holds:

- (1)  $[1 - (h'xh'y/[h']_{xy}^2)] = O(|x-y|)$ ,
- (2)  $\log(h'xh'y/[h']_{xy}^2) = O(|x-y|)$ .

**PROPOSITION.** *If  $h$  satisfies the local Koebe condition, then the  $h$  distortion of cross ratios of small standard 4-tuples is commensurable to the size of the 4-tuple.*

**PROOF.** The proposition follows from the above calculations. Q.E.D.

Calculating the log in (2) we get

$$\log h'x + \log h'y - 2 \log [h']_{xy}.$$

Let us replace the last term with the average taken after the log to obtain (a)  $(\log h'x + \log h'y - 2[\log h']_{xy})$  with an error of twice (b)  $(\log \text{average}_{[x,y]}(h') - \text{average}_{[x,y]}(\log h'))$ .

**NOTE.** If both (a) and (b) are  $O(|x-y|)$ , then (1) and (2) hold.

Expression (a) suggests the Zygmund condition on continuous functions:

$$Z: \varphi(x) + \varphi(y) - 2\varphi\left(\frac{x+y}{2}\right) = O(|x-y|).$$

**PROPOSITION.** *If  $\varphi$  satisfies  $Z$  on an interval  $J$ , then the average of  $\varphi$  over  $J$  is the value of  $\varphi$  at the midpoint with an error  $O(\text{length } J)$ .*

**PROOF.** Think of the uniform measure on  $J$  as two dirac masses moving out uniformly from the center. Use the  $Z$  condition to replace the average of  $\varphi$  at the moving points by the value at the center. Q.E.D.

**COROLLARY.** *If  $\log h'$  is Zygmund, then expression (a) is  $O(|x-y|)$ .*

**PROOF.** Use the proposition, then the definition of  $Z$  again.

There is a converse to the corollary. Say  $\varphi$  satisfies the *average property* if  $\text{average}_{[x,y]} \varphi = \frac{1}{2}(\varphi(x) + \varphi(y)) + O(|x-y|)$ .

**PROPOSITION.** *If  $\phi$  satisfies the average property for all intervals  $J \subset I$ , then  $\phi$  satisfies the Zygmund property for all pairs  $x, y$  in  $I$ .*

**PROOF.** Apply the average property to  $[x, (x+y)/2]$ ,  $[(x+y)/2, y]$ , and  $[x, y]$ , and combine averages of averages to get the Zygmund property for  $x, y$ .

**COROLLARY.** *The Zygmund property is equivalent to the average property.*

**PROOF.** The corollary follows from the propositions above.

**CONCLUSION A.** Expression (a) is  $O(|x-y|)$  iff  $\log h'$  is Zygmund.

Now we consider when expression (b) is  $O(|x-y|)$ . We are concerned with small intervals  $J$  and we assume  $h'$  is continuous. Then  $h'$  varies only a little from one of its values  $h'(x_0) = a$ . Expression (b) is unchanged if we multiply  $h'x$  by  $1/a$ . Write  $(1/a)h'$  on  $J$  as  $1 + \varepsilon$ , where  $\varepsilon$  is a small function. Expand the two terms of (b) as:

$$\begin{aligned} & \log \frac{1}{|J|} \int_J (1 + \varepsilon) - \frac{1}{|J|} \int_J \log(1 + \varepsilon) \\ &= \left( \frac{1}{|J|} \int_J \varepsilon - \frac{1}{2} \left( \frac{1}{|J|} \int_J \varepsilon \right)^2 \cdots \right) - \left( \frac{1}{|J|} \int_J \varepsilon - \frac{\varepsilon^2}{2} \cdots \right) \\ &= -\frac{1}{2} \left( \frac{1}{|J|} \int_J \varepsilon \right)^2 + \frac{1}{|J|} \int_J \varepsilon^2/2 \cdots \end{aligned}$$

Here the first term could be zero so there would be no cancellation. Thus, we estimate each brutally with absolute values. Assume  $\varepsilon$  is Hölder of order  $\frac{1}{2}$  on  $J$ ,  $|\varepsilon(x) - \varepsilon(y)|^2 \leq C_J |x - y|$ . Since  $\varepsilon$  is zero at  $x_0$ , we get the estimate  $C_J \cdot \text{length } J$  for the sum of the absolute values. Also, if  $C_J \cdot \text{length } J$  is sufficiently small, then the higher order terms can be ignored.

**CONCLUSION B.** Expression (b) is  $O(|x-y|)$  if  $h'$  is Hölder of order  $\frac{1}{2}$ . The coefficient for  $|x-y| < \varepsilon$  is estimated by the *normalized  $\frac{1}{2}$ -Hölder norm*: take the sup over all intervals  $J$  of length  $\leq \varepsilon$  of  $C_J$  above, where  $1 + \varepsilon = h'(x)/h'(x_0)$  for convenient  $x_0$  in  $J$  and we assume  $C_J |\text{length } J|$  is sufficiently small.

Let us note that Zygmund functions are  $\alpha$ -Hölder for all  $\alpha < 1$ . However, the  $\alpha$ -Hölder constants are not determined by the Zygmund norm. Let us also note the normalized  $\frac{1}{2}$ -Hölder norm of  $h'$  can be estimated by the square of the usual  $\frac{1}{2}$ -Hölder norm of  $\log h'$  — the best  $C$  such that

$$|\log h'x - \log h'y|^2 \leq C|x-y|.$$

Now we can summarize the above by the following theorem:

**THEOREM.** (a) *If  $\log h'$  is Zygmund, then  $h$  satisfies the local Koebe distortion condition. The coefficient is controlled by the Zygmund norm of  $\log h'$  and the  $\frac{1}{2}$ -Hölder norm of  $\log h'$ . Conversely,*

(b) if  $\log h'$  is  $\frac{1}{2}$ -Hölder, then the local Koebe condition for  $h$  implies  $\log h'$  is Zygmund. (See Remark below for a stronger statement.)

PROOF. The above discussion has been a proof of (a). For part (b) recall from above that the local Koebe inequality implies that expression (a) plus expression (b) is  $O(|x - y|)$ . The  $\frac{1}{2}$ -Hölder norm implies expression (b) is  $O(|x - y|)$ . Thus, expression (a) is  $O(|x - y|)$ . But this implies  $\log h'$  is Zygmund by the third proposition above. Q.E.D.

Other results about cross ratio distortion can be found in [MV].

PROBLEM. Derive necessary and sufficient conditions for the integral distortion to be commensurable to the linear scale. (In the above discussion we have estimated the integral by the integrand.)

REMARK (added December 1990). Actually we can solve this problem. By a standard 4-triple  $(a, b, c, d)$  we mean one where  $(b - a) = (c - b) = (d - c)$ . The solution of the problem is to show a homeomorphism  $h$  which distorts a standard tiny 4-triples' cross ratio by  $O(\text{scale})$  is a diffeomorphism with  $\log h'$  Zygmund, and conversely. We sketch this.

The same method shows  $O(\text{scale})^\alpha$  cross ratio distortion is equivalent to  $\log h'$  is  $C^\alpha$ ,  $0 < \alpha < 1$ , and  $C^{1, \alpha-1}$ ,  $1 < \alpha \leq 2$ , while distortion  $o(h^2)$  is equivalent to  $h$  being Moebius.

The proof consists of studying for a fine grid of intervals  $I_\beta$  the approximate derivatives  $d_\beta = |hI_\beta|/|I_\beta|$ , the ratio distortions  $r_\beta = |hI_{\beta'}|/|hI_\beta|$  for consecutive intervals, and the cross ratio distortions  $c_\beta = \text{change in } \ln(1 + (|I_{\beta'}|(|I_\beta| + |I_{\beta'}| + |I_{\beta''}|))/|I_\beta||I_{\beta''}|)$ , where  $I_\beta, I_{\beta'}, I_{\beta''}$  are consecutive intervals. If  $l_\beta = \frac{1}{2} \log d_\beta$  and  $\varepsilon_\beta = \frac{1}{2} \log r_\beta$ , then

- (1)  $\varepsilon_\beta = l_{\beta'} - l_\beta$  exactly,
- (2)  $c_\beta = \varepsilon_{\beta'} - \varepsilon_\beta$  modulo higher order terms in  $\varepsilon_{\beta'}$ .

One may also compare maximum ratio distortions  $\varepsilon(t)$  at two adjacent scales and find

- (3)  $\varepsilon(2t) \geq 2\varepsilon(t) + c(t) + \text{higher order terms in } \varepsilon(t)$ , where  $c(t)$  is the maximum cross ratio distortion at scale  $t$ .

To use these relationships there are two important preparation lemmas:

LEMMA 1. If  $c(t)$  is bounded on some open interval, then  $\varepsilon(t)$  is bounded on every closed subinterval.

LEMMA 2. If  $c(t)$  tends to zero on some open interval, then  $\varepsilon(t)$  tends to zero in every closed subinterval.

Lemma 1 follows from the *four intervals remark*: if for four consecutive equal intervals the middle two are mapped to very disparate intervals, then the Poincaré length  $\ln(1 + \frac{MT}{LR})$  of one interval  $M$  in three others  $T = L + M + R$  is greatly increased by the map (see §2).

Lemma 2 uses the exact relationship between  $c_\beta$  and  $\varepsilon_\beta$  to say that if  $\varepsilon_\beta > 0$ , then  $\varepsilon_{\beta'} > \varepsilon_\beta$  mod terms the size of  $c_\beta$ . Then at twice the scale one



gets at least twice the ratio distortion. Using this amalgamation many times contradicts the boundedness of ratio distortion unless  $\varepsilon_\beta \rightarrow 0$  with the scale.

Now if  $c(t) = O(t^\alpha)$ ,  $0 < \alpha < 1$ , one uses (3) to get  $\varepsilon(t) = O(t^\alpha)$ , which is equivalent to  $h$  being a  $C^{1+\alpha}$  diffeomorphism.

If  $c(t) = O(t)$  we use (1) and (2) to get

(4)  $l_{\beta''} + l_\beta - 2l_{\beta'} = O(t)$  plus terms  $O(t^{2\alpha})$ ,  $\alpha < 1$ . Since  $\varphi = \log h'$  is  $\alpha$ -Hölder for all  $\alpha < 1$ , we have  $|\frac{1}{|I|} \int \log h' - \log \frac{1}{|I|} \int h'| = O(t^{2\alpha})$  for all  $\alpha < 1$ .

Thus for  $\log h'$  (4) implies that the average over an interval  $I$  is the average over the middle  $\frac{1}{3}$  with error  $O(|I|)$ . Iterating this yields that the average of  $\varphi$  over  $I$  equals the value of  $\varphi$  at the midpoint with error  $O(|I|)$ . Thus,  $\varphi = \log h'$  satisfies the Zygmund condition,

(5)  $\varphi(x+t) + \varphi(x-t) - 2\varphi(x) = O(t)$ .

Continuing, if  $c(t) = O(t^\alpha)$  for  $1 < \alpha < 2$ , one gets versions of (4) and then (5) with  $O(t)$  replaced by  $O(t^\alpha)$ . Dividing by  $t$  and looking at a geometric series over the scales yields  $\varphi'x$  is  $C^{\alpha-1}$ . Now if  $\alpha = 2$ , we can have (4) with error  $O(t^2)$  and then (5) with error  $O(t^2)$ , which by the same argument gives  $\varphi'$  is Lipschitz.

If  $c(t)$  is  $O(t^2)$ , one can try to define the Schwarzian as  $\lim c(t)/t^2$ . Thus,  $c(t) = o(t^2)$  means  $h$  is Moebius.

REMARK. These calculus results can be used to prove familiar results about circle diffeomorphisms  $f$  for optimal levels of smoothness in the above scale. The following is a scorecard:

	$C^{1+\alpha}, 0 < \alpha < 1$	$C^{1+\text{Zygmund}}$	$C^{1+\text{zygmund}}$
cross ratio distortion	$O(t^\alpha)$	$O(t)$	$o(t)$
Denjoy's theorem	no	yes	yes
ergodicity of $f$	no	yes	yes
renormalization is bounded	no	yes	yes
renormalization limits are rotations	no	no	yes
M. Hermann ratio rigidity for bounded type	no	no	yes

See forthcoming notes from CUNY.

**2. The Koebe distortion argument of Denjoy, de Melo-Van Strien, Swiatek, Yoccoz, et al. and Zygmund smoothness.** Consider a composition  $g$  of many diffeomorphisms  $f_i$  between tiny intervals  $J_i$  all lying disjointly in some big interval  $I$ ,  $f_i: J_i \rightarrow J_{i+1}$ .

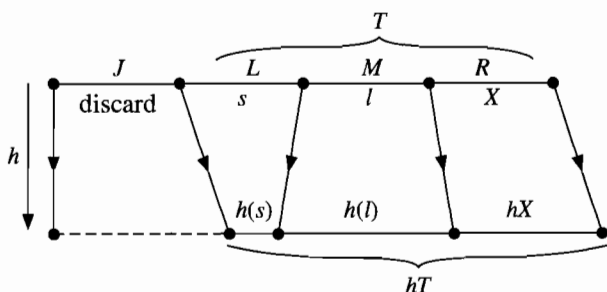


FIGURE 1

The classical Denjoy argument estimates  $\log |g'x/g'y|$ ,  $x, y \in \text{domain } g$ , in terms of the  $\sum_i$  total variation  $|\log f'_i|$ . This will be finite, say, if  $f_i = f/I_i$  and  $\log f'$  has bounded variation on  $I$ . The proof is by the chain rule.

The new argument, called the Koebe principle for one-dimensional real dynamics, treats the case when the factors can be divided into two groups so that relative to some coordinate system on  $I$

(i) for one group a Denjoy type argument can be used at least to study cross ratios.

(ii) the factors in the other group decrease Poincaré length (a type of cross ratio) (because of a positive Schwarzian condition) even though  $\log f'_i$  can have unbounded variation.

Here if  $L, M, R$  is a partition of an interval  $T$  into three consecutive subintervals (the left, the middle, and the right) the *Poincaré length* of  $M$  in  $T$  is  $\log(1 + \frac{MT}{LR})$ . It is the length of  $M$  in the Riemannian metric on  $T = [a, b]$  corresponding to the form  $|dx|/(x-a) + |dx|/(b-x)$ .

The additive change of Poincaré length (P-length) along a composition is additive over the factors. In a decomposition such as (i), (ii) above, the increase in P-length is controlled by the factors of type (i) because there is a decrease for the factors of type (ii). This is the first idea (cf. Świątek [Sw]).

The second idea is the four intervals argument. Let  $J, L, M, R$  be contiguous equal length intervals and let  $h$  be a homeomorphism of the union into the real line so that one of  $hL$  and  $hM$  is much smaller than the other. Discard from the original four intervals the outer interval next to the one of  $L$  or  $M$  made smaller, called  $s$ . Let  $T$  denote the union of the remaining three  $L, M, X$  and let  $l \subset T$  be the one of  $L$  or  $M$  made larger. The P-length of  $l \subset T$  is  $\log 4$ . The P-length of  $h(l) \subset hT$  is very large because  $h(l)$  is much larger than  $h(s)$  and  $h(T)$  is of course greater than  $hX$ . Thus, one has the analogue of complex Koebe distortion:

**REAL KOEBE DISTORTION.** *If a homeomorphism  $h: I \rightarrow \text{reals}$  does not increase unit P-lengths too much the quasimetric distortion of interior symmetric triples is controlled.*

More precisely, if  $x, y \in I$  satisfy that  $|x - y|$  is as small as the distance to  $\partial I$  and  $z = (x + y)/2$ , then  $1/M \leq (h(x) - h(z))/(hy - hz) \leq M$ , where

$M$  can be calculated from the bound  $B$  on the additive increase of Poincaré length of unit Poincaré length subintervals  $J \subset T$  where  $T \subset I$ , i.e., the  $B$  defined by

$$(\text{P-length of } hJ \subset hT - \text{P-length } J \subset T)_+ \leq B$$

for all  $J \subset T$  so that  $\text{P-length } J \subset T = 1$ .

REMARK. The point here as in Koebe distortion for schlicht mappings is we go from one analytic condition (in that case holomorphic; in this case positive Schwarzian or controlled P-length increase) to interior control on the nonlinearity.

We describe the dynamic Koebe distortion principle for a rather general class of dynamic systems. Let  $M$  be a compact one-manifold provided with a differentiable structure, where overlap homeomorphisms  $h_{\alpha\beta}$  are continuously differentiable and the  $\log h'_{\alpha\beta}$  have bounded Zygmund norm (see §1).

Suppose  $f: M \rightarrow M$  is a smooth mapping with finitely many critical points where  $f' = 0$ . At a nonsingular point assume  $\log f'$  is Zygmund. At a singular point  $c_i$  suppose there are coordinate systems in the  $(1 + \text{Zygmund})$  structure so that  $f$  takes the form  $x \rightarrow |x|^{r_i} + v_i$  or  $x \rightarrow (\text{sign } x)(|x|^{r_i}) + v_i$ , where  $r_i > 1$ .

Assume we have a long composition  $g$  of diffeomorphisms  $f_i: J_i \rightarrow J_{i+1}$  where the  $J_i$  are disjoint in  $M$  and  $f_i^{-1} = f$  restricted to  $J_{i+1}$ . Quasisymmetric distortion is defined informally above and formally in §3.

THEOREM. *For the composition  $g$  the increase in Poincaré length and therefore the interior quasisymmetric distortion of  $g$  in domain  $g$  is controlled by constants of the coordinate systems and local models of  $f$ , independent of the length of the composition  $g$ .*

PROOF. We first need a lemma.

LEMMA. *If  $h$  is a diffeomorphism of the unit interval  $I$ ,  $\log h'$  is Zygmund,  $T \subset I$  is a tiny interval, and  $J \subset T$  has unit Poincaré length, then the Poincaré length of  $hJ \subset hT$  is  $1 + O(\text{length } T)$ . The coefficient is controlled by the Zygmund norm of  $\log h'$  and the  $\frac{1}{2}$ -Hölder norm of  $\log h'$  squared.*

PROOF OF THE LEMMA. We have proved this in §1 when  $J$  sits in the middle of  $T$ . In general  $J$  may be tiny and near one end of  $T$ . We have to calculate the integral of §1 over the rectangle  $R$  of Figure 2. Control on the integral yields that the control on P-length changes for 4-triples of P-length  $\sim 1$ .

Using the local Koebe condition, and the fact that for a point in  $R$  the distance to the diagonal and the vertical distance to the diagonal are equivalent, the integral takes the form

$$a \cdot \int_a^b \frac{1}{t^2} O(t) dt,$$

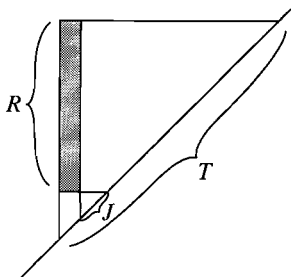


FIGURE 2

where  $a \sim \text{length } J \sim \text{distance}(J, \partial T)$  and  $b \sim \text{length } T$ . This yields  $a \log b/a$  which has order  $b$  when  $a$  and  $b$  are commensurable. This is the case already discussed. Otherwise, if  $a \ll b$ ,  $a \log b/a$  is much smaller than  $b$ . This proves the lemma.

**PROOF OF THEOREM.** (i) As we go along the composition a Poincaré length is decreased if we are entirely within one of the coordinate systems for the singular point models because  $f^{-1}$  has positive schwarzian there and maps of positive schwarzian decrease Poincaré length (de Melo and Van Strien [MV1], [MV2]).

(ii) There are finitely many possible transitional cases for long intervals which do not fit inside one model or the other. We will not discuss these further. They are finite.

(iii) Finally we have the factors where the lemma applies. We view the lemma as saying intervals  $J \subset T$  of any P-length  $\geq 1$  cannot increase by more than the multiplicative factor  $1 + O(\text{length } T)$ . By disjointness of the orbit of  $T$  this effect is controlled by the total length of  $M$ . Q.E.D.

For an alternative exposition of the theorem see [MV].

**3. The a priori real bounds** (proof of Theorem 1).. Let  $f: I \rightarrow I$ ,  $f(a) = f(b) = a$ ,  $I = [a, b]$ , be a smooth quadratic-like mapping, i.e.,  $f = Qh$ , where  $Q: I \rightarrow I$  is a quadratic polynomial and  $h: I \rightarrow I$  is a diffeomorphism with  $\log h'$  Zygmund. Recall we bound  $h$  in terms of  $\varphi = \log h'$  by the best  $B$  so that for all  $x, y$  in  $I$ ,

$$\left| \frac{\varphi(x) + \varphi(y)}{2} - \varphi\left(\frac{x+y}{2}\right) \right| \leq B|x-y|, \quad |\varphi(x) - \varphi(y)|^2 \leq B|x-y|.$$

We say that  $f$  is bounded by  $B$ .

**THEOREM 1** (“beau” property of renormalization). *Any sequence of renormalizations of  $f$  is bounded in terms of  $B$  and after a number of renormalizations depending on  $B$  the bound is universal and independent of  $B$ .*

**PROOF.** We will continually meet bounds with the properties of the theorem. We say such a quantity is “beau” (bounded and eventually universal). We will number some useful statements developed along the way. Let

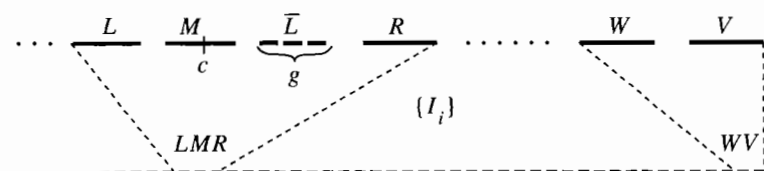


FIGURE 1

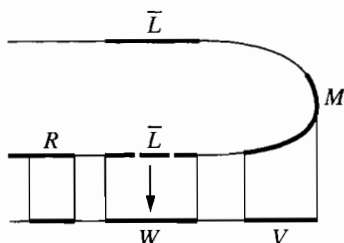


FIGURE 2

$\{1, 2, 3, \dots\}$  denote the forward orbit of the critical point  $c$  and let  $I_j$  or  $I(j)$  for  $j = 1, 2, \dots, q_n$  denote the intervals bounded by  $\{j, j + q_n\}$ , where the  $n$ th renormalization of  $f$  is defined by  $R^n f = f^{q_n}/I'$  and  $I' \supset I(q_n)$ . In Figure 1  $M$  is  $I(q_n)$ , the interval containing the critical point,  $L$  is its neighbor among the  $I_j$  mapping to the neighbor  $W$  of  $V = I_1$ , the interval containing the critical value, and  $R$  is the other neighbor among the  $I_j$  of  $M$ .  $\bar{L}$  is the mirror image of  $L$  and it lies in the gap  $g$  between  $M$  and  $R$  (see Figure 2).

To derive the “beau” property of the bounds we first go through all the steps to get some bound, then we go back through the steps again to achieve the eventually universal property. Let  $(I, J)$  denote the smallest interval containing  $I$  and  $J$ .

(1) *There is  $\lambda > 1$  independent of  $n$  so that  $\text{diameter}(L, M) \geq \lambda \text{length } M$ .*

PROOF OF (1). Suppose  $I_s$  among the  $I_j$  is smallest, and  $s > 2$ . By the next item (2)  $f^{s-1}: I_1 \rightarrow I_s$  has an inverse branch defined on  $I_s$  and its immediate neighbors, which are longer than  $I_s$ . Using Koebe (§2),  $\text{diameter}(W, V) \geq \lambda' \text{length } V$ .

Now we use the fact that  $h^{-1}$  controlled by  $B$  and  $Q^{-1}$  has bounded quasimetric distortion to do the one more preimage required and to treat the cases  $s = 1$  and  $s = 2$ .

DEFINITION. The *quasisymmetric distortion*  $M(q)$  of a homeomorphism  $q$  of an interval  $I$  is the sup of  $\log |(q(x) - q(y))/(q(y) - q(z))|$  over all symmetric triples  $x, y, z$  in  $I$ .

(2)  *$f^j: I_1 \rightarrow I_{j+1}$  has a continuous inverse branch defined on the span of  $I_{j+1}$  and its immediate neighbors among the  $I_\alpha$ .*



FIGURE 3

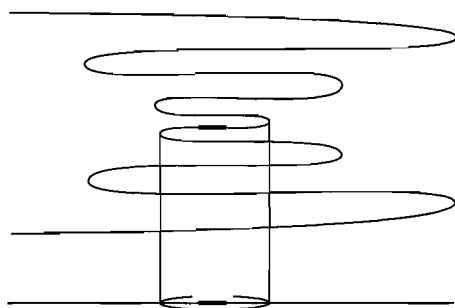


FIGURE 4

PROOF OF (2). Each endpoint of an  $I_j$  is the image of the fold and these are oriented as in Figure 3. This is true for  $I_n$  because  $g = f^{q_n}/I_n$  is a unimodal map and its endpoints are the critical value of  $g$  and its image. It is then seen for the other intervals by applying the dynamics.

Now any inverse branch defined near  $x$  for the  $k$ th power of a folding map has a maximal interval of definition which is bounded by either endpoints of the original interval  $I$  or by the forward images of the critical point whose folds point away from  $x$ . See Figure 4.

For  $f_j$  these limiting points cannot be the endpoints of  $I_{j+1}$  which are  $\{j+1, j+1+q_n\}$  because  $j$  is smaller than each. Thus, these can only be the outer endpoints of the neighbors or further away. Q.E.D.

(3) REMARK. It seems we cannot proceed without the combinatorial fact (2) which was the breakthrough point for Theorem 1.

(4) Any composition of inverse branches of length  $l$  starting at  $I_j$  for  $l < j$  has quasisymmetric distortion bounded on  $I_j$  plus a definite proportional neighborhood.

PROOF OF (4). Any  $f^{-j}$  for  $j < q_n$  is defined on span (LMR) by (2) so by (1) and Koebe all these have bounded quasisymmetric distortion on a definite neighborhood of  $M$ . Call this extra room the "Koebe space." Now we can move the Koebe space around  $M$  around to each of the intervals and apply Koebe again to obtain the proof of (4).

(5) Now we would like to repeat the above argumentation using the bigger (official) renormalization intervals  $I'_\alpha$ , namely the appropriate subcollection of the inverse orbit of  $I' \subset I$ , where  $g = f^{q_n}/I'$  is a smooth quadratic-like mapping, in the above sense,  $g(\partial I') \subset \partial I'$ . These intervals are interior disjoint and bounded by the points of a periodic orbit and some of their preimages. A modification of the above is required. Point (2) fails literally. However it only fails finitely. Namely, in the inverse of  $f^j: I'_1 \rightarrow I'_{j+1}$  there

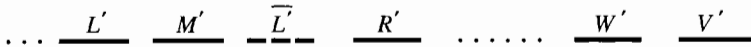


FIGURE 5

may be one or two  $Q^{-1}$  factors with poles in images of the neighbors of  $I'_{j+1}$ . Thus, we may factor  $f^{-j}$  on the span of  $I'_{j+1}$  and its immediate neighbors into at most three compositions where Koebe (§2) applies, interposing at most two quasimetric maps given by the  $Q^{-1}$  factors. There is one dangerous possibility that must be excluded. If the first  $Q^{-1}$  factor cuts away most of the Koebe space it follows that  $\text{diam}(L', M')$  is very large compared to  $\text{diam } M'$ , where we use the prime analogue of the notation of Figure 1 (see Figure 5). If so we pick up the previous argumentation at point (4). Then if at some point Koebe space is cut away as we pull back the triple of intervals  $(L', M', \bar{L}')$ , then for the composition up to this point we must have  $L' \rightarrow \cdots \rightarrow V'$  and  $M' \rightarrow \cdots \rightarrow W' \rightarrow L'$ . The second composition is qs by Koebe. Thus, the relevant critical point is quasicentered in  $L'$  because the critical point of  $M'$  is centered in  $M'$ . It follows by Koebe the first composition is qs on part of  $L'$  between the critical point and  $M'$ . Then the critical point of  $V'$  is not too close to  $W'$  relative to  $\text{diam}(W')$ . Thus, the loss of Koebe space is controlled at this one dangerous moment, which is the only point different from the previous argumentation.

- This proves we have statement (4) with  $I'_\alpha$  replacing  $I_\alpha$ .
- (6) Note we have shown the analogue of statement (1) with the corresponding prime notation: *There is a  $\lambda' > 1$  independent of  $n$  so that diameter  $(L', M') \geq \lambda'$  length  $M'$ .*
- (7) *The total length of the intervals  $I'_\alpha$  decreases by a definite factor each time we renormalize.*
- PROOF OF (7). Consider a sequence of renormalizations of combinatorics  $\sigma_1, \sigma_2, \dots, \sigma_k$  of degrees of  $n_1, n_2, \dots, n_k$ . Apply point (6) to the composed renormalization to find a gap next to the critical point interval of size comparable to it. By point (5) we can move this gap around to be adjacent to each of  $n_1, n_2, \dots, n_k$  intervals.
- Now think of this  $\sigma_1, \sigma_2, \dots, \sigma_k$  renormalization as a  $\sigma_k$  renormalization of the  $\sigma_1, \dots, \sigma_{k-1}$  renormalization. We see these moved around gaps among the new gaps of the  $\sigma_k$ -renormalization. This shows the length decreases by a definite factor because of our constructed gaps in each  $\sigma_k$  packet.
- (8) *Each of the bounds is “beau.”*
- PROOF OF (8). The total length of the intervals goes to zero at an exponential rate whose constants depend on  $B$ . But then the accumulated effect of  $h$  on the above bounds tends to zero exponentially fast at a rate depending on  $B$ . All the other considerations were independent of  $f$  and  $B$ . Thus, any bound derived as above is “beau.”
- (9) *In a linear coordinate system where  $x = 0$  is the critical value define*

the nonlinearity of an interval  $J$  to be  $\int_J |dx/x| = n(J)$ . Then  $\sum' n(I'_\alpha)^2 = O(1)$ , where we sum over all the intervals except the critical value interval—which has infinite nonlinearity.

PROOF OF (9). If we move around through the entire circuit we can study how the Poincaré lengths of  $I'_{q_n}$ , in a neighborhood where the inverse branches are injective, are altered. We can choose this neighborhood to be definite since each  $I_j \subset I'_j$  contains a point roughly centered (because this is true for  $j = q_n$ , then use (5)) and we have (2). The effect on Poincaré length is bounded by (5) on the one hand and can be calculated in terms of the sum and a bounded effect due to  $h$ . By the same reasoning as in (8) the bound is “beau.”

(10) Now look at the entire renormalization down to some level. We see a long composition of  $h$ 's and  $Q$ 's making up the diffeomorphism  $h_n$  in the decomposition  $R^n f = Qh_n$ . All the partial compositions are uniformly quasimetric by (5). Thus, they are uniformly Hölder of some exponent. A  $Q^{-1}$  factor on  $I_\alpha$  has  $B$ -norm  $\sim (nI_\alpha)^2$  and the total  $B$ -norm is bounded by (9).

Consider each factor of the composition as acting on a definition neighborhood of its dynamical interval among the  $I'_\alpha$ . Rescale each of these larger intervals to be a standard interval and consider an exponentially small subinterval  $J$ . For each partial composition the image of  $J$  is also exponentially small by the uniform Hölder property coming from the uniform quasimetric bound.

Applying the Zygmund control on  $h$  and (9) we see the distortion of standard cross ratio on the scale of  $J$  is exponentially small. This implies (last remark §1) that on the rescaled dynamic interval inside the standard interval the quasimetric distortion is exponentially small. This calculus exercise shows such homeomorphisms are uniformly  $C^{1+\alpha}$ , for appropriate  $\alpha$  (and conversely). We take from this that the first derivatives of all partial compositions of the rescaled dynamic intervals  $I'_\alpha$  to itself are on the order of one. Repeating the cross ratio argument using this Lipschitz control yields  $O(\text{scale})$  distortion of cross ratio of standard 4-tuples. Lipschitz plus  $O(\text{scale})$  control implies  $B$ -bounded on a closed subinterval (§1). Q.E.D.

**4. Renormalization limits and schlicht mappings—the Epstein class.** Let us go one step beyond the “beau” property of the sequence of renormalizations of  $f$ , discussed in the previous section. We assume  $f: I \rightarrow I$  is a smooth quadratic-like mapping bounded by  $B$  which is infinitely renormalizable with combinatorics  $\sigma = (\sigma_1, \sigma_2, \dots)$ . Let  $f_1, f_2, \dots$  be the sequence of renormalizations of  $f$  and write  $f_n = Q_n h_n$ , where  $Q_n$  is the quadratic polynomial  $Q_n: I \rightarrow I$  satisfying  $Q_n(a) = Q_n(b) = a$  and  $Q_n(c') = f_n(c)$ , where  $c'$  is the critical point of  $Q_n$  and  $c$  is the critical point of  $f$ ,  $h_n: I \rightarrow I$  is a diffeomorphism, and  $\partial I = \{a, b\}$ .



**THEOREM.** *The family of renormalizations  $\{f_n = Q_n h_n\}$  is precompact in the sense that the critical value of  $Q_n$  is bounded away from  $a$  and  $\{h_n\}$  is precompact in the  $C^{1+\alpha}$  topology on diffeomorphisms for any  $\alpha < 1$ . Any  $C^0$  limit of the  $f_n$  is a folding mapping in the symmetric form  $hQ$  and  $h^{-1}$  has a complex analytic injective extension to  $\mathbb{C} - \{x \text{ real but not in a universal neighborhood of } I\}$ . Also,  $Q_n \rightarrow Q$  and  $h_n \rightarrow h$  in the  $C^{1+\alpha}$  topology for any  $\alpha < 1$ .*

**DEFINITION.** If  $J \supset I$  is the neighborhood of  $I$  where  $h^{-1}$  of the theorem is defined we say  $f = hQ$  belongs to the Epstein class and we write  $f \in E(J)$ .

**REMARK.** We will see later the limit only depends in the bounded combinatorics case on the type  $\sigma = (\sigma_1, \sigma_2, \dots)$  of  $f$ .

**PROOF.** In the long composition

$$R^n f = Qh \cdots QhQhQh/I_{q_n}$$

( $2q_n$  factors) all the  $h$  factors are becoming linear exponentially fast in  $n$  in the sense that the sum of all their  $B$ -bounds is tending to zero exponentially fast (see (9), §3). Also all the partial compositions up to the last  $Q$  on the right are  $B$ -bounded by the work of the last section. In fact all this was seen to be true in a definitely larger neighborhood. Using the Lipschitz property of composition as a map from  $C^{1+\alpha} \times C^{1+\alpha} \rightarrow C^\alpha$  we can remove the  $h$  factors one at a time. Thus,  $R^n f$  can be written in the Epstein form with an exponentially small error in every  $C^\alpha$ ,  $\alpha < 1$ . Note the resultant composition of  $Q$ 's being schlicht is controlled by its action on two interior points of the interval.

Then we use the corollary below and the boundedness of  $h_n$  to see that  $Q'(a)$  is bounded from below. Thus,  $Q_n$  is bounded away from the zero quadratic polynomial. We can form limits and the above estimates prove what we want. Q.E.D.

Now we turn to the corollary which needs a lemma:

Let  $f: I \rightarrow I$  be any folding mapping with  $f(a) = f(b) = a$ ,  $\partial I = \{a, b\}$ .

**LEMMA.** *Either (a) there is no fixed point between the boundary fixed point and the critical value; (b) there is a smaller box (as in Figure 1(c) of the introduction understood for  $n = 1$ ); or (c) no renormalization is possible, for  $n > 1$ .*

**PROOF.** If there is a fixed point  $p$  between  $a$  and  $c$ , consider the box on  $p$ . If it contains the graph we have a smaller box. Otherwise, the critical value in two iterates lands in the invariant interval  $[(a, p)]$ . In this second case there is no interval about the critical point whose images are interior disjoint and which returns to itself under some iterate.

**COROLLARY.** For a minimal renormalization  $f$  which is further renormalizable,  $f(x) > x$  for  $a < x < c$  (if  $f(c)$  is a maximum).

**PROOF.** By the lemma  $f(x) = x$  for some  $a < x < c$  implies that either there is a smaller box or no further renormalization is possible. The other possibility  $f(x) < x$  for  $a < x < c$  implies the critical value is forward asymptotic to  $a$ , so no further renormalization is possible.

(CONVERSE) **REMARK.** A box for  $f^n$  satisfying  $f^n(x) > x$  for  $a_n < x < c$  has interior disjoint images under the preimages of  $f$  following the critical orbit backwards.

**PROOF.** These preimage boxes define conjugate boxes. These can only overlap at the orientation preserving fixed point.

**5. Composition of roots and the sector theorem.** Let  $[a, b] = I$  be an interval on the real axis. Let  $S = S(a, b)$  be the set of injective holomorphic mappings (schlicht mappings) of  $\mathbb{C} - \{x \text{ real not in } [a, b]\}$  which are homeomorphisms of  $[a, b]$  to itself and preserve the two half planes. Then  $S(a, b)$  contains left and right square roots, branches of  $\sqrt{z}$  followed by linear transformations defined on  $\mathbb{C} - \text{slit}$ , where the slit is a real ray complementary to  $[a, b]$ .

We consider a composition of elements from  $S(a, b)$  of the form

$$A_n C_n \cdots A_2 C_2 A_1 C_1$$

satisfying:

(i)  $A_1, A_2, \dots, A_n$  are left square roots and  $C_1, C_2, \dots$  are general elements of  $S(a, b)$ .

(ii)  $A_1$  has a singularity at  $a$  (i.e., the slit for  $A_i$  is  $(-\infty, a)$ ), and  $a_i =$  singularity of  $A_i$  moves away to the left exponentially fast, i.e., if  $|a-b| = 1$ , then  $k \leq |a_{i+1} - a|(|a_i - a|)^{-1} \leq K$  for  $i = 2, 3, \dots$ ,  $1 < k < K < \infty$ .

(iii) If  $I_i$  denotes the maximal open interval on which  $C_i$  extends to a diffeomorphism into the reals, then  $C_i I_i$  contains  $(a_i, a)$ . Moreover, if  $J'_i = J_i$  plus  $\lambda$ -proportional space on either side, where  $J_i = C_i^{-1}(a_i, a)$ , then  $J'_i \subset I_i$ .

**SECTOR THEOREM.** There is a  $\theta$  depending only on  $(k, K, \lambda)$  so that the image of the upper half plane by the composition  $A_n C_n \cdots A_1 C_1$  is contained in the sector  $0 \leq \arg(z - a) \leq \pi - \theta$  (see Figure 1 on next page).

**PROOF.** (i) The regions of the upper half plane cut out by circles passing through  $a$  and  $b$  are Poincaré metric distance  $R$  neighborhoods of the geodesic  $(a, b)$  in  $\mathbb{C} - \{x \text{ real but not in } (a, b)\}$ . By Schwarz's lemma these are mapped into themselves by any element or any composition from  $S$  (Figure 2 on next page).

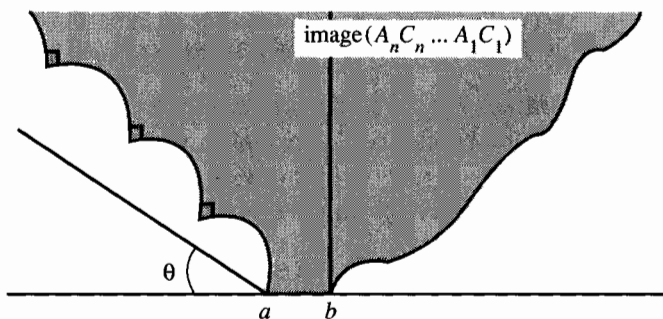


FIGURE 1

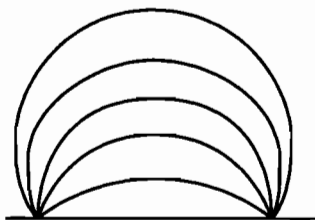


FIGURE 2

(ii) The composition  $C_i$  is injective on  $\mathbb{C} - \{\text{real } x \text{ not in } I_i\}$  (assumption (ii)) and  $J'_i \subset I_i$ . So by Koebe distortion  $C_i$  yields a bounded from linear distortion mapping of a bounded shape neighborhood in the upper half plane  $U_i$  of  $J_i$  to a bounded shape neighborhood in the upper half plane  $V_i$  of  $(a_i, b)$  (Figure 3). Since  $C_i$  fixes the two points  $\{a, b\}$  it only distorts the Euclidean metric on the neighborhood by a bounded factor.

Now we prove the sector theorem. The statement is obvious for  $n = 1$ , so assume  $n \geq 2$ .

Start with any point  $p_1$  in the upper half plane and define  $p_2$  such that  $p_2 = A_1 C_1 p_1, \dots, p_{i+1} = A_i C_i p_i$  for  $i \geq 2$ . First note that  $p_2$  lies to the right or on a vertical line at  $a$ . There are two cases as  $i$  increases:

*Case 1.*  $p_i$  is far from  $(a, b)$  relative to the singularity  $a_i$  of  $A_i$ . Precisely  $p_i$  does not belong to  $U_i$ . Then  $p_{i+1}$  is not in  $A_i C_i U_i = A_i V_i$ . Now  $A_i V_i$  contains all the points of the strip  $\{y > 0, x \text{ in } (A_i a_i, a)\}$  in a rectangle resting on  $(A_i a_i, a)$  with height a definite fraction of the base. Thus, the angle of  $p_i$  as viewed from  $a$  is large and stays large after application of  $A_i C_i$ . This is so because  $A_i C_i p_i = p_{i+1}$  lies to the right of the vertical line at  $A_i a_i$  and above the rectangle at the bottom of the strip (Figure 4).

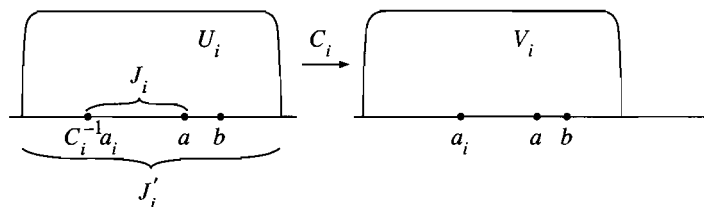


FIGURE 3

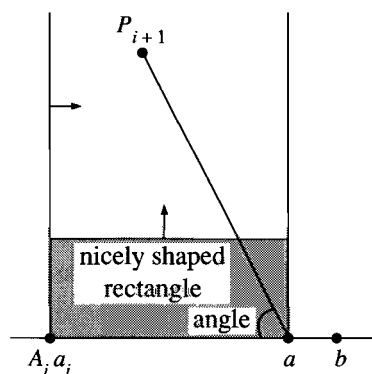


FIGURE 4

*Case 2.* There is a first  $i$  so that  $p_i \in U_i$ . At this point we may assume the angle as viewed from  $a$  is large using Case 1 and the fact that  $p_2$  is to the right or on the vertical line at  $a$ . Then one applies a fixed number  $l$  of the factors  $A_i C_i, A_{i+1} C_{i+1}, \dots$  until  $a_{i+l}$  is much farther away from  $a$  than  $p_{i+l}$ . This happens because of assumption (ii) and statement (i) of the proof in case  $p_i$  lies in the Poincaré ball (Figure 2) of scale  $p_i$ . (The contrary case can be reduced by Koebe and induction to this case, and we leave this for the reader.) During these  $l$  iterations the angle as viewed from  $a$  is only boundedly distorted elementary geometry and the Remark below show. After that, the subsequent factors  $A_j C_j$  for  $j > i + l$  only cause a sequence of distortions decaying geometrically by assumption (ii) (see Remark below). Thus, the angles of  $p_2, p_3, \dots$  remain large in all cases.

**REMARK.** If a holomorphic mapping is schlicht on  $\mathbb{C} - \{x \text{ real but not in an interval } I\}$ , then it has bounded distortion on any region as  $R_0$  in Figure 5 on next page. The constants depend on the geometry of  $R_0$ . Also, it has exponentially small nonlinearity on a region such as  $R_n$  which is exponentially small.

Applying the first remark  $l$  times ( $l$  fixed) yields a bounded distortion in the above paragraph. Applying the second yields the geometric series of distortions used above. Q.E.D.

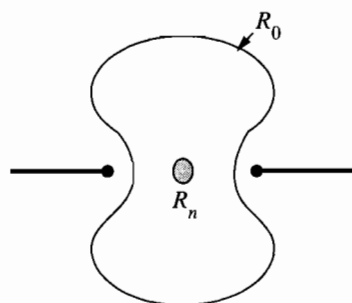


FIGURE 5

**6. The factoring of the sector theorem.** Consider an infinitely renormalizable mapping  $f: I \rightarrow I$  of combinatorial type  $\leq T$  with critical point  $c$  and critical value  $v$ . Let  $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_n, \dots$  be the interval collection at each level. Let  $I_j(c) \in \mathcal{E}_j$  and  $I_j(v) \in \mathcal{E}_j$  denote the intervals containing the critical point  $c$  and critical value  $v$  respectively (Figure 1). Assume  $f = hQ$ , where  $h^{-1}$  has a complex analytic injective extension to  $\mathbb{C} - \{x \text{ real not in } J \supset I\}$  and  $Q$  is a quadratic polynomial, i.e.,  $f$  belongs to the Epstein class (§4).

Consider the basic backwards composition  $f(n)$  from  $I_n(c)$  to  $I_n(v)$  passing through each  $n$ th level interval in  $\mathcal{E}_n$ . Define the *scale of a factor*  $f^{-1}$  of  $f(n)$  to be the largest  $s$  so that its domain interval at level  $n$  belongs to  $I_s(v)$ .

Divide the composition  $f(n)$  into epochs by: epoch  $n-1$  is from the beginning of the composition up to and including the last factor of scale  $n-1$ , epoch  $n-2$  is from there up to and including the last factor of scale  $n-2$ , etc.

In epoch  $j$  mark all the left intervals of scale  $j$ , where a left (right) interval of  $\mathcal{E}_n$  is one dynamically related to  $I_n(v)$  by an orientation reversing (preserving) map. A left (right) root is by definition the part of the factor  $f^{-1}$  starting at a left (right) interval corresponding to  $Q^{-1}$  in the factoring  $f = hQ$ . Let  $C_\alpha$  denote the part of basic composition between two marked left roots.

The backwards composition from  $I_s(c) \rightarrow I_s(v)$  at level  $s$  restricted to an  $n$ th level interval in  $I_s(c)$  is called a *basic map at level  $s$* .

**PROPOSITION.** Suppose the marked left root just after some composition  $C_\alpha$  has scale  $r$ . Then  $C_\alpha$  is a finite composition of basic maps at level  $r$ , right roots at scale  $r$ , and restrictions of  $h^{-1}$ . The number of each is bounded in terms of the bound  $T$  on the combinatorics.

**PROOF. Claim:** The last visit during epoch  $j$  to an  $n$ th level interval in  $I_j(v)$  lands on a left interval. (This useful observation was made by Wellington de Melo.)

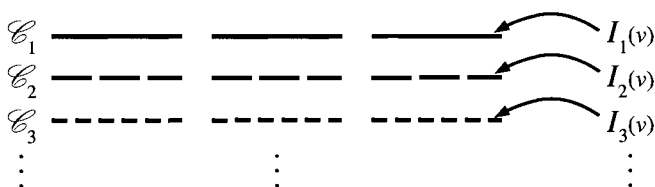


FIGURE 1

**PROOF OF CLAIM.** Consider the  $j$ th renormalization  $f_j$  of  $f$  preserving  $I_j(v)$ . Now  $f_j$  is unimodal so it maps  $I_n(v)$  to the interval  $J_n$  furthest to the left in  $I_j(v)$  by an orientation reversing mapping. Thus,  $J_n$  is a left interval and we see that under the backwards composition starting at  $I_j(c)$  the inverse branches of  $f_j$  run through all the intervals of  $\mathcal{C}_n$  in  $I_j(v)$  arriving last at  $J_n$  and then going on to  $I_n(v)$ . This proves the claim.

Now by construction each epoch  $j$  is decomposed cleanly into basic maps of level  $j$  and single factors  $f^{-1}$  starting at intervals of scale  $j$ . By the claim the  $C_\alpha$  either run between two left factors at scale  $j$  or start just after the last visit to  $I_j(v)$  and run to the first left factor at scale  $j - 1$ . In either case we have the structure required by the proposition. Q.E.D.

**COROLLARY.**  $C_\alpha$  satisfies property (iii) of the sector theorem relative to the immediately following left root.

**PROOF.** We use the above proposition. Let  $a_i$  be the center of the left root following  $C_\alpha$  at scale  $r$ . In §5 we have linearly renormalized all factors to fix  $\{a, b\}$  and  $|a - b| = 1$ . In these terms the right roots of  $C_\alpha$  at scale  $r$  are to the right of  $(a, b)$  at distance commensurable to  $|a_i - a|$  because of bounded geometry of the Cantor set. Also, by the bounded structure of basic maps at level  $r$  these have bounded distortion between intervals containing the dynamic intervals at level  $r$  with space on either side. Thus, the region of control for each basic map covers the interval  $(a_i, a)$  plus space on either side because  $a_i$  is in the dynamic interval at scale  $r$ .

The fixed number of right roots, the fixed number of  $h$  factors, and the fixed number of basic maps only disturb this control a fixed amount. Q.E.D.

Finally we make explicit the connection between the marked left roots here, the  $C_\alpha$  between, and the factoring

$$A_n C_n \cdots A_2 C_2 A_1 C_1$$

of the sector theorem (§5). We let  $A_1$  be the marked left root at scale  $n - 1$  closest to  $I_n(v)$ . We let  $A_2, A_3$ , etc. be the subsequent marked left roots. Finally, the  $C_\alpha$  are the in between compositions, as expected.

**PROPOSITION.** Assume  $f = hQ$  is of the Epstein class. Then the above composition satisfies hypotheses (i), (ii), and (iii) of the sector theorem, at level  $n$ .

REMARK. The dependence of the constants  $(k, K, \lambda)$  on  $f = hQ$  and  $n$  is "beau" as in §3.

PROOF. (i) is true by definition. (iii) is the above proposition. (ii) follows directly from bounded combinatorics and bounded geometry of the Cantor set.

We note only for (ii) that there are always left roots of scale  $j$  in epoch  $j$  (see the proof of the claim above).

Also, bounded combinatorics controls the number of marked left roots in epoch  $j$  because they lie in disjoint intervals at level  $j$ . Q.E.D.

**7. The sector inequality.** Suppose  $c < c' < b' < b < a$  in the reals and  $F$  is a schlicht mapping of a hemidisk  $D$  of radius  $R$  and center  $b'$  to a sector with angles  $(\theta, \pi/2)$  resting on  $(a', b')$  in the upper half plane  $H$ .

Let  $N$  be the geodesic neighborhood of  $(a, c)$  in  $H$  corresponding to the Euclidean disk of radius  $\bar{R}$  whose boundary passes through  $\{a, c\}$  (Figure 1). Assume  $\theta$  is fixed, all nonzero distances between  $\{a, b, c, a', b', c'\}$  are of order 1, and  $F$  carries  $\{a, b, c\}$  to  $\{a', b', c'\}$  in order.

**THEOREM.** For  $\bar{R}$  sufficiently large and  $R/\bar{R}$  sufficiently large compared to  $\bar{R}$ ,  $N$  contains  $F(N)$  plus all the points of the upper half plane within a definite Euclidean distance to  $F(N)$ .

PROOF. (1) Let  $\psi$  be the Riemann mapping of the upper half plane to the sector carrying  $(\infty, a, c)$  to  $(\infty, a', c')$ . Let  $U(R)$  denote  $\psi(D(R))$ , where  $D(R)$  is the hemidisk of radius  $R$  centered at  $b'$ .

(2) If  $U$  is a simply connected domain with arc  $\gamma$  on its boundary and  $F$  is a complex analytic mapping of  $U \rightarrow U$  which is continuous at  $\gamma$  and preserves  $\gamma$ , let  $\tilde{U}$  and  $\tilde{F}$  denote the double of  $U$  and  $F$  along  $\gamma$ . Then  $\gamma$  is a geodesic in the Poincaré metric of  $\tilde{U}$  preserved by  $\tilde{F}$ . By Schwarz's lemma  $\tilde{F}$  preserves the neighborhoods of  $\gamma$  of Poincaré distance  $\leq c$  for all  $c > 0$ .

Thus,  $F$  preserves these regions intersect  $U$ . Abusing language we call them Poincaré geodesic neighborhoods of  $\gamma$  in  $U$  (see Figure 2).

(3) The Riemann map  $\psi$ , outside a fixed neighborhood of the corners  $(a, b, c)$  and  $(a', b', c')$ , is the composition of a fractional power of  $z$ ,  $z^\alpha$ , and a mapping of uniformly bounded distortion of the Euclidean metric.

We will comment on the geometry ignoring this map of bounded distortion. For example,  $U(R)$  is the intersection of the sector with the hemidisk of center  $b'$  and radius  $R^\alpha$  (here  $\alpha = 1/2 - \theta/\pi$ ).

(4) Our schlicht mapping  $F$  is then the composition of this power law and some schlicht mapping  $G$  of  $U(R)$  into the sector which fixes  $\{a', c'\}$ .

(5) Let  $\gamma$  be the arc of the boundary of the sector from  $a'$  to  $c'$ . Then by the argument of (2)  $G$  takes the Poincaré geodesic neighborhoods of  $\gamma$

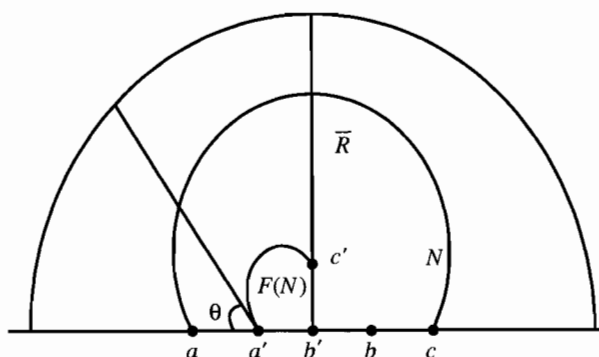


FIGURE 1

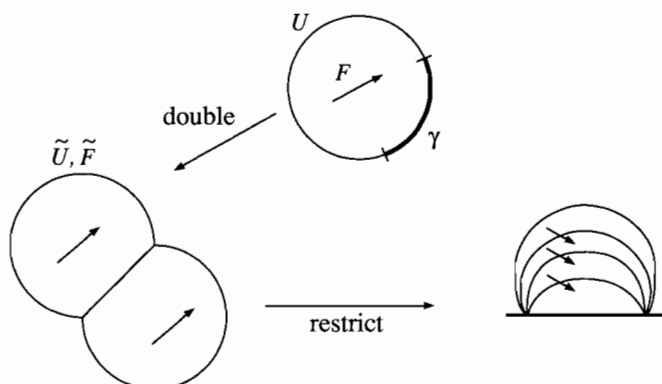


FIGURE 2

in  $U(R)$  into the Poincaré geodesic neighborhoods of  $\gamma$  in the sector  $S$  preserving or decreasing “distance to  $\gamma$ .”

(6) The boundary of  $S$  and boundary of  $U(R)$  only start to differ at Euclidean distance  $R^\alpha$  from  $\gamma$ . Thus, the geodesics much closer to  $\gamma$  than  $R^\alpha$  are about the same for  $U(R)$  or for  $S$ . So  $G$  does not move these neighborhoods too much. (To see this one can use the fact that the Poincaré metric of a simply connected plane domain is comparable to the Euclidean metric times the reciprocal of the Euclidean distance to the boundary.)

(7) Now a fairly large geodesic neighborhood  $N$  of  $(c, a)$  in the upper half plane is carried by  $\psi$  well within the sector (using (3) and our lower bound in the angle  $\theta$ ) to a geodesic neighborhood of  $\gamma = (c', a')$  in the sector. By (6)  $G$  does not move it very much. Thus, the composition  $F$  does what we want in the geodesic neighborhood  $N$  corresponding to the circle of radius  $R$  so large that the power law beats the bounded distortion part of (3). Q.E.D.



*Modified sector inequality.* Replace the hemidisk  $D$  of radius  $R$  in the previous section by the largest "Poincaré neighborhood"  $M$  of  $(c, a)$  in the upper half plane contained in  $D$  (see the above proof). Suppose  $F$  is a schlicht mapping of  $M$  into the sector which is continuous on  $(c, a)$  and carries  $\{c, b, a\}$  in order to  $\{c', b', a'\}$ . Let  $N$  as above be the Poincaré neighborhood of  $(c, a)$  of Euclidean diameter  $\bar{R}$ .

**MODIFIED THEOREM.** *For  $\bar{R}$  large and  $R/\bar{R}$  sufficiently large compared to  $\bar{R}$ ,  $N$  contains  $F(N)$  plus all the points of the upper half plane within a definite Euclidean distance to  $F(N)$ .*

**PROOF.** Let  $\infty \in \partial M$  be the highest point of  $M$ . Let  $\bar{\psi}$  be the Riemann mapping of  $M$  to the sector carrying  $\{\infty, c, a\}$  in order to  $\{\infty, c', a'\}$ . As  $R$  approaches  $\infty$ ,  $\bar{\psi}$  approaches the Riemann mapping  $\psi$  of the previous proof uniformly on  $N$  since  $\bar{R}$  is fixed. Then the proof above can be modified by continuity considerations to work here.

**8. The complex quadratic-like mapping produced by renormalization.** Let us work with renormalizable mappings  $f = hQ: I \rightarrow I$  of combinatorics  $\leq T$  of the Epstein class  $E(J)$ ,  $I \subset J$ . After some renormalization we can assume  $J$  contains a definite neighborhood of  $I$  and that the real bounds on the critical orbit hold §§3, 4.

**THEOREM.** *For any  $n > N(T)$  the  $n$ th renormalization  $g = R^n f$  has a complex analytic extension  $G$  to some disk  $D \subset \mathbb{C}$  so that  $D \rightarrow G(D)$  is proper of degree two and  $G(D) - D$  has conformal modulus  $> m(T) \geq 0$  (see Figure 1).*

**PROOF.** Let  $f(n)$  denote the backwards branch going from the critical point interval  $I_n \in \mathcal{E}_n$  to the critical value interval  $I_n v \in \mathcal{E}_n$ .

Let  $(a, c)$  denote the maximal interval where  $f(n)$  is a diffeomorphism into the reals and let  $(\bar{a}, \bar{c}) = f(n)(a, c)$ . Let  $\{\pm a, \pm(c')\}$  be  $f^{-1}\{a, c\}$ .

By the sector theorem and §6 the image by  $f(n)$  of the upper half plane is in the sector  $0 \leq \arg(z - \bar{a}) \leq \pi - \theta(T)$  since we have arranged that  $J$  contains a definite neighborhood of  $I$  and the real bounds on the critical orbit hold.

Now apply two branches of  $f^{-1}$  to the sector (obtaining Figure 3(a), and (b)) applied to any geodesic disk on  $(a, c)$  of radius less than that of one contained in the region of bounded nonlinearity of  $h^{-1}$  (e.g.  $\leq$  scale of  $J$ ).

There are several points to make.

(1) It is simple to see that  $g_+^{-1}M$  in Figure 3(a) lies in a sector  $\pi/2 \leq$

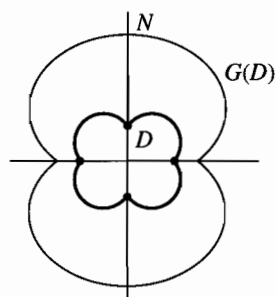


FIGURE 1

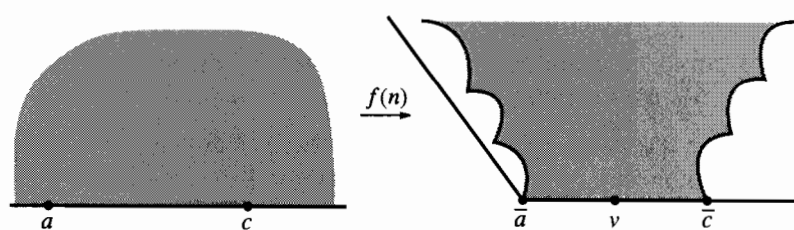


FIGURE 2

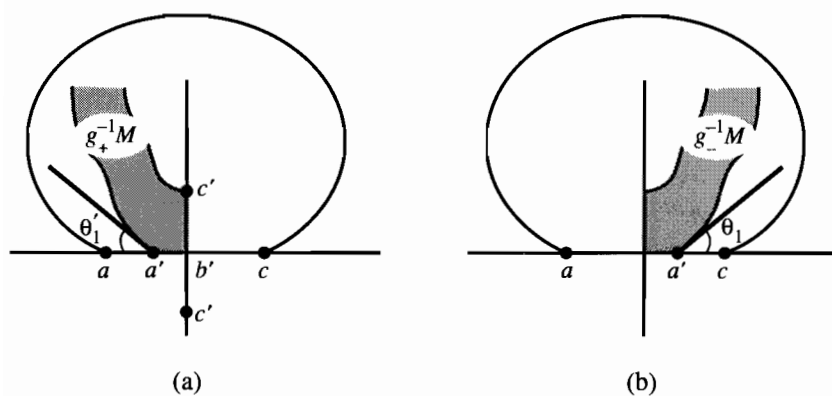


FIGURE 3

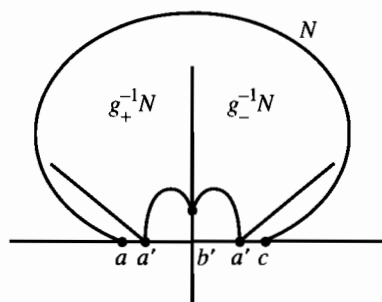


FIGURE 4

$\arg(z - a') \leq \pi - \theta_1(\theta_1 < \theta)$  because we have only applied a bounded distortion map  $h^{-1}$  to a piece of the sector in Figure 2 on previous page and then a right root. Here  $g_+^{-1} = f_+^{-1}f(n)$ .

(2) The same remark applies to  $f_-^{-1}f(n) = g_-^{-1}$  and  $g_-^{-1}(M)$  (Figure 3(b)).

(3) By the proposition below,  $\text{distance}(c, b')$  and  $\text{distance}(a, b')$  are each a definite factor greater than  $\text{distance}(a', b')$ , where  $b'$  is the critical point of  $R^n f$ .

(4) If  $n(T)$  is large enough, then  $M$  can be taken large relative to  $N$  in the sector inequality for some  $N \subset M$  with  $N$  large relative to  $(a, c)$ . This yields Figure 4 on previous page.

Now we reflect across the  $x$  axis to obtain Figure 1 and the result. Q.E.D.

**PROPOSITION.** *Distance( $a, b'$ ) and distance( $b', c$ ) are each greater than distance( $a', b'$ ) by a definite factor.*

**PROOF.** We use what de Melo calls the basic fact,  $(a, c)$  contains the critical point interval  $I_n \in \mathcal{C}_n$  and its two immediate neighbors. Call these the *three intervals*. Also  $a'$  is a critical point of  $f^{q_n}$ , where  $R^n f = f^{q_n}/I_n$  which is closest to  $b'$ . Thus,  $a'$  lies in one of the preimages of  $I_n$  by  $f^{q_n}$  which is closest to  $I_n$ . The proposition follows from the bounded geometry of the *three intervals* and their critical points. This follows using §3 in the manner of the self-contained first paragraph of §15.

**9. Douady-Hubbard theory and Riemann surface laminations.** Consider the set  $\mathbb{Q}_\mathbb{C}$  of complex quadratic-like mappings  $F: D \rightarrow FD$  with connected invariant set  $K_F = \bigcap F^{-n}D$  up to  $\mathbb{C}$ -analytic conjugacy near  $K_F$ . Two  $F, G$  are  $h$ -equivalent [DH1] if they are qc conjugate near  $K_F$  and  $K_G$  by a map which has no conformal distortion (i.e.,  $\bar{\partial}$  map = 0, a.e.) between  $K_F$  and  $K_G$  a.e. Lebesgue.

**THEOREM (Douady-Hubbard).** (a) *The quadratic polynomials with connected invariant set cut each  $h$ -equivalence class in one point.*

(b) *Each  $h$ -equivalence class is bijectively equivalent to the set of real analytic degree two expanding maps of  $S^1$  up to real analytic equivalence.*

**REMARK.** The quadratic polynomial in part (a) is called the internal class of  $q \in \mathbb{Q}_\mathbb{C}$  while the expanding mapping in part (b) is called the external class of  $q \in \mathbb{Q}_\mathbb{C}$ . Thus the internal class is a point on the Mandelbrot set. We refer to the various “submanifolds” of  $\mathbb{Q}_\mathbb{C}$  with constant internal class, i.e., constant label in the Mandelbrot set, as the *prestable manifolds* of renormalization. In any  $\mathbb{C}$ -analytic family, fixing the internal class defines a  $\mathbb{C}$ -analytic subspace [DH1]. Therefore in the context of *symmetric*  $\mathbb{C}$ -analytic mappings the stable manifolds of renormalization (Theorems 2 and 2') have at least a real analytic structure.

Now we associate to any *smooth* degree two expanding map  $f$  of the circle

( $f' > 1$ ,  $f'$  Hölder) a Riemann surface lamination (appendix) whose point in Teichmüller space determines the smooth conjugacy class of  $f$ . For real analytic  $f$  the point determines the external class.

Form the inverse limit space, a dyadic solenoid,

$$\bar{S} = \varprojlim \{\cdots \rightarrow S \xrightarrow{f} S \xrightarrow{f} S\}$$

with induced mapping  $\tilde{f} = \varprojlim \{f\}$ . By the theorem below leaves of  $\tilde{S}$  carry unique affine structures compatible with their smooth structure so that  $\tilde{f}$  becomes affine. Now attach upper half spaces of the leaves of  $\tilde{S}$  using these affine structures to obtain a Riemann surface lamination  $\tilde{L}$ . Then  $\tilde{f}$  on  $\tilde{S}$  extends affinely to a holomorphic mapping  $\tilde{F}: \tilde{L} \rightarrow \tilde{L}$ . Remove the boundary and quotient by the group generated by  $\tilde{F}$  to obtain  $L_f = \tilde{L}/\{\tilde{F}\}$ .

The Riemann surface lamination  $L_f$  up to Teichmüller equivalence (appendix) remembers the conformal structure on leaves up to qc isotopy. Lifting the  $\tilde{L}$  and recalling  $\tilde{F}^{-1}$  is contracting we see the qc isotopy converging to the identity on the  $\partial$  solenoid. Thus, the affine structure on the leaves is determined by  $L_f$  up to Teichmüller equivalence. (The unique field of affine structures on the dynamic solenoids  $(\tilde{S}, \tilde{f})$  are continuously varying. Thus, the structure on even one leaf determines the field of structures.)

The eigenvalues of  $f$  can be read off from the affine expansion factors of the  $\tilde{f}$  periodic leaves. The eigenvalues of  $f$  determine the sizes (up to a bounded factor) of the intervals in the  $n$ th level Markov grids  $f^{-n}$  (fixed point of  $f$ ). The Markov grids for  $f$  and  $g$  determine a unique conjugacy  $h$  which is quasimetric in general and obviously Lipschitz if the corresponding sizes are in bounded ratio. A Lipschitz conjugacy  $h$  between  $f$  and  $g$  is differentiable at most points. It follows that  $h'$  exists and is Hölder using the approximate formula  $h \sim g^n Df^{-n}$  when  $f^{-n}$  is chosen to converge to a point of differentiability of  $h$ . When  $f$  and  $g$  are real analytic, domains of analyticity are easily controlled and  $h$  is seen to be real analytic.

Here is the discussion for the theorem below. The natural projection  $\tilde{S} \xrightarrow{\pi} S$  is used to induce smooth structures and smooth Riemannian metrics in the one-dimensional leaves of  $\tilde{S}$ . The induced mapping  $\tilde{f}$  on  $\tilde{S}$  then inherits the same geometric structure as  $f$  had. In particular,  $\tilde{g} = \tilde{f}^{-1}$  is uniformly contracting on leaves. To construct a leaf translation  $x$  to  $y$  on the same leaf, approximately, we apply  $g^n$  until these points become  $\bar{x}$  and  $\bar{y}$ , very close together. We translate  $\bar{x}$  to  $\bar{y}$  using the smooth structure and transport this back by  $g^n$ . The limit of this construction defines the translation pseudogroup of the affine structure on each leaf. It is invariant by construction. The construction also shows any invariant structure must be this one. The field of translation pseudogroups varies continuously because we have an a priori estimate on the error of the  $g^n$  approximation. This proves

**THEOREM** (global linearization). *Each leaf of  $\tilde{S}$  carries a translation pseudogroup or affine structure compatible with its smooth structure so that  $\tilde{f}$  between leaves is affine. The affine translations in the leaves vary continuously in the transverse direction, and they are uniquely specified by continuity and  $\tilde{f}$  invariance.*

**COROLLARY.** *The external classes  $\{f\}$  can be embedded  $f \rightarrow L_f$  in the Teichmüller space of a surface lamination (appendix).*

**REMARK.** In the real analytic case there is a direct construction of  $L_f$  and the linearization in terms of a holomorphic extension  $F$  of  $f$  to some neighborhood  $F^{-1}U$  of  $\{|z|=1\}$  in  $\{|z|>1\}$ . It is:

Form  $\tilde{V} = \varprojlim \{\cdots F^{-2}U \rightarrow F^{-1}U \rightarrow U\}$  and  $\tilde{F}: \tilde{W} \rightarrow \tilde{V}$ , where  $\tilde{W} = \varprojlim \{\cdots F^{-2}U \rightarrow F^{-1}U\}$ . This construction of passing to the inverse limit converts  $F$ , which is neither globally defined on its image or injective, into  $\tilde{F}$ , which is injective but still not globally defined on its image. So form  $G = \tilde{F}^{-1}: \tilde{V} \rightarrow \tilde{V}$ , which is globally defined and injective but not onto. Now make  $G$  onto by passing to the direct limit  $\tilde{U} \rightarrow \varinjlim \{\tilde{V} \rightarrow \tilde{V} \rightarrow \cdots\}$  and  $\tilde{G} = \varinjlim G$  to obtain the bijection  $\tilde{G}: \tilde{U} \rightarrow \tilde{U}$ . (The direct limit step was suggested by Lyubich.)

The leaves of the Riemann surface lamination  $\tilde{U}$  are upper half planes permuted holomorphically by  $\tilde{G}$ . This proves anew the linearization theorem. The quotient of  $\tilde{U}$  by the group generated by  $\tilde{G}$  is the orbit lamination  $L_f$ .

**10. The modulus function on external classes.** Let  $m(x)$  denote the supremum of the conformal modulus of  $N$  for some representative  $F: F^{-1}N \rightarrow N$ , where  $x$  denotes the  $\mathbb{C}$ -analytic equivalence class of germs of degree two holomorphic mappings of a neighborhood of  $\{|z|=1\}$  in  $\{|z|\geq 1\}$ . Let  $\text{distance}(x, y)$  denote the infimum of qc distortion of conjugacies between representatives on some neighborhood of  $\{|z|=1\}$  in  $\{|z|\geq 1\}$ .

**THEOREM.** *If  $m(x) = \infty$ , then  $x$  is represented by  $z \rightarrow z^2$ . If  $x$  and  $y$  satisfy  $m(x) \geq \varepsilon$  and  $m(y) \geq \varepsilon$  then  $\text{distance}(x, y) \leq d(\varepsilon) < \infty$ .*

**PROOF.** Choose coordinates so  $N$  is a standard annulus. The modulus of  $F^{-1}N$  is  $\frac{1}{2}$  modulus  $N$ . Thus  $F^{-1}N$  contains a definite neighborhood of  $|z|=1$  in  $|z|\geq 1$ . We can apply Koebe distortion to control the nonlinearity of  $F$ . Then by replacing  $N$  by a smaller concentric annulus  $N'$  we have on  $F^{-1}(N')$  a completely controlled analytic mapping. In particular, the geometry of the glued fundamental domain is controlled. Thus, for two such maps we can construct by pull back a quasiconformal conjugacy (§11) with controlled distortion. This proves the second statement. The first statement is standard.

**11. Thurston equivalences and the pull back conjugacy.** Start with two complex quadratic-like mappings  $F: D \rightarrow F(D)$  and  $G: D \rightarrow G(D)$ . A *Thurston equivalence* between  $F$  and  $G$  is a certain kind of homotopy class of pairs

$$(X_F, C_F) \xrightarrow{H} (X_G, C_G),$$

where  $X_F$  is a contractible region containing the positive critical orbit  $C_F = \{1, 2, 3, \dots\}$  of  $F$ ,  $X_G$  is a contractible region containing the positive critical orbit  $\{1, 2, 3, \dots\}$  of  $G$ ,  $H(1) = 1$ ,  $H(2) = 2$ , etc. and the phrase "homotopy class" is defined by isotopies of the contractible regions  $X_F$ ,  $X_G$  fixing  $\{1, 2, 3, \dots\}$  and homotopies of restrictions of  $H$ 's to common smaller regions also fixing  $\{1, 2, 3, \dots\}$ .

To define which homotopy classes are Thurston equivalences consider Figure 1.

Start with a homotopy class  $H_0$  and lift it through the branch cover so that 1 goes to 1. We assume (a) the lift  $H_1$  carries 2 to 2, 3 to 3, etc. and (b) the homotopy class of  $H_1$  equals the homotopy class of  $H_0$  (rel  $\{1, 2, 3, \dots\}$ ). In (b) we have used the (dynamic) point that the covering spaces are subsets of their respective bases. Note also that property (a) only depends on the homotopy class of  $H_0$ . In effect we have defined a *Thurston map on homotopy classes*  $H_0 \mapsto H_1$  and a *Thurston equivalence* is a homotopy class which is fixed by this Thurston map.

For the *pull back conjugacy theorem* we start with a certain representative of a Thurston equivalence  $H_0: F(D) \rightarrow G(D)$  which (a) is also a conjugacy between the maps  $F: \partial D \rightarrow \partial F(D)$  and  $G: \partial D \rightarrow \partial G(D)$  and (b) is a quasiconformal homeomorphism.

Then  $H_0: F(D) \rightarrow G(D)$  and the pull back  $H_1: D \rightarrow D$  restrict to maps  $\partial H_i: \partial D \rightarrow \partial D$  satisfying  $G \cdot \partial H_i = \partial H_0 \cdot F$ ,  $i = 0, 1$ , so  $\partial H_0 = \partial H_1$  or  $\partial H_0 = \partial H_1 \cdot \tau$ , where  $\tau$  is the involution of the double cover  $F: \partial D \rightarrow \partial G(D)$ . If we lift the homotopy asserting  $H_0$  and  $H_1$  have the same homotopy class rel  $\{1, 2, 3, \dots\}$ , then we see the first possibility holds (Figure 2 on next page), i.e.,  $\partial H_0 = \partial H_1: \partial D \rightarrow \partial D$ .

Thus, we may add to  $H_1$  the restriction of  $H_0$  to the outer annulus  $F(D) - D$  to get an extension of  $H_1$  to a homeomorphism between  $F(D)$  and  $G(D)$ . This homeomorphism, called  $H_1: F(D) \rightarrow G(D)$ , will be quasiconformal if we make the technical assumption that  $\partial D$  in  $f(D)$  is a quasicircle. Its

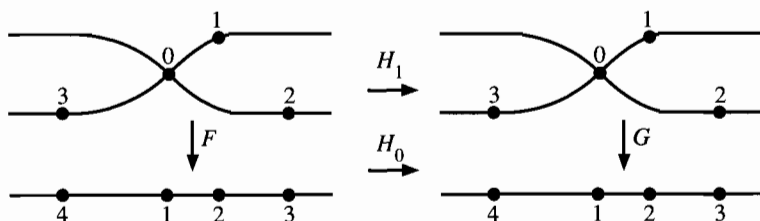


FIGURE 1

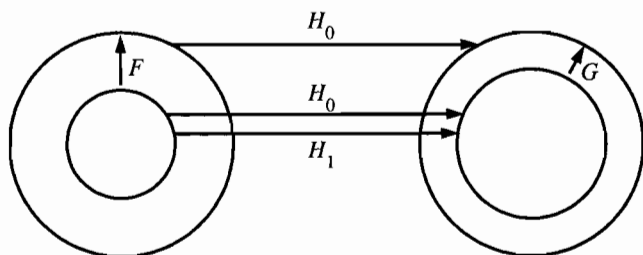


FIGURE 2

distortion will be the same as that of  $H_0$  because it is the union of a piece of  $H_0$  (on  $F(D) - D$ ) and a “complex analytic conjugate” of  $H_0$  (since  $GH_1 = H_1F$  on  $D$ ). (This step was suggested by Curt McMullen.)

We iterate the process to construct an infinite sequence of  $K$ -quasiconformal homeomorphisms with fixed values on  $\{1, 2, 3, \dots\}$  and not changing on the outer rings  $F(D) - D$ , then  $D - F^{-1}D$ , then  $F^{-1}D - F^{-2}D$ , etc. By *quasiconformality*, we can extract convergent subsequences on all of  $F(D)$ . The limit actually exists on the union of the outer rings. So if this union is dense all limits over subsequences are equal and equal to the continuous extension of what is on the outer rings. If the union is not dense there is some ambiguity in the limit on the interior of the invariant set. This interior is classified in types by Sullivan [S5] and homotopies to conjugacies and discussed in Mane-Sad-Sullivan [MSS] and McMullen-Sullivan [MS] (see also [ST]). We continue this discussion in the more rigid case when there is no interior and the limiting map is unique and is a conjugacy.

The conformal distortion of the limiting conjugacy depends on the distortion of  $H_0$  on the fundamental domain and the distortion on the Julia set. The latter may be zero by construction as in §14 where internal classes are discussed. Or it may be zero because the Julia set may not support a new measurable invariant conformal structure. An example of the latter is the infinitely renormalizable bounded type symmetric complex quadratic-like mapping (Theorem 6, §15). We express this last property by there is no “measurable invariant line on the Julia set” or “no measurable line field.” All the above amounts to the

**THEOREM.** *Suppose  $F$  and  $G$  are complex quadratic-like mappings which are Thurston equivalent (along their positive critical orbits) via a quasiconformal homeomorphism. Then if  $\partial D \subset F(D)$  and  $\partial D \subset G(D)$  are quasicircles there is a quasiconformal conjugacy  $H: F(D) \rightarrow G(D)$  between  $F$  and  $G$ . The conformal distortion of  $H$  depends only on the pairs  $(F(D) - D, F/\partial D)$  and  $(G(D) - D, G/\partial D)$  if the invariant set has no interior and the Julia set has no “measurable line field.”*

**REMARK.** Otherwise the distortion depends also on the construction of the conjugacy on the invariant set. (a) In case there is no interior for the invariant

set but the Julia set has a “measurable line field,” the construction is canonical and leads to a specific distortion on the Julia set parametrized by a complex number in the unit disk (see [MSS] for more details, however conjecturally this case does not exist). (b) In case there is interior there must be either (i) a super attracting cycle, (ii) an attracting cycle, (iii) an indifferent periodic point, or (iv) a Siegel disk [S5]. In cases (i) and (iii), the construction can be made so that the limit has no conformal distortion on the filled in Julia set [MSS]. In case (ii) the distortion required is essentially  $|\log \lambda_1 / \lambda_2|$ , where  $\lambda_i$  are the eigenvalues at the attracting cycle [MSS]. In case (iv) it is reasonable to think the Thurston equivalence makes the eigenvalues equal and then the construction can be made with no distortion on the filled in Julia set (see [MSS, ST, S5]). This point does not concern us in this paper where we study mappings symmetric about the real axis.

PROOF. Besides the discussion before the statement of the theorem we need to add that a quasiconformal homeomorphism representing a Thurston equivalence can be restricted, deformed by an isotopy, and extended to fit with a quasiconformal equivalence between the pairs  $(F(D) - D, F/\partial D)$  and  $(G(D) - D, G/\partial D)$  to satisfy (a) and (b) of the second paragraph of this section. This is elementary planar topology. Q.E.D.

**12. Renormalization of complex quadratic-like mappings.** One says a complex quadratic-like mapping  $F$  with connected invariant set is *renormalizable* if there is an  $n$  and a disk  $D$  containing the critical point so that  $F^n: D \rightarrow F^n D$  is complex quadratic-like with connected invariant set. Let  $RM \subset M$  be the subset of points of the Mandelbrot set  $M$  which have representative quadratic-like mappings which are renormalizable. Recall the Douady-Hubbard theorem (§9) that germ equivalence classes of quadratic-like mappings are isomorphic to  $M \times T = \{\text{internal class, external class}\}$ . We can define a renormalization operator  $R$  on representatives by  $F \rightarrow F^n/D$ , where  $n$  is minimal.

**THEOREM.** *Renormalization defines a mapping  $RM \times T \xrightarrow{R} M \times T$  respecting the prestable manifolds  $\{\text{pt} \times T\}$ , namely  $R$  equals the union over  $RM$  of mappings  $\{m, T\} \xrightarrow{R_m} \{R(m), T\}$  for an induced surjective operator  $RM \xrightarrow{R} M$ . The individual mappings  $R_m$  are induced by mappings on the spaces of special conformal structures preserving special Beltrami paths.*

**DEFINITION.** A *special Beltrami path* is one coming from a Beltrami path of invariant conformal structures outside the filled in Julia set. The latter are the *special conformal structures*.

PROOF. The theorem follows, once  $m \in M$  is fixed, from the definitions and the picture of renormalization, as merely the restriction of the variable conformal structure exterior to the Julia set of  $f$  to the exterior of the Julia set of  $Rf = f^n/D$ .



*Addendum.* (Renormalization of vector fields on the laminations.) A backward orbit of  $G$ , the renormalized mapping of  $F$ , extends in a natural way to a backward orbit of  $F$ . Thus we have an embedding of  $L_G$ , the lamination of  $G$ , minus the contribution  $\tilde{K}_F$  coming from the germ of  $K_F$  near  $K_G$ , in  $L_F$ , the lamination of  $F$ . Thus a continuous vector field on  $L_F$  with  $\bar{\partial}$  in  $L^\infty$  can be restricted to obtain one on  $L_G - \tilde{K}_F$  which has  $\bar{\partial}$  in  $L^\infty$  and is uniformly bounded in the  $P$ -metric coming from (nghd of  $L_G - \tilde{K}_F$ ). Extending by zero on  $\tilde{K}_F$  we obtain a continuous vector field on  $L_G$  with  $\bar{\partial}$  in  $L^\infty$ . (This may be shown most easily using the equivalence  $(\bar{\partial} \text{ in } L^\infty) \sim (\text{Zygmund in conformal metric})$ .)

**13. Teichmüller contraction of renormalization for symmetric complex quadratic-like mappings.** If a real analytic folding mapping  $f$  has a complex quadratic-like extension we have two notions of renormalization—real renormalization (of the introduction and §§1–4) and complex renormalization (of §12). The following theorem is proved below.

**THEOREM 13.1.** *These two notions are compatible in the sense that whenever real renormalization is possible for  $f$  for some  $n$ , then for the same  $n$  so is complex renormalization for the complex extension  $F$  of  $f$ . Moreover, the complex extension of  $Rf$  is a quadratic-like mapping which is  $\mathbb{C}$ -analytically equivalent to  $RF$  on neighborhoods of their respective connected invariant sets.*

Let  $f$  have a complex quadratic-like extension  $F: D \rightarrow F(D)$  so that the conformal modulus of  $D - F(D) \geq \varepsilon > 0$ . Suppose  $f$  admits  $n \geq n(T, \varepsilon, l)$  renormalizations of return times  $\leq T$  for a certain function  $n(T, \varepsilon, l)$ , where  $l$  is chosen below. Let  $\mu$  be a Beltrami coefficient defined on  $F(D) - D$  which is symmetric about the real axis and let  $|\mu|_T$  denote the Teichmüller length of  $\mu = \sup_{|\phi|=1} \int \phi \mu$  (see appendix). Let  $l'$  be determined by the universal bounds of §§8 and 10. Choose  $l > l'$  and let  $\lambda(l, l') < 1$  be determined by the Grötzsch inequality (see appendix). Then we have renormalization qua Teichmüller contraction.

**THEOREM 13.2.**  $|R^n \mu|_T \leq \lambda(l, l') |\mu|_T$  for  $n \geq n(T, \varepsilon, l)$ .

**COROLLARY.** *Under bounded return time renormalization of symmetric Mandelbrot internal classes, the Teichmüller distance between external classes decreases exponentially fast.*

**PROOF OF THEOREM 13.1.** Now suppose for some  $n$ ,  $g^n$  is renormalizable on the real axis, i.e., there is the little box on the diagonal of Figure 1(c) of the introduction which encloses the graph of  $Rg$ . Consider the connected component  $D_R$  in  $G^{-n}U$  of the critical point in the complement of  $G^{-2}\gamma, G^{-3}\gamma, \dots, G^{-n}\gamma$ . Then by construction  $G^n$  has one critical point inside  $D_R$ . The intersection  $J_R$  of  $D_R$  with the real axis is the interval

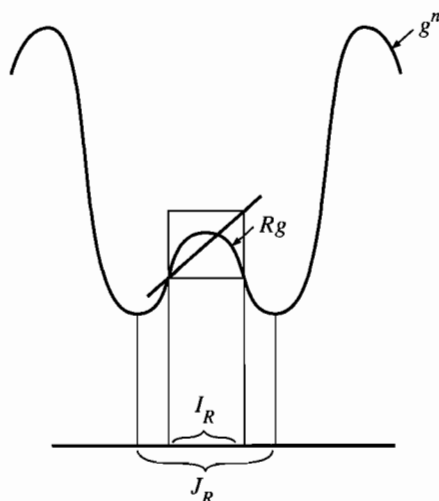


FIGURE 1

between the next two critical points of  $g^n$  out from the central one again by construction: thus, it contains the dynamic interval  $I_R$  of  $Rg$  (Figure 1).

The image of  $D_R$  by  $g^n$  is  $U = \mathbb{C} - \{x \text{ real not in } J\}$  further slit in from either side by the images under  $G^n$  of the arcs  $G^{-2}\gamma, G^{-3}\gamma, \dots, G^{-n}\gamma$ . These are  $\gamma, g\gamma, \dots, g^{n-2}\gamma$ . By the discussion of §8, based on point (2) of the proof of Theorem 1 in §3, the unslit interval covers the small dynamic interval  $[R_g(c), (R_g)^2c]$  and its two immediate neighbors in the  $n$ th level collection of small dynamic intervals (see Figure 3, §3). And these intervals cover  $J_R$  as discussed in §8. Thus,  $G^n: D_R \rightarrow (U \text{ with further slits})$  is a quadratic-like mapping—in the 4-fold symmetric Epstein form.

**PROOF OF THEOREM 13.2.** We choose an extremal  $\tilde{\mu}$ , representing the same tangent vector to  $L$  as  $\mu$  defined on the same fundamental domain (see Remark below). We deform along a Beltrami path a distance  $l$  greater than  $l'$  (see below). We may assume  $\tilde{\mu}$  is also symmetric about the real axis. Then the two  $\mathbb{C}$ -analytic systems at the endpoints have a definite modulus (§10). So there is a definite number of renormalizations required so that the modulus is greater than the universal constant of §8. Then the qc distance is at most a certain constant, call it  $l'$  (§10). Now we apply the almost geodesic lemma (appendix) to see that the renormalized tangent vector is reduced in Teichmüller length by a universal factor  $\lambda(l, l') < 1$ .

**PROOF OF COROLLARY.** Choose a Beltrami path between two real analytic external classes. The tangent vectors along these paths are represented by  $\mu$ 's to which Theorem 13.2 applies. As we renormalize this continues to be true by the complex bounds (§8). Thus, the Teichmüller arc length of this path decreases exponentially by integrating the inequalities of Theorem 13.2.

REMARK. If a tangent vector to  $T(L_f)$  is defined by an  $F$ -invariant  $\mu$  on a neighborhood  $U$  of the circle, we can push forward the  $\tilde{F}$  invariant holomorphic quadratic differentials on  $\tilde{L}$  to obtain holomorphic quadratic differentials  $\varphi$  on  $U$  satisfying  $F_*\varphi = \varphi$ .

The holomorphic invariants  $(\varphi, \mu)$  are just  $\int \varphi \mu$  on any fundamental domain of  $F$ . By Hahn-Banach there is a  $\tilde{\mu}$  on this domain with  $L^\infty$ -norm  $\tilde{\mu} = \sup(\varphi, \mu)$ . This is the definition of an extremal  $\tilde{\mu}$ . By the appendix there is a vector field  $V$  on the lamination so that  $\bar{\partial}V$  is  $\mu - \tilde{\mu}$  pulled up to the lamination. We can renormalize  $V$  (see Addendum §12) to see that  $R^n\mu$  and  $R^n\tilde{\mu}$  pulled to the appropriate lamination differ by  $\bar{\partial}$  (vector field). Thus they continue to have the same  $T$ -norm by the integration by parts formula (appendix). Alternatively (McMullen), we can see they continue to have the same  $T$ -norm by pushing forward quadratic differentials using the map of the Addendum §12.

**14. Proof of Theorem 2'.** Let  $\{f_n\}$  be an inverse chain related by renormalization,  $\cdots \rightarrow f_n \xrightarrow{\sigma_n} f_{n-1} \rightarrow \cdots \rightarrow f_2 \xrightarrow{\sigma_2} f_1 \xrightarrow{\sigma_1} f_0$ , where  $\sigma_i$  is bounded in size by  $T$ . Then if all the  $f_i$  are smooth quadratic-like mappings which are uniformly bounded, by §4 they are all Epstein and by §8 they are all complex quadratic-like with a definite modulus (§10). Let  $c(f)$  denote the internal class of Douady-Hubbard (§9). Suppose  $\{g_n\}$  is another such inverse chain with the same combinatorics and suppose  $c(f_0) = c(g_0)$ . We want to show  $\{f_n\} = \{g_n\}$  as complex analytic mappings up to affine rescaling.

$$(1) \quad c(f_n) = c(g_n).$$

PROOF OF (1). Start with a quasiconformal conjugacy between  $f_0$  and  $g_0$  expressing the fact that  $c(f_0) = c(g_0)$  (§11). By a finite construction on the real axis then the two half planes, this can be promoted to a qc homotopy conjugacy between  $f_1$  and  $g_1$  which is a conjugacy between the forward critical orbits of  $f_1$  and  $g_1$  (§11).

Now perform the pull back conjugacy construction of §11 to obtain a quasiconformal conjugacy between  $f_1$  and  $g_1$  which is a.e. conformal on the saturation of the filled in Julia sets of  $f_0$  and  $g_0$ . The measure of what is left in the filled in Julia set of  $f_1$  or  $g_1$  is zero (see Remark below). Thus,  $c(f_1) = c(g_1)$ . Continuing in this way  $c(f_n) = c(g_n)$ .

(2) Now work in the topology of uniform convergence on a definite neighborhood of the dynamic interval where all maps have the Epstein form and are normalized to have the same dynamic intervals. We will choose this neighborhood once and for all to include the Julia sets of all the maps appearing in inverse chains. If this were not possible the modulus bound on the annuli (§8) would be violated.

The closure  $K$  of these maps in this topology is compact because they are actually bounded in the space defined by sup norm on a larger neighborhood.

It is also clear that the set of elements of bounded inverse chains is closed.

(3) By Theorem 13.2 the Teichmüller distance  $d$  between  $f_n$  and  $g_n$  must be zero.

(4) Let  $(F, G)$  be a limit point of  $(f_n, g_n)$ . Then  $F, G$  have the same internal class. This continuity of the internal class is proven in [DH1]. Since  $F$  and  $G$  are members of inverse chains we must also have  $d(F, G) = 0$ .

(5) Take a  $\mathbb{C}$ -analytic conjugacy between  $F$  and  $G$  on some neighborhood of the filled in Julia sets (see §9). This means for a sequence of  $n \rightarrow \infty$  we can have  $K_n$  qc conjugacies between  $f_n$  and  $g_n$  on definite neighborhoods of filled in Julia sets and  $K_n \rightarrow 1$ .

Now we view  $(f_0, g_0)$  as lying deep inside  $(f_n, g_n)$  and take a limit of the above conjugacies. The fixed neighborhood of  $f_n$  becomes a huge neighborhood of  $f_0$  and the limiting conjugacy is a  $\mathbb{C}$ -analytic conjugacy between  $f_0$  and  $g_0$  defined on all of  $\mathbb{C}$ . Q.E.D.

REMARK. McMullen offers this proof: It is known (Lyubich) that *almost every point in a Julia set of a complex quadratic-like mapping is forward asymptotic to the closure of the forward critical orbit*. If we apply this statement to  $f_1$  we see almost every point in the Julia set of  $f_1$  eventually lands on the Julia set of  $f_0$  by the renormalization disk picture (§12).

From the classification [S5] the interior of the filled in Julia set of  $f_1$  is the union of the preimages of the interior of the filled in Julia set of  $f_0$ .

**15. Proof of Theorem 2.** Theorem 2 follows from Theorem 2' and the following (see §9).

**THEOREM 6.** *If two symmetric complex quadratic-like maps  $F$  and  $G$  have the same bounded type  $(\sigma_0, \sigma_1, \dots)$ , then they have the same internal class.*

PROOF. All the renormalizations are bounded smooth quadratic-like mappings by Theorem 1. Because the combinatorics is bounded this means the geometry of the interval collections at one level inside an interval at the previous level is bounded. Otherwise, we could change the kneading sequence in a limit of bounded shape examples with converging combinatorics, and this would contradict the continuity of kneading sequences in the  $C^1$  topology.

Bounded geometry of the interval collections means we can present the critical orbit Cantor sets as an intersection of symmetric pictures in a plane of disks within disks of bounded geometry (Figure 1 on next page). But then we can construct a qc mapping between two such critical orbit Cantor sets by choosing the standard symmetric rigid maps between corresponding circles and extending them to bounded distortion diffeomorphisms between intermediate “pairs of pants” regions. Using §11 we can promote this to a symmetric qc conjugacy between  $F$  and  $G$ .

The complex polynomial  $z \rightarrow z^2 + c$ ,  $c$  real equivalent to  $F$ , admits a symmetric invariant conformal structure on its Julia set. This conclusion is valid for all the polynomials  $z \rightarrow z^2 + c$  with this kneading sequence (thinking

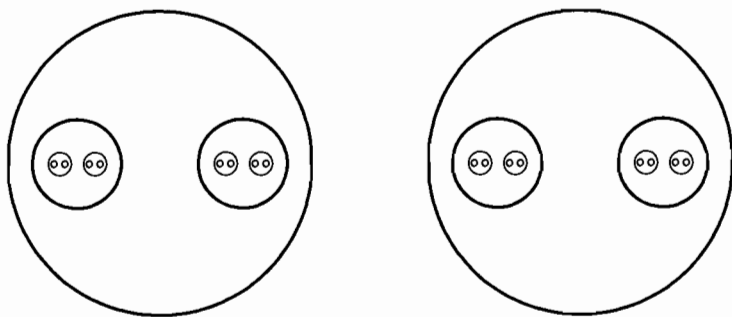


FIGURE 1

of them as  $G$  above, e.g.). There is a closed interval of such  $c$ 's if Theorem 6 is false. Apply this statement about conformal structures to an endpoint of this interval. Construct a quasiconformal deformation of this quadratic polynomial Sullivan [S5] to see it is not the endpoint. Contradiction.

**REMARK.** The theorem is unknown for unbounded type because even though we have by Theorem 1 uniform bounds on the renormalizations we do not know there is a qs conjugacy between critical orbits of two  $f$  and  $g$  of the same unbounded type. This gap, because of Yoccoz's recent work, and the above Theorem 6 is the only one left to settle the celebrated question of density of hyperbolic systems in the quadratic family.

**Appendix. Riemann surface laminations and their Teichmüller theory.** Here is what we need to provide a basis for the argumentation of §§9–15.

(1) The notions of Riemann surface laminations, Beltrami tensor, and integrable quadratic differential have to be defined.

(2) (a) For a real analytic degree two expanding map  $f$ ,  $L_f$  constructed as in §9 or as the space of orbits of  $\bar{F}$  on  $\tilde{U}$ , where  $\tilde{U} = \varprojlim \{\cdots F^{-2}U \rightarrow F^{-1}U \rightarrow U\}$ ,  $\bar{F} = \varprojlim F$ , and  $F$  is a  $\mathbb{C}$ -analytic extension of  $f$  to a neighborhood  $U$ , should be a Riemann surface lamination.

(b) The integrable quadratic differentials  $\bar{\varphi}$  on  $L_f$  lift to  $\tilde{U}$  and then project to  $\varphi$  on  $U$ , which satisfy  $F_*\varphi = \varphi$ . (This means a system  $\{\varphi_m\}$  with  $\varphi_m$  on  $F^{-m}U$  and  $F_*\varphi_m = \varphi_{m-1}$ .)

(c) If a Beltrami tensor  $\bar{\mu}$  on  $L_f$  comes from an  $F$ -invariant Beltrami coefficient on  $U$ , then the pairing  $(\bar{\varphi}, \bar{\mu})$  is computed as an integral  $\int \varphi \mu$  over a fundamental domain of  $F$  in  $U$ .

(3) Two Beltrami tensors  $\mu$  and  $\gamma$  on  $L_f$  differ by a trivial deformation  $\mu - \gamma = \bar{\partial}V$  for a continuous vector field on  $L_f$  tangent to the leaves iff the holomorphic invariants are equal,  $(\mu, \varphi) = (\gamma, \varphi)$ , for all integrable

holomorphic quadratic differentials  $\varphi$  on  $L_f$ .

(4) The notion of holomorphic quadratic differential also has to be such that the Grötzsch inequality is valid. If  $\psi_t$ ,  $0 \leq t \leq 1$ , is a quasiconformal isotopy in the leaves of the lamination between conformal structures  $c_0$  and  $c_1$ , where  $c_1$  is obtained by stretching  $c_0$  by a factor  $l$  along the trajectories of  $\varphi$ , then  $l \leq \int K(x) d|\varphi|$ , where  $K(x)$  is the conformal dilatation of  $\psi_1$  at  $x$ .

(5) The Grötzsch inequality leads directly to the almost geodesic lemma. We say  $\mu$  is  $\varepsilon$ -extremal if  $|\mu|_\infty \leq (\sup \int \varphi \mu)(1 + \varepsilon)$ ,  $|\varphi| = 1$ . (Extremal means  $\varepsilon$ -extremal for  $\varepsilon = 0$ .) This allows one to prove the *almost geodesic* property of the global deformation determined by stretching a distance  $l$  along an  $\varepsilon$ -extremal  $\mu$ . Let  $\psi_t$ ,  $0 \leq t \leq 1$ , be a qc isotopy which compares the initial conformal structure  $c_0$  and the conformal structure  $c_1$  obtained by stretching along  $\mu$ , which is  $\varepsilon$ -extremal, a distance  $l$ . Let  $K$  be the maximum dilatation of  $\psi_1$ , then there is a universal function  $\delta(\varepsilon, l)$ , where  $\delta(\varepsilon, l) \rightarrow 0$  as  $\varepsilon \rightarrow 0$  so that  $l \leq K(1 - \delta)$  (*almost geodesic lemma*).

Here we go.

(1) and (2) A closed Riemann surface lamination  $L$  will be a compact space so that each point has a neighborhood (open disk  $\times$  transversal) with overlap homeomorphisms  $F(z, \lambda)$  preserving the disk factors and holomorphic in  $z$ . Beltrami tensors from the point of view of functional analysis are bounded Borel measurable functions modulo equality a.e. in each disk. We assume in addition that as a function of  $\lambda$ ,  $\mu(z, \lambda)$  varies continuously in the topology of convergence against each element in  $L^1$  (disk). For changing coordinates,  $\mu(z, \lambda)$  is the coefficient of the tensor  $d\bar{z}/dz$ . Quadratic differentials analytically, in a product chart, are elements in a direct system of  $L^1$  spaces. The direct system is all  $\sigma$ -finite measure classes on the transversal. For each of these we form the product measure class with Lebesgue measure on the disk, form the  $L^1$  space, and take the union (or rather direct limit). For changing coordinates these objects can be viewed for each measure as  $L^1$  cross sections of the line bundle whose fiber is (volume elements of measure class on the transversal)  $\otimes (dz^2$  on disk).

We lose no generality by restricting attention to transversally invariant measure classes because these are cofinal in the directed set of all transversal measure classes.

We have a pairing  $(\varphi, \mu)$  between Beltrami coefficients and quadratic differentials by integration over  $L$ . With these definitions the requirements of (2) are satisfied.

(3) (a) Using the formula on one disk

$$V = \int_D \mu \frac{d\zeta d\bar{\zeta}}{\zeta - z} + \int_{\partial D} V \frac{d\zeta}{z - \zeta},$$

where  $\mu = \bar{\partial} V$  shows  $V$  is bounded near  $O \in D$  with a modulus of con-

tinuity of the form  $\delta \log \delta$  near there with constants depending on  $|\mu|_\infty$  whenever  $V$  vanishes on  $\partial D$ .

(b) Now a continuous vector field  $V$  on a compact Riemann surface lamination  $L$  is bounded by compactness. We now make a hyperbolic assumption on  $L$ : there is a continuous leafwise Riemannian metric on  $L$  which is conformal and which on the universal cover of each leaf is uniformly quasi-isometric to the hyperbolic metric on the cover which is assumed to be the disk. Then  $V$  vanishes at the boundary of the cover and by (a) we have bounds and a leafwise modulus of continuity (in the hyperbolic metric and therefore the Riemannian metric) controlled by  $|\bar{\partial}V|_\infty$ .

(c) Suppose  $\bar{\partial}V_i = \mu_i$  and  $\mu_i \rightarrow \mu$  weakly in  $L^\infty$  on each leaf. By (a) and (b) we can in each leaf take pointwise limits of the formula (a) with no boundary term. The limit  $V$  satisfies  $\bar{\partial}V = \mu$  and the formula with no boundary term for every point in each leaf. The argument works for nets as well as sequences.

*Claim.*  $V$  is a continuous vector field on  $L$ .

PROOF OF CLAIM. We combine four points

(i) The formula (a) can be restricted to a large metric disk in a leaf to compute  $V$  approximately at the center.

(ii) If  $x_i \rightarrow x$  in  $L$ , large metric disks about  $x_i$  converge isometrically to a large metric disk about  $x$  (or at least a covering of such a metric disk). This is the basic property of laminations.

(iii) The kernel  $dz/z$  of formula (a) is almost fixed in  $L^1$  by a mapping fixing  $O$  of a large hyperbolic disk with small isometric distortion.

(iv) Since  $\mu$  is weakly continuous in the  $\lambda$  variable (by assumption) and  $L^\infty$  bounded, its integral against a  $\lambda$ -continuous family of  $L^1$  functions is continuous in  $\lambda$ .

(d) By (c) the bounded  $\mu$ 's of the form  $\bar{\partial}V$  for continuous  $V$ 's on  $L$  is closed for the weak topology defined by integration against integrable quadratic differentials. Thus, these  $\mu$ 's are precisely those which annihilate all the  $\varphi$ 's which annihilate all the  $\bar{\partial}V$ . But the integration by parts formula  $\int_L \bar{\partial}V\varphi = \int V\bar{\partial}\varphi$  shows  $\int V\bar{\partial}\varphi = 0$  for all  $V$ , or  $\varphi$  is holomorphic on leaves. This proves (3).

(4) A holomorphic quadratic differential can be viewed locally as a positive measure on the transversal times an  $L^1$  function in the holomorphic quadratic differentials on the disks. With this convention in mind we associate to a holomorphic integrable quadratic differential  $\varphi$  (i) a metric on leaves on curvature  $\leq 0$  (the coordinates where  $\varphi$  is locally  $dz^2$  or  $z^k dz^2$ ), (ii) the measured lamination on each leaf whose trajectories are tangent to the line elements so that  $\varphi$  (line element)  $> 0$  and whose transverse measure is determined by the metric of (i), and (iii) a further transverse measure to these trajectories in the transverse direction defined by the (transverse) measure defining  $\varphi$ .

These trajectories so transversally measured defined a generalized closed geodesic curve  $|\varphi|$  which is of course tangent to the leaves. If we deform this curve by an isotopy  $\psi_t$ ,  $0 \leq t \leq 1$ , in the leaves the length can only increase because the curvature is  $\leq 0$  and we start from a geodesic. Let  $c_o$  be the original conformal structure and let  $c_l$  be the conformal structure associated to the metric obtained by stretching the  $\varphi$ -metric by a factor  $l$  along the trajectories of  $\varphi$ .

Let  $K(x)$  denote the conformal dilatation of  $\psi_1$  between  $c_o$  and  $c_l$ . Let  $J(x)$  be the Jacobian of  $\psi_1$  between these two metrics. Then we compute the length of  $\psi_1$  (curve) in the stretched metric

$$\left| \int D(x) d\varphi \right| \leq \int K^{1/2}(x) J^{1/2}(x) d\varphi \leq \left( \int K d\varphi \right)^{1/2} \left( \int J d\varphi \right)^{1/2},$$

where  $D(x) \leq K(x)^{1/2} J(x)^{1/2}$  is a derivative of  $\psi_1$ .

Assume  $|\varphi| = 1$ ; then  $\int J d\varphi = l$  and recall the length of the image can only be longer than the homotopic closed geodesic whose length is  $l|\varphi| = l$ . We deduce the Grötzsch inequality

$$(d) \quad l \leq \int |K(x)| d|\varphi|.$$

REMARK. One can write an exact formula

$$D(x) = K^{1/2}(x) J^{1/2}(x) (\text{angle factor})^{1/2}$$

and get the better inequality (Reich Strebel)

$$l \leq \int |K(x) (\text{angle factor})| d|\varphi|,$$

which implies if  $\sup K(x) = l(1 + \varepsilon)$  for  $\varepsilon$  small, then the angle factor must be near 1 on a fraction of points near 1 relative to the measure  $|\varphi|$ .

(5) Now we prove the almost geodesic lemma mentioned above. Suppose  $\mu$  is  $\varepsilon$ -extremal and choose an integrable holomorphic quadratic differential  $\varphi$  of norm 1 so that  $|\int \varphi \mu| (1 + \varepsilon) \geq |\mu|_\infty$ .

It follows by elementary arithmetic that  $\mu$  must line up in measure (relative to  $|\varphi|$ ) with the trajectories of  $\varphi$ , and (relative to  $|\varphi|$ ) have essentially constant  $L^\infty$  norm. So stretch the original conformal structure in the  $\mu$  direction by this essentially constant factor  $l$  to obtain  $c_l$ .

Now stretch the conformal structure  $c_o$  by a factor  $l$  along  $\varphi$ , where  $l = |\mu|_\infty$ , to obtain  $\bar{c}_l$ . Consider the map  $\psi$  between  $\bar{c}_l$  and  $c_o$  which is  $\bar{c}_l \xrightarrow{\text{identity}} c_l \xrightarrow{\psi_l^{-1}} c_o$ . If  $K$  is dilatation of  $\psi_1$  between  $c_l$  and  $c_o$ , then the dilatation of the composition above is at most  $2l + K$  at almost all points and  $\leq K + o(1)$  at most points because  $\mu$  and  $\varphi$  line up in measure relative to  $|\varphi|$ .

Now  $c_o$  is obtained from  $\bar{c}_l$  by stretching along the orthogonal trajectories by a factor  $l$ . By Grötzsch,  $l \leq \int |\text{dilatation } \psi| d|\varphi|$ . This is a contradiction if  $K = l'$  is less than  $l$  by an amount which is independent of  $\varepsilon$  for  $l$  and  $l'$  fixed.



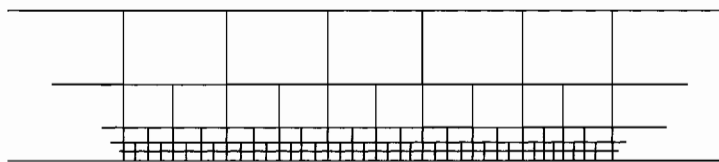


FIGURE 2

(6) The Teichmüller metric is defined infinitesimally by  $\sup \int_L \mu \varphi$ ,  $\varphi$  holomorphic and  $|\varphi| = 1$ , for all Beltrami coefficients  $\mu$  which define tangent vectors to the space of conformal structures on  $L$ . We integrate to get the length of paths of conformal structures and define the resulting metric on leafwise isotopy classes of conformal structures on a fixed background quasiconformal model  $L$ . This defines the *Teichmüller metric space* of  $L$ .

In the  $L_f$  case,  $f$  real analytic, the Teichmüller metric is related to the qc conjugacy metric by the following:

REMARK. All the  $f^{-n}$  have small ratio distortion at small enough scales, with constants depending on the Hölder constants. Similarly, all the  $F^{-n}$  for some  $\mathbb{C}$ -analytic extension  $F$  of  $f$  to a neighborhood  $U$  of  $S = \{|z| = 1\}$  have small nonlinearity on a small enough neighborhood of  $S$ , with constants depending on  $U$ .

The  $T$  metric  $d$  between  $f$  and  $g$  estimates linearly log ratio discrepancies between the Markov grids for  $f$  and  $g$  at arbitrarily fine scales. (Proof: these ratios are organized by a Hölder continuous scaling function on the Cantor set of ends of the tree of inverse branches [S3] and  $d$  estimates the speed of change of the scaling function; (see example below). Thus, if  $d(f, g) \leq \varepsilon$  and  $F$  and  $G$  are defined on definite neighborhoods  $U$ , then all corresponding consecutive ratios between the grids for  $f$  and  $g$  are estimated by  $O(\varepsilon)$  below a scale depending on  $U$  and  $\varepsilon$ . Also, we can construct for  $F$  and for  $G$  an invariant system of vertices for a system of Carleson boxes starting at a definite scale (see Figure 2).

We divide each box into three triangles and construct an almost simplicial conjugacy between fundamental domains of  $F$  and  $G$ . We pull this back to obtain a qc conjugacy between  $F$  and  $G$ .

CONCLUSION. If the Teichmüller  $d(f, g) = O(\varepsilon)$ ,  $f$  and  $g$  have  $\mathbb{C}$ -analytic extensions to a definite neighborhood  $U$  of  $\{|z| = 1\}$ , then on smaller definite neighborhoods (depending on  $\varepsilon$  and  $U$ ) there is a qc conjugacy between  $F$  and  $G$  with conformal distortion,  $O(\varepsilon)$ .

EXAMPLE (quadratic differentials on  $L_f$ ). An integrable holomorphic quadratic differential  $\tilde{\varphi}$  on an upper half plane leaf of  $\bar{U}$  can be pushed down to  $L_f$ . For example if we take  $\tilde{\varphi}$  to have poles at infinity plus three points of the Markov grid of a solenoidal leaf (the grid is pulled up from the circle) we obtain a  $\varphi$  which measures the change in one of the asymptotic

ratios which are recorded by the scaling function [S3].

REMARK. (1) The Teichmüller discussion above goes through for locally compact laminations. We only need to require that the quasiconformal vector fields are continuous and uniformly bounded. Then the proofs of (3), (4), (5), and (6) are unchanged. In this way we have a generalization of classical Teichmüller space by considering the case of a lamination with one leaf.

(2) There are enough integrable holomorphic differentials to have a version of (3) in the *measurable theory* where  $\mu(x, \lambda)$  and  $V(x, \lambda)$  are defined, uniformly bounded, and measurable in  $\lambda$  for each transversal measure class in a consistent manner. Then the Beltrami tensors form the complete dual of measurable integrable quadratic differentials.

The proof of (3) is the same and in addition there is the formula for the Teichmüller norm (measurable theory)  $\inf_V |\mu + \bar{\partial} V| = \sup_\varphi \int \mu \varphi \equiv |\mu|_T$ ,  $\varphi$  holomorphic of mass one. We used this formula in the *continuous theory* for  $L_f$  for those tangent vectors  $\mu$  coming from fundamental domains of  $F$ .

## REFERENCES

- [A] L. V. Ahlfors, *Conformal invariants*, McGraw-Hill, New York, 1973.
- [AAC1] R. Artuso, E. Aurell, and P. Cvitanović, *Recycling of strange sets: I. Cycle expansions*, *Nonlinearity* **3** (1990), 325–359.
- [AAC2] —, *Recycling of strange sets: II. Applications*, *Nonlinearity* **3** (1990), 361–386.
- [BL1] A. M. Blokh and M. Yu. Lyubich, *Measurable dynamics of S-unimodal maps of the interval*, preprint, Stony Brook, 1990/2.
- [BL2] —, *Measure and dimension of solenoidal attractors of one-dimensional dynamical systems*, *Comm. Math. Phys.* **127** (1990), 573–583.
- [C] M. Cosnard, *Etude des solutions de l'équation fonctionnelle de Feigenbaum*, *Bifurcations, Théorie Ergodique et Applications*, Astérisque **98–99** (1982), 143–162.
- [Cv] P. Cvitanović, *Recycling chaos*, *Nonlinear Physical Phenomena* (Brasilia 1989 Winter School) (A. Ferraz, F. Oliveira, and R. Osorio, eds.), World Scientific, Singapore, 1990.
- [CCR] F. Christiansen, P. Cvitanović, and H. Rugh, *The spectrum of the period doubling operator in terms of cycles*, *J. Phys. A* **23** (1990), L713–L717.
- [CE] P. Collet and J.-P. Eckmann, *Iterated maps of the interval as dynamical systems*, Birkhäuser, Boston, 1980.
- [CEK1] P. Collet, J.-P. Eckmann, and H. Koch, *Period-doubling bifurcations for families of maps on  $R^n$* , *J. Statist. Phys.* **25** (1980), 1–14.
- [CEK2] —, *On universality for area preserving maps of the plane*, *Phys. D* **3** (1981), 457–467.
- [CEL] P. Collet, J.-P. Eckmann, and O. E. Lanford, III, *Universal properties of maps on the interval*, *Comm. Math. Phys.* **76** (1980), 211–254.
- [CER] M. Campanino, H. Epstein, and D. Reulle, *On Feigenbaum's functional equation*, *Topology* **21** (1982), 125–129; *On the existence of Feigenbaum's fixed point*, *Comm. Math. Phys.* **79** (1981), 261–302.
- [CT1] P. Coullet and C. Tresser, *Itération d'endomorphismes et groupe de renormalisation*, *J. Phys. Colloq. C 539* (1978), C5–25; *C. R. Acad. Sci. Paris Sér. A* **287** (1978).
- [CT2] —, *Preturbulent states and renormalization group for simple models*, *Rep. Math. Phys.* **17** (1980), no. 2, 189–203.
- [D] W. F. Donoghue, Jr., *Monotone matrix functions and analytic continuation*, Springer-Verlag, Berlin, 1974.
- [deF] Edson de Faria, Thesis, City University of New York, 1991.
- [DGP] B. Derrida, A. Gervois, and Y. Pomeau, *Universal metric properties of bifurcations of endomorphisms*, *J. Phys. A* **12** (1979), 269.

- [DH1] A. Douady and J. H. Hubbard, *On the dynamics of polynomial-like mappings*, Ann. Sci. École Norm. Sup. (4) **18** (1985), 287–343.
- [DH2] —, *Étude dynamique des polynômes complexes*. I and II, Publ. Math. Orsay, Univ. Paris XI, Orsay.
- [E1] H. Epstein, *New proofs of the existence of the Feigenbaum functions*, Comm. Math. Phys. **106** (1986), 395–426.
- [E2] —, *Fixed points of composition operators*, Non-linear Evolution and Chaotic Phenomena (P. Zweifel, G. Gallavotti, and M. Anile, eds.), Plenum Press, New York, 1988.
- [E3] —, *Fixed points of composition operators*. II, Nonlinearity **2** (1989), 305–310.
- [EE1] J.-P. Eckmann and H. Epstein, *On the existence of fixed points of the composition operator for circle maps*, Comm. Math. Phys. **107** (1986), 213–231.
- [EE2] —, *Fixed points of composition operators*, VIIIth International Congress on Mathematical Physics (Marseille, 1986) (M. Mebkhout and R. Seneor, eds.), World Scientific, Singapore, 1987.
- [EE3] —, *Bounds on the unstable eigenvalue for period doubling*, Comm. Math. Phys. **128** (1990), 427–435.
- [EEW] J.-P. Eckmann, H. Epstein, and P. Wittwer, *Fixed points of Feigenbaum's type for the equation  $f^p(\lambda x) \equiv \lambda f(x)$* , Comm. Math. Phys. **93** (1984), 495–516.
- [EL] H. Epstein and J. Lascoux, *Analyticity properties of the Feigenbaum function*, Comm. Math. Phys. **81** (1981), 437–453.
- [EW1] J.-P. Eckmann and P. Wittwer, *Computer methods and Borel summability applied to Feigenbaum's equation*, Lecture Notes in Phys., vol. 227, Springer-Verlag, Berlin, 1985.
- [EW2] —, *A complete proof of the Feigenbaum conjectures*, J. Statist. Phys. **46** (1987), 455–477.
- [F1] M. J. Feigenbaum, *Quantitative universality for a class of non-linear transformations*, J. Statist. Phys. **19** (1978), 25–52.
- [F2] —, *Universal metric properties of non-linear transformations*, J. Statist. Phys. **21** (1979), 669–706.
- [Fa] C. Falcolini, *Some solutions of Feigenbaum's functional equation*, Boll. Un. Mat. Ital. A(7) **1** (1987).
- [FKS] M. J. Feigenbaum, L. P. Kadanoff, and S. J. Shenker, *Quasi-periodicity in dissipative systems: A renormalization group analysis*, Phys. D **5** (1982), 370–386.
- [G] Fred Gardiner, *Teichmüller theory and quadratic differentials*, Pure Appl. Math., Wiley, 1987.
- [Gr] J. Groeneveld, *On constructing complete solution classes of the Cvitanović-Feigenbaum equation*, Phys. A **138** (1986), 137–166.
- [G1] J. Guckenheimer, *Renormalization of one dimensional mappings and strange attractors*, Contemp. Math., vol. 58, part III, 1987, pp. 143–160.
- [G2] —, *Limit sets of S-unimodal maps with zero entropy*, Comm. Math. Phys. **110** (1987), 655–659.
- [GJ] J. Guckenheimer and S. Johnson, *Distortion of S-unimodal maps*, Ann. of Math.
- [H] M. Herman, *Quasicircles and Siegel disks*.
- [JR] L. Jonker and D. Rand, *Universal properties of maps of the circle with  $\varepsilon$ -singularities*, Comm. Math. Phys. **90** (1983), 273–292.
- [JS1] M. Jakobson and G. Świątek, *Induced hyperbolicity for one dimensional maps* (to appear).
- [JS2] —, *Metric properties of non-renormalizable S-unimodal maps*, I. Induced expansion and invariant measures (to appear).
- [Ly1] M. Yu. Lyubich, *On typical behavior of the trajectories of a rational mapping of the sphere*, Soviet Math. Dokl. **27** (1983), 22–24.
- [Ly2] —, *Non-existence of wandering intervals and structure of topological attractors of dimensional dynamical systems*, I. The case of negative Schwarzian derivative, Ergodic Theory Dynamical Systems **9** (1989), 737–750.
- [Ly3] —, *Ergodic theory for smooth one dimensional dynamical systems* (to appear).
- [L1] O. E. Lanford III, *Remarks on the accumulation of period-doubling bifurcations*, Mathematical Problems in Theoretical Physics, Lecture Notes in Phys., vol. 116, Springer-Verlag, Berlin, 1980, pp. 340–342; *A computer-assisted proof of the Feigenbaum conjectures*, Bull. Amer. Math. Soc. (N.S.) **6** (1984), 127.
- [L2] O. E. Lanford, *Critical circle mappings* (to appear).

- [L3] —, *Finite analysis in the Feigenbaum problem* (to appear).
- [L4] O. E. Lanford, III, *Renormalization group methods for circle mappings*, Proc. Conf. on Statistical Mechanics and Field Theory: Mathematical Aspects (Groningen, 1985), Lecture Notes in Phys., Springer-Verlag (to appear).
- [L5] —, *A shorter proof of the existence of the Feigenbaum fixed point*, Comm. Math. Phys. **96** (1984), 521–538.
- [L6] —, *Functional equations for circle homeomorphisms with golden ratio rotation number*, J. Statist. Phys. **34** (1984), 57–73.
- [LL] O. E. Lanford, III and R. de la Llave, unpublished.
- [M] Marco Martens, *Intervals dynamics*, Thesis, Delft, 1990.
- [Mc] Curt McMullen, *Rational maps and Kleinian groups*, Proc. 1990 Internat. Congr. Math. (Kyoto, Japan).
- [Me] B. Mestel, Ph.D. Dissertation, Department of Mathematics, Warwick University, 1985.
- [Mi] C. Mira, *Complex dynamics in two dimensional endomorphisms*, Nonlinear Anal. **4**, 1167.
- [MJ] W. de Melo, *Lectures on one-dimensional dynamics*, IMPA, Rio de Janeiro, 1989.
- [MMV] M. Martens, W. de Melo, and S. Van Strien, *Fatou Julia Sullivan theory*, Acta Math. (to appear).
- [MS] C. McMullen and D. Sullivan, *Quasiconformal homeomorphism and dynamics*, III. *The Teichmüller space of a holomorphic dynamical system* (to appear).
- [MSS] R. Mane, P. Sad, and D. Sullivan, *On the dynamics of rational maps*, Ann. Sci. École Norm. Sup. (4) **16** (1983), 193–217.
- [MT] J. Milnor and W. Thurston, *Mappings of the interval*, Lecture Notes in Math., Springer-Verlag, 1988.
- [MV] W. de Melo and S. van Strien, *One dimensional dynamics*, Ergebrisse Series, Springer-Verlag, 1991.
- [MV1] W. de Melo and S. Van Strien, *A structure theorem in one-dimensional dynamics*, Ann. of Math. (2) **129** (1989), 519–546.
- [MV2] —, *Schwarzian derivative and beyond*, Bull. Amer. Math. Soc. (N.S.) **18** (1988), 159–162.
- [N] M. Nauenberg, *On fixed points for circle maps*, Phys. Lett. AB **92** (1982), 319–320.
- [ORSS] S. Ostlund, D. Rand, J. Sethna, and E. Siggia, *Universal properties of the transition from quasiperiodicity to chaos in dissipative systems*, Phys. D **8** (1983), 303–342.
- [Pa] W. Paluba, Thesis, City University of New York, 1991.
- [Po] M. Pollicott, *A note on the Artuso-Aurell-Cvitanović approach to the Feigenbaum tangent operator*, J. Statist. Phys. (to appear).
- [R1] D. Rand, *Universality and renormalization in dynamical systems*, New Directions in Dynamical Systems (T. Bedford and J. W. Swift, eds.), Cambridge Univ. Press, 1987.
- [R2] —, *Global phase space universality, smooth conjugacies and renormalisation. The  $C^{1+\alpha}$  case*, Nonlinearity.
- [Ru] D. Ruelle, *Thermodynamic formalism*, Addison-Wesley, Reading, MA, 1978.
- [S1] D. Sullivan, *Conformal dynamical systems*, Internat. Conf. on Dynamical Systems (Rio de Janeiro, 1976), Princeton Univ. Press, Princeton, NJ, 1980.
- [S2] —, *Quasiconformal homeomorphisms in dynamics, topology, and geometry*, Proc. Internat. Congr. Math. (Berkeley, 1986).
- [S3] —, *Differentiable structures on fractal like sets*, H. Weyl Symposium (Duke University, 1988), Amer. Math. Soc., 1990.
- [S4] —, *Bounded structure of infinitely renormalizable mappings*, Universality in Chaos, 2nd ed. (P. Cvitanović, ed.), Adam Hilger, Bristol, 1989.
- [S5] —, *Quasiconformal homeomorphisms and dynamics*, I. *Fatou Julia problem on wandering domains*, Ann. of Math. (2) **122** (1985), 401–418.
- [Sh] A. N. Sharkovskii, *Coexistence of cycles of a continuous map of a line into itself*, Ukrain. Mat. Z.
- [SS] M. Shub and D. Sullivan, *Expanding mappings of the circle revisited*, Ergodic Theory Dynamical Systems, 1985.
- [ST] D. Sullivan and W. Thurston, *Extending holomorphic motions*, Acta Math. **157** (1986), 243–257.

- [Sv1] Sebastian van Strien, *Hyperbolicity and invariant measures for general  $C^2$  interval maps satisfying the Misiurewicz condition*, Comm. Math. Phys. **128** (1990), 437–496.
- [Sv2] ———, *On the bifurcations creating horseshoes*, London Math. Soc., 1991, pp. 316–351.
- [Sw] G. Świątek, *Critical circle maps*, Comm. Math. Phys. **119** (1989), 109–128.
- [V] G. Valiron, *Fonctions analytiques*, Presses Univ. Paris, France, 1954.
- [VSK] E. B. Vul, Ya. G. Sinai, and K. M. Khanin, *Feigenbaum universality and the thermodynamical formalism*, Uspekhi Mat. Nauk **39** (1984), 3–37.
- [Y1] J. C. Yoccoz, private communication.
- [Y2] ———, *On local connectivity of the Mandelbrot set* (to appear).

GRADUATE SCHOOL AND UNIVERSITY CENTER, CITY UNIVERSITY OF NEW YORK, NEW YORK,  
NEW YORK 10036

INSTITUTE DES HAUTES ETUDES SCIENTIFIQUE, BURES SUR YVETTE, FRANCE 91440

## Instantons and Their Relatives

KAREN UHLENBECK

**Introduction and a glimpse of history.** It was a great honor to be invited to give an hour talk at the Centennial Celebration of the American Mathematical Society. I have been reminded many times both before and after that I was the one woman speaker among the twenty odd speakers of this conference. Unlike some of my younger colleagues who gave addresses, I myself cannot expect to be present at the 150th anniversary celebration, although I very much hope a world sufficiently similar to ours exists in 2038 for such a celebration to take place. I also hope that no comments on the place of women in mathematics are even relevant at the time of this next celebration.

In preparing and writing up my talk, which was introduced by the legendary differential geometer S. S. Chern, I have been very aware that I am the only speaker representing the exciting developments which have taken place in global differential geometry in the last fifteen years. The technical understanding of elliptic partial differential equations has led to unprecedented understanding of the global aspects of diverse basic ideas in geometry such as minimal surfaces and Riemannian curvature equations. Applications in topology, algebraic geometry, and applied mathematics are very striking and important. My talk, however, concentrated on a completely new subject: the study of curvature or field equations linked not to the geometry of the manifold but to the extrinsic geometry of objects with the technically obscure name of principal bundles. Their structure groups appear in particle physics as the  $SU(2)$  is isospin; the  $SU(3)$ 's of isospin and strangeness, color, and charm; the  $SU(5)$  of unified field theories; and the  $E_8$ 's of string theory. The classical equations of the gauge theory of theoretical physicists entered pure mathematics. In a very few years they have become central to mathematicians' understanding of objects such as smooth four-manifolds and stable bundles. This development has been the event of greatest intellectual excitement in my career. I am left with the feeling that the few small contributions I made in gauge field theory were right in the center of intellectual progress.

Hence my continued pursuit of a chimera: an understanding of mathematical ideas as they come from outside mathematics itself. My energies have taken me only to the mathematical edge of physics.

Before I begin the official write-up of my talk, I would like to take this opportunity to thank all the people and institutions who have encouraged and supported my mathematical career thus far. Mention of a special sort should go to my thesis advisor, Richard Palais, and to my good friend, S. T. Yau. Finally, my junior colleague, Dan Freed, has been of great help in preparing this manuscript.

**1. The birth of gauge theory.** How did gauge theory appear and become successful in mathematics in the space of a few years? The fundamental mathematical ingredients were in place. The basics of fibre and vector bundles and their connections were in daily use by geometers. Chern-Weil theory (and even Chern-Simons invariants) were studied in most graduate courses in differential geometry. De Rham cohomology and its realization via the Hodge theory of harmonic forms were standard items in differential topology. In hindsight, the Yang-Mills equations were waiting to be discovered. Yet mathematicians were in themselves unable to create them. Gauge field theory is an adopted child.

Physicists Yang and Mills wrote down their equations in 1954, referring explicitly to isotopic spin as the group invariance. Some ingredients of gauge theory became incorporated gradually into the theory of the electro-weak interactions in physics over the next twenty years. These are not particularly recognizable or striking to mathematicians as gauge theories because of their "broken symmetry." There were isolated cases of mathematicians noticing the importance of these equations, but no essential impression was made on the mathematical community as a whole.

The original Yang-Mills equations are nonlinear extensions of Maxwell's equations in space-time. This means they are a system of second-order partial differential equations in the four variables of space and time ( $3 + 1$  dimensions). In the process of studying the quantum theory, the solutions to the Euclidean four-dimensional second-order equations become important. A series of important papers in the seventies starts with the discovery of the first-order self-dual equations and the single instanton solution (which one can think of as rotationally symmetric in  $\mathbb{R}^4$ ) by the Russian physicists Belavin, Polyakov, Schwarz, and Tyupkin in 1975. Conformal invariance produces a five-dimensional family, which has turned out to be a complete family of solutions of energy  $8\pi^2$ . A larger  $5|k|$  parameter family of energy  $|k|8\pi^2$  for all  $k$  was almost immediately discovered by a number of physicists (Wilcek; Corrigan and Fairlie; Jackiw, Nohl, and Rebbi). The form of the solutions employed the "'t Hooft Ansatz" of reducing the equations to the linear equation  $\Delta\phi = 0$ . Almost immediately, mathematicians were able to contribute information on solutions using an amazing variety of techniques

from modern mathematics. Many of the original ideas were due to Michael Atiyah, and in those early years he lectured all over the world on instantons and gauge theories. Much of the purely mathematical development is surely due to the interest and excitement he conveyed to his mathematical audiences at that time.

I have left the rest of the early mathematical development to be inferred from the table of papers in the subject. Simon Donaldson received the Field's Medal in 1986 primarily for the work in his Ph.D. thesis, published in 1983 in the *Journal of Differential Geometry*. Gauge theory has proved itself an important tool of mathematics, which I for one believe will last. The following list of Early Papers in Gauge Theory and the bibliography are but a small part of the evidence. Mathematical gauge theory provides the best understood invariants for topological 4-manifolds [D1, A1], more invariants for homology 3-spheres [A2], a description of the moduli space of stable holomorphic bundles over Kähler manifolds [D2], a tool for uniformization theorems [S], concrete descriptions of cosmological objects, as well as a special model toy for sophisticated mathematicians investigating abstract mathematical phenomena.

It is an important part of physics? Has the child adopted by mathematicians been rejected by its natural parents? The physics papers in 1975–77 listed in the following table are in general part of a scheme to describe quantum chromodynamics. They represent failed attempts to understand strong interactions by “tunnelling effects.” The physics behind these equations has in general been completely mysterious to mathematicians, who are continually frustrated in their attempts to either understand or believe even the simplest calculations in quantum field theories of this geometric sophistication. The failure of this model does not mean they are not present in physics. In their broken form they are used in calculation for the standard model of particles. Lattice gauge theory is a thriving user of CRAY time. String theory interactions pre-suppose quantized gauge theories. Those who study the physics referred to in the talks of Vaughn Jones, Victor Kac, and Ed Witten will find them very much behind the group representation theory which has become so important. So far we mathematicians have been able to make use only of the classical theory which was not of much use in theoretical physics. Good geometric mathematical models for quantizing gauge field theories promise to be interesting to both sets of parents.

### EARLY PAPERS IN GAUGE THEORY

(1954) C. N. Yang and R. Mills, *Conservation of isotopic spin and isotopic gauge invariance*, Phys. Rev. **96**, 1–9.

(1975) A. A. Belavin, A. M. Polyakov, A. S. Schwarz, and Yu. S. Tyupkin, *Pseudo-particle solutions of the Yang-Mills equations*, Phys. Lett. B **59**, 85–87.

(1976) G. 't Hooft, *Computation of the quantum effects due to a four-dimensional pseudo-particle*, Phys. Rev. D **14**, 3432–3450.

(1976) —, *Symmetry breaking through Bell-Jackiw anomalies*, Phys. Rev. Lett. **37**, 8–11.



- (1977) E. Corrigan and D. B. Fairlie, *Scalar field theory and exact solutions to a classical  $SU(2)$  gauge theory*, Phys. Lett. B **67**, 67–71.
- (1977) Richard Ward, *On self-dual gauge fields*, Phys. Lett. A **61**, 81–82.
- (1977) M. Atiyah and R. Ward, *Instantons and algebraic geometry*, Comm. Math. Phys. **55**, 97–118.
- (1977) M. Atiyah, N. Hitchin, and I. Singer, *Deformation of instantons*, Proc. Nat. Acad. Sci. U.S.A. **74**, 2662–2663.
- (1977) M. Atiyah, N. Hitchin, V. Drinfeld, and Y. Manin, *Construction of instantons*, Phys. Lett. **65A**, 185–187.
- (1978) M. Atiyah, N. Hitchin, and I. Singer, *Self-duality in four-dimensional Riemannian geometry*, Proc. Roy. Soc. London Ser. A **362**, 425–461.
- (1978) M. Atiyah and J. D. Jones, *Topological aspects of Yang-Mills theory*, Comm. Math. Phys. **61**, 97–118.
- (1979) J. P. Bourguignon, H. B. Lawson, and J. Simons, *Stability and the gap phenomenon for Yang-Mills*, Proc. Nat. Acad. Sci. U.S.A. **76**, 1550–1553.
- (1982) C. H. Taubes, *Self-dual Yang-Mills connections on non-self-dual four-manifolds*, J. Differential Geom. **17**, 139–170.
- (1982) K. Uhlenbeck, *Connections with  $L^p$  bounds on curvature*, Comm. Math. Phys. **83**, 11–29.
- (1983) M. Atiyah and R. Bott, *The Yang-Mills equations over Riemann surfaces*, Philos. Trans. Roy. Soc. London Ser. A **308**, 523–615.
- (1983) S. Donaldson, *An application of gauge theory to the topology of four-manifolds*, J. Differential Geom. **18**, 269–316.

**2. What is an instanton?** What is all the excitement about? What is an instanton, and why is it important?

Instantons are 4-dimensional objects, and share a lot of properties with vortices (two-dimensional objects used to describe superconductivity), monopoles (three-dimensional cosmological objects), and Hermitian Yang-Mills metrics (complex objects of any complex dimension). These are all close relatives in a large diverse family of global geometric mathematical objects defined by partial differential equations. We can include in this extended family objects like geodesics, minimal and constant curvature surfaces, and black holes. All these objects extend simple well-understood physical models, satisfy nonlinear equations obtained from variational principles, exhibit certain topological properties, have a gauge or coordinate invariance, and have important modern applications in topology, algebraic geometry, and applications. There are many good descriptions of the basic equations of gauge theory in the list of basic reference books in §4. Here we give a more impressionistic view by comparing the characteristic properties and behavior of instantons with similar phenomena exhibited by minimal surfaces. It is a powerful mathematical fact that intuition and techniques can be passed back and forth between the two.

*Physical origins.* The minimal area surface equation was written down by the Belgian physicist Plateau as an equation satisfied approximately by soap films. Since nearly all of us played with soap films and bubbles as young or not so young children, we think we understand minimal area surfaces conceptually. One imagines the surface in the familiar three-space we live in, and then transposes it into a curved space created by the imagination.

Instantons represent tunnelling from flat three-dimensional space to itself. No one plays with quantum effects as a child, and even the geometric connection or vector potential which replaces the concept of surface is hard to imagine. I think of the vector potential as having some of the stretchy properties of surfaces, but sitting over the spacial manifold, not in it.

*Variational formulation.* The concept of least area is familiar to me from simple calculus problems. However, the formula for area of a general surface in a three-manifold is quite complicated. The minimal area principle is often replaced by minimal energy, or the  $L^2$  norm of the derivative of the embedding,  $s: \Sigma \rightarrow X$ .

$$E(s) = \int_{\Sigma} |ds|^2 (du)^d.$$

The Yang-Mills integral is simply the  $L^2$  norm of the curvature (or field):

$$\text{YM}(A) = \int_X |F_A|^2 (du)^d.$$

For  $\dim \Sigma = 2 = d$  and  $\dim X = 4 = d$ , both integrals are conformal invariants. This allows one to compactify  $\Sigma = \mathbb{R}^2 \cup \{\infty\} = S^2$  and  $X = \mathbb{R}^4 \cup \{\infty\} = S^4$  in exactly the same way, and produces the same borderline behavior for Morse theory. We think of the lack of compactness in Yang-Mills at points much like we think of the “bubbling” of minimal surfaces. Both are caused by scale or conformal invariance [SU, Se].

*Linear models.* Both the minimal surface (or harmonic map equations) and the Yang-Mills equations can be thought of as nonlinear generalizations of Hodge-de Rham theory. Harmonic  $p$ -forms  $\alpha \in \Omega^p(M, \mathbb{R})$  satisfy the closed condition  $d\alpha = 0$  and the Hodge equation  $d * \alpha = 0$ .

If  $s: M \rightarrow N$  is a map, then  $ds = \alpha$  is a one-form with values in  $s^*TN$ . In this context  $d_s \alpha = 0$  is an identity and  $d_s * \alpha = 0$  is the harmonic equation. Likewise, if  $F_A$  is the curvature of a connection  $A$ ,  $D_A F_A = 0$  is the Bianchi identity and  $D_A * F_A = 0$  is the Yang-Mills equation.

*First-order equations.* Both the harmonic map equation and Yang-Mills are  $d$ -dimensional equations, where  $d$  is arbitrary. However, in the scale invariant case, we have special equations. For  $\alpha = ds$ , a one-form, and  $X$  a complex Kähler manifold with complex structure operator  $J$ , a special very tractable class of minimal surfaces or harmonic maps are the holomorphic or antiholomorphic ones. They satisfy the Cauchy-Riemannian equations

$$J(s)\alpha = \pm * \alpha.$$

Likewise, instantons and anti-instantons satisfy a similar first-order equation:

$$F_A = \pm * F_A.$$

*Complex equations.* For  $\Sigma$  and  $X$  arbitrary complex manifolds, important examples of harmonic maps (or minimal surfaces) are generated by holomorphic maps  $\Sigma \rightarrow M$ . For Yang-Mills, if  $X$  is complex, there is a special form of the Yang-Mills equations which requires the curvature  $F_A$  to be a two-form of type  $(1, 1)$  which is traceless with respect to the Kähler form.

*Gauge invariance.* The concept of a minimal surface does not carry with it a preferred coordinate chart. Jesse Douglas' original solution to the Plateau problem (for which he received half the first Field's medal) uses conformal coordinates obtained via the Riemann mapping theorem and replaces area by energy [Do]. No such elegant global solution for gauge fixing has emerged for the coordinate problem in gauge theory. However, locally on the manifold or locally in the space of connections, harmonic slices are used for technical constructions.

*Topological applications.* Minimal surfaces can be used to study the topology of three-manifolds [MY]. The solutions to the instanton equation have emerged as the main tool for studying the topology of differential four-manifolds [A1].

*Moduli spaces.* In studying moduli spaces of minimal surfaces, one is forced to look to the Riemann moduli space for the models. Moduli spaces of solutions to Yang-Mills exhibit many similarities with these same model spaces of complex structures on Riemann surfaces. Similar compactification phenomena exist.

*Examples.* One of the most satisfying aspects of the study of minimal surfaces in 3-space is the existence of many immediate examples via the Weierstrass representation. If  $f$  and  $g$  are any two holomorphic functions on  $\Omega$ , then the piece of surface

$$\Sigma = \{\operatorname{Re}(f(z)(1 - g^2(z))), if(z)(1 + g^2(z)), 2f(z)g(z)), z \in \Omega\} \in \mathbb{R}^3$$

is a minimal surface. This, and some hard to come by expertise with computer graphics are all that one needs to draw many beautiful pictures of minimal surfaces [H].

The Penrose transform converts solutions of  $F_A = -*F_A$  to holomorphic bundles over  $\mathbb{CP}^3$ . However, the complex analysis is considerably more complicated! Fortunately, the 't Hooft Ansatz produces solutions  $A = \operatorname{Im}(\frac{\partial}{\partial \bar{q}}(\ln \varphi)dq)$  to Yang-Mills from solutions to  $\Delta \varphi + \lambda \varphi^3 = 0$  on  $\mathbb{R}^4$ , if we regard  $\mathbb{R}^4 = \mathbb{H}$  as the quaternions. Some simple examples are available, although I do not know what computer pictures would look like. I have seen some elegant pictures of vortices and monopoles however [HMRVW].

**3. Abelian vortices.** The importance of the class of equations found in gauge theory lies almost entirely in the structure of their moduli spaces of

solutions. Of course, the moduli space of solutions to the self-dual Yang-Mills equations on a four-manifold  $X$  is very complicated. However, in some cases, special simpler solutions can be constructed using symmetries [T1]. Here we present these simpler equations and describe the moduli space of solutions on a Riemann surface. I feel this is appropriate. Clifford Taubes' description in his thesis of the solutions to the vortex equation on  $\mathbb{R}^2$  was one of the first analytical results on moduli spaces in gauge field theory [T2]. The present result is contained in a 1988 Ph.D. thesis of a Ph.D. student of mine, Steven Bradlow [B].

In this application, we let  $X$  be a Kähler manifold (it will specialize at the end to a Riemann surface). The integral for the coupled Yang-Mills-Higgs equations has the general form

$$\int_X \left[ |F_A|^2 + |D_A \Phi|^2 - \frac{\lambda}{4} (t - |\Phi|^2)^2 \right] (du)^d.$$

Here  $F_A$  is the curvature in a principal bundle and  $\Phi$  is a section of an associated bundle. The parameters  $\lambda$  and  $t$  are real. This integrand is particularly easy to describe if the group is  $U(1)$  and  $\Phi$  is a section of a line bundle  $E$ . Fix a base connection  $D_0$  on  $E$ . The unknowns are a complex-valued functions  $\Phi$  twisted to lie in  $E$  and an ordinary one-form  $A$ ,

$$F_A = F_0 + d(iA), \quad D_A \Phi = D_0 \Phi + iA \cdot \Phi.$$

The Euler-Lagrange equations are second order and have the form

$$d * F_A + \text{Im}(*D_A \Phi, \Phi) = 0, \quad \Delta_A \Phi + \frac{\lambda}{2} (-t + |\Phi|^2) \Phi = 0.$$

One family of solutions can be obtained by considering the solutions

$$\Phi = 0, \quad d * F_A = 0.$$

These are the usual Yang-Mills equations and linear Hodge-de Rham theory provides an analysis of these.

However, on a Kähler manifold we can use the complex structure to integrate the functional by parts to obtain an equivalent integral

$$\int_X \left[ |H_A|^2 + 4|F_A^{0,2}|^2 + 2|\bar{\partial}_A \Phi|^2 - H_A |\Phi|^2 + \frac{\lambda}{4} (t - |\Phi|^2)^2 \right] (du)^d - 8\pi^2 \text{ch}_2 E.$$

Here  $F_A^{0,2}$  is the  $(2, 0)$  part of the curvature,  $H_A = (\omega, F_A)$  is the contraction of the curvature two-form with the Kähler form ( $*F_A$  in two real dimensions), and  $\text{ch}_i E$  indicate topological contributions from the Chern-Weil formulas. The first-order equations come from this integration by parts exactly as they do for Yang-Mills or monopoles.

**THEOREM.** *For  $\lambda = 1$ , we have that the Yang-Mills-Higgs integral is bounded below by the number  $-8\pi^2 \text{ch}_2(E) + 2\pi \text{ch}_1 E$ . This topological minimum is taken on by solutions to the first-order equations  $\bar{\partial}_A \Phi = 0$ ,  $F_A^{0,2} = 0$ , and  $2H_A - |\Phi|^2 + t = 0$ .*

The surprise is that it is completely straightforward (given a little standard complex differential geometry) to find all the solutions of this equation.

**THEOREM (Bradlow).** *For fixed  $t > t_0 = 4\pi \text{ch}_1 E (\text{vol } X)^{-1}$ , the solutions of the first-order equations correspond in a one-to-one fashion to holomorphic sections of holomorphic line bundles.*

This gives a really simple description of the moduli space of solutions over a Riemann surface  $\Sigma$ . For  $\text{ch}_1 E = k$ , specifying the  $k$  zeros of a holomorphic section on  $\Sigma$  gives both the section and the line bundle.

**COROLLARY.** *The moduli space of solutions to the first-order vortex equations on a Riemann surface in a bundle with  $\text{ch}_1 E = k$  corresponds to the space  $\mathcal{S}^k(\Sigma)$  of  $k$  unordered points on  $\Sigma$ , possibly with multiplicity. These points correspond to zeros of the Higgs field in the solution.*

The analysis in this sample example reduces to the solution of an elliptic equation of the form

$$-\Delta u + |\Phi|^2 e^u - (t - t_0) = 0$$

for a change of metric  $e^u$  in the bundle  $E$ . This equation was studied by Kazdan and Warner in conjunction with their investigation of conformal deformations of metrics in two dimensions. Of course, the interesting and new nonlinear analysis is in the extensions to the nonabelian case, where contact is made with notions of stability in algebraic geometry. This work is also in Bradlow's thesis [B]. However, this simple abelian example serves to illustrate what we can expect moduli spaces to look like.

**4. And if there is a twenty-first century . . .** I am a pessimist. If I think about the future, I think mainly about overpopulation, AIDS, fiscal instability, the threat of nuclear war, and myriads of different seemingly unsolvable social and environmental problems. In making my predictions, I must confess that I worry there will be no twenty-first century suitable for the pursuit of mathematics.

However, I am a mathematical optimist. It seems to me that Mathematics is intellectually in great shape. Current developments are exciting. The problems of the world are not reflected as problems within the world of mathematics. This is certainly one of its attractions for me. But I see it as more than a refuge from real life. To me real progress has been made in mathematics in the twenty years I have been a member of the community, whereas I am not so sure about progress in the real world. During these years the world of mathematics has opened up to make contact with neighboring intellectual disciplines. We have been influenced by our old friends from theoretical physics—but even more by greater changes such as the advent of the computer age and by the daily use of mathematics in technology. The response within the discipline of mathematics has been very positive, if a bit

slow and conservative. As a result, the content of mathematics in the form of its fundamental ideas seems to me to be much richer. I look forward to the next decades in the development of mathematics with great curiosity and hope.

PREDICTION 1. *Simplicity through complexity.*

There are two basic approaches to simplicity in mathematics. One approach struggles with the choice of description of the mathematical object via bases or coordinates until one obtains the right and self-evident minimalist description. On the other hand one can throw in all possible descriptions (as my mother used to say, everything and the kitchen sink) and then divide out by equivalences to obtain a simple classification of objects. This suits today's complex world. The success of gauge theory is via this second complexification route. I think we are not done with this trend of enlarging the class of mathematical objects to unreal proportions and then dividing out by even larger equivalence classes. One sees this scheme working in the successful BRS quantization methods of physicists, and I predict we will continue to follow this pattern, at least in geometry [FGZ].

PREDICTION 2. *More beyond partial differential equations.*

The last fifteen years has seen the domination of differential geometry by techniques from partial differential equations. This might appropriately be called the "Yau School" of differential geometry [Y]. If the influences from physics continue, I think the era of domination will end. We should ask ourselves, "What will the discovery of a unified field theory in theoretical physics mean for differential geometry? The goal of a unified field theory is to meld the theory of gravity (geometry) with particles (groups). Geometry via algebra? It is not a completely new idea, of course.

PREDICTION 3. *Geometric understanding of quantum field theory.*

I hope that we mathematicians will soon have done with our fundamental difficulties with quantum theory. We have had fifty years of lack of success in explaining why Feynman diagrams work. I agree with many other speakers that we will soon decode the complexity of conformal field theories and redefine them as basic and simple mathematical objects. Topological quantum field theory holds out a really hopeful new possibility for axiomatic approaches to quantum field theory which will capture the essential geometric ideas for mathematicians without bogging down in analytical contradictions.

PREDICTION 4. *And after theoretical physics ?*

We are going through a period where the primary outside influence on differential geometry has been either cosmology or fundamental physics. We learn the importance of 1, 2, 3, 4, 10, 26, and  $\infty$  dimensions (i.e., we study small and infinite-dimensional manifolds). We should remember there are

other sources of inspiration. How about robotics, which must be done in large but finite dimensions and which is too complicated to be exact?

PREDICTION 5. *Return to Boubaki.*

My last comment is on style. I was generally taught in the famous Boubakist style of definition, theorem, proof, and maybe example if there is time. Coordinate descriptions were out: abstraction was in. Generalization abounded. If you cannot do it in  $n$  dimensions, do not bother. This approach dates back at least to Hilbert and his famous problems. My own mathematical interests and the predominant mathematical style today has become far more oriented towards the particular. One can also identify this trend in the talks in this Centennial Celebration. Coordinate descriptions are a universal language. Low dimensional topology is central. Lots of us think by example. I do consider it unfortunate that examples are basically the only thing in many useful subjects in mathematics which I personally do understand. This evolution to the particular has been an important part of the opening up and reaching out of the last twenty years.

Many scientists complain to me of the mathematical style of teaching of twenty years ago. The outside world has barely yet understood the change! I think this change has evolved along with the practical experiences of our universal teaching experience. It is here the world may have had its effect on mathematics: through our calculus students.

I see signs of reversion to the general in the next generation of students. They again think coordinates, low dimensions, and  $SU(2)$  are old-fashioned. This is as it should be. They have their own mathematics and its style to discover.

### BASIC REFERENCE BOOKS

- M. Atiyah, *Geometry of Yang-Mills fields*, Academia Nazionale-S. N. S., Pisa, 1979.
- A. Jaffe and C. Taubes, *Vortices and monopoles*, Progress in Physics, no. 2, Birkhäuser, 1981.
- H. B. Lawson, *Theory of gauge fields in 4-D*, CBMS, no. 58, Amer. Math. Soc., Providence, RI, 1985.
- D. Freed and K. Uhlenbeck, *Instantons and four manifolds*, MSRI publication, no. 1, Springer, 1990.
- S. Kobayashi, *Differential geometry of complex vector bundles*, Princeton Univ. Press, Princeton, NJ, 1987.
- M. Atiyah and N. Hitchin, *Scattering of magnetic monopoles*, Princeton Univ. Press, Princeton, NJ, 1989.
- S. Donaldson and P. Kronheimer, *Geometry of four-manifolds*, Clarendon Press, Oxford, 1990.
- M. Atiyah, *Collected works*, Vol. 5, Oxford Univ. Press, 1988.
- Y. Manin, *Gauge field theory and complex geometry*, Grundlehren Math. Wiss., no. 289, Springer-Verlag, 1988.
- Y. T. Siu, *Lectures on Hermitian-Einstein metrics for stable bundles and Kähler-Einstein metrics*, DMV Seminar 8, Birkhauser, 1989.

## REFERENCES

- [A1] M. Atiyah, *The Yang-Mills equations and the structure of 4-manifolds*, Durham Sympos. on Global Riemannian Geometry, Ellis Horwood, 1984, pp. 11–17.
- [A2] —, *New invariants of 3 and 4 manifolds*, The Mathematical Heritage of Hermann Weyl (R. O. Wells, Jr., ed.), Proc. Sympos. Pure Math., vol. 48, Amer. Math. Soc., Providence, RI, 1988, pp. 285–300.
- [B] S. Bradlow, *Vortices on Kähler manifolds*, Ph.D. Thesis, University of Chicago, August 1988.
- [D1] S. K. Donaldson, *The Yang-Mills Equations on Euclidean space*, Perspectives in Mathematics, Birkhauser-Verlag, 1984, pp. 93–109.
- [D2] —, *Infinite determinates, stable bundles and curvatures*, Duke Math. J. **54** (1987), 231–247.
- [Do] J. Douglas, *Solution to the problem of Plateau*, Trans. Amer. Math. Soc. **33** (1931), 263–321.
- [FGZ] I. Frenkel, H. Garland, and G. Zuckerman, *Semi-infinite cohomology and string theory*, Proc. Nat. Acad. Sci. U.S.A. **83** (1986), 8442–46.
- [FM] R. Friedman and J. Morgan, *Algebraic surfaces and four-manifolds, some conjectures and speculations*, Bull. Amer. Math. Soc. (N.S.) **18** (1988), 1–20.
- [HMRVW] A. Hey, J. Merlin, M. Ricketts, M. Vaughn, and D. Williams, *Topological solutions in gauge theory and their computer graphic representation*, Science **240** (May 1988), 1163–68.
- [H] D. Hoffman, *The computer-aided discovery of new embedded minimal surface*, Math. Intelligencer **9** (1987), 8–21.
- [MY] W. Meeks III and S. T. Yau, *Topology of three dimensional manifolds and the embedding problems in minimal surface theory*, Ann. of Math. (2) **112** (1980), 441–485.
- [SU] J. Sacks and K. Uhlenbeck, *The existence of minimal 2-spheres*, Ann. of Math. (2) **113** (1981), 1–24.
- [Se] S. Sedlacek, *A direct method for minimizing the Yang-Mills functional on 4-manifolds*, Comm. Math. Phys. **86** (1982), 515–527.
- [S] C. Simpson, *Constructing variations of Hodge structure using Yang-Mills theory*, J. Amer. Math. Soc. **1** (1988), 867–918.
- [T1] C. Taubes,  *$O(2)$  Symmetric connections in an  $SU(2)$  Yang-Mills theory*, Comm. Math. Phys. **69** (1979), 179–193.
- [T2] —, *Arbitrary  $N$ -vortex solutions to the first order Ginzburg-Landau equations*, Comm. Math. Phys. **72** (1980), 277–292.
- [UY] K. Uhlenbeck and S. T. Yau, *On the existence of Hermitian Yang-Mills connections*, Comm. Pure Appl. Math. **34** (1986), S257–S293.
- [Y] S. T. Yau (ed.), *Seminar on differential geometry*, Ann. of Math. Stud., no. 102, Princeton Univ. Press, Princeton, NJ, 1982.

DEPARTMENT OF MATHEMATICS, THE UNIVERSITY OF TEXAS, AUSTIN, TEXAS 78712



## Geometry and Quantum Field Theory

EDWARD WITTEN

**1. Introduction.** First of all, I would like to thank the American Mathematical Society for inviting me to lecture here on this occasion, and to thank the organizers for arranging such a stimulating meeting. And I would like to echo the sentiments of some previous speakers, who expressed the wish that we will all meet here in good health on the 150th anniversary of the American Mathematical Society, to hear the younger mathematicians explain the solutions of some of the unsolved problems posed this week.

It is a challenge to try to speak about the relation of quantum field theory to geometry in just one hour, because there are certainly many things that one might wish to say. The relationship between theoretical physics and geometry is in many ways very different today than it was just ten or fifteen years ago. It used to be that when one thought of geometry in physics, one thought chiefly of classical physics—and in particular of general relativity—rather than quantum physics. Geometrical ideas seemed (except perhaps to some visionaries) to be far removed from quantum physics—that is, from the bulk of contemporary physics. Of course, quantum physics had from the beginning a marked influence in many areas of mathematics—functional analysis and representation theory, just to mention two. But it would probably be fair to say that twenty years ago the day to day preoccupations of most practicing theoretical elementary particle physicists were far removed from considerations of geometry.

Several important influences have brought about a change in this situation. One of the principal influences was the recognition—clearly established by the middle 1970s—of the central role of nonabelian gauge theory in elementary particle physics. The other main influence came from the emerging study of supersymmetry and string theory. Of course, these different influences are inter-related, since nonabelian gauge theories have elegant supersymmetric

---

1980 *Mathematics Subject Classification* (1985 Revision). Primary 81E13, 81E99.

Research supported in part by NSF Grant 86-20266 and NSF Waterman Grant 88-17521.

©1992 American Mathematical Society  
0-8218-0167-8 \$1.00 + \$.25 per page

generalizations, and in string theory these appear in a fascinating new light. Bit by bit, the study of nonabelian gauge theories, supersymmetry, and string theory have brought new questions to the fore, and encouraged new ways of thinking.

An important early development in this process came in the period 1976–77 with the recognition that the Atiyah-Singer index theorem was the proper context for understanding some then current developments in the theory of strong interactions. (In particular, the solution by Gerard't Hooft [1] of the “U(1) problem,” a notorious paradox in strong interaction theory, involved Yang-Mills instantons, originally introduced in [2], and “fermion zero modes” whose proper elucidation involves the index theorem.) Influenced by this and related developments, physicists gradually learned to think about quantum field theory in more geometrical terms. As a bonus, ideas coming at least in part from physics shed new light on some mathematical problems. In the first stage of this process, the purely mathematical problems that arose (at least, those that had motivations independent of quantum field theory, and in which progress could be made) involved “classical” mathematical concepts—partial differential equations, index theory, etc.—where physical considerations suggested new questions or a new point of view.

In the talk just before mine, Karen Uhlenbeck described some purely mathematical developments that at least roughly might be classified in this area. She described advances in geometry that have been achieved through the study of systems of nonlinear partial differential equations. Among other things, she sketched some aspects of Simon Donaldson’s work on the geometry of four-manifolds [3], in which dramatic advances have been made by studying the moduli spaces of instantons—solutions, that is, of a certain nonlinear system of partial differential equations, the self-dual Yang-Mills equations, which were originally introduced by physicists in the context of quantum field theory [2].

If “classical” objects (such as instantons) that arise in quantum field theory could be so interesting mathematically, one might well suspect that mathematicians will soon find the quantum field theories themselves, and not only the “classical” objects that they give rise to, to be of interest. Such a question was indeed raised by Karen Uhlenbeck at the end of her talk, and is much in line with the perspective offered by Michael Atiyah in [4], which was the starting point for many of my own efforts.

I will talk today about three areas of recent interest where quantum field theory seems to be the right framework for thinking about a problem in geometry:

(1) Our first problem will be to explain the unexpected occurrence of modular forms in the theory of affine Lie algebras. This problem, which was described the other day by Victor Kac, has two close cousins—to explain “monstrous moonshine” in the theory of the Fischer-Griess monster group [5, 6], and to account for the surprising role of modular forms in algebraic

topology [7], about which Raoul Bott spoke briefly at the end of his talk. Quantum field theory supplies a more or less common explanation for these three phenomena, but the first requires the least preliminary explanation, and it is the one that I will focus on.

(2) The second problem is to give a geometrical definition of the new knot polynomials—the Jones polynomial and its generalizations—that have been discovered in recent years. The essential properties of the Jones polynomial have been described to us the other day by Vaughn Jones.

(3) The third problem is to get a more general insight into Donaldson theory of four-manifolds—which was sketched in the last hour by Karen Uhlenbeck—and the closely related Floer groups of three-manifolds. Here again there are lower dimensional cousins, namely the Casson invariant of three-manifolds, Gromov's theory of maps of a Riemann surface to a symplectic manifold, and Floer's closely related work on fixed points of symplectic diffeomorphisms. But among these formally rather analogous subjects, I will concentrate on Donaldson/Floer theory.

**2. Physical Hilbert spaces and transition amplitudes.** Let us sketch these three problems in a little more detail. In the first problem, one considers the group  $\mathcal{L}G$  of maps  $S^1 \rightarrow G$ , where  $G$  is a finite-dimensional compact Lie group, and  $S^1$  is the ordinary circle. The representations of  $\mathcal{L}G$  with “good” properties, analogous to the representations of compact finite-dimensional groups, are the so-called integrable highest weight representations (see [8, 9] for introductions.) These representations are rigid (no infinitesimal deformations). From this it follows that any connected group of outer automorphisms of  $\mathcal{L}G$  must act at least projectively on any integrable highest weight representation  $\mathcal{R}$  of  $\mathcal{L}G$ . In fact, the group  $\text{diff} S^1$  of diffeomorphisms of  $S^1$  acts on  $\mathcal{L}G$  by outer automorphisms and acts projectively on the integrable highest weight representations. Thus, in particular, the vector field  $d/d\theta$  that generates an ordinary rotation of  $S^1$  is represented on  $\mathcal{R}$  by some operator  $H$ .

One computes in such a representation the “character”

$$(2.1) \quad F_{\mathcal{R}}(q) = \text{Tr}_{\mathcal{R}} q^H$$

(here  $q$  is a complex number with  $|q| < 1$ ), and one finds this to be a modular function with a simple transformation law under a suitable congruence subgroup of the modular group. Setting  $q = \exp(2\pi i\tau)$ , the modular group is of course the group  $\text{PSL}(2, \mathbb{Z})$  of fractional linear transformations

$$(2.2) \quad \tau \mapsto \frac{a\tau + b}{c\tau + d},$$

of the upper half-plane, with  $a, b, c, d \in \mathbb{Z}$  and  $ad - bc = 1$ . I will not enter here into the complicated question of exactly what kind of modular functions the characters (2.1) are. (One simple, general statement, which

from one point of view is the statement that comes most directly from quantum field theory, is that the  $F_{\mathcal{R}}(q)$ , with  $\mathcal{R}$  running over all highest weight representations of fixed "level", transform as a unitary representation of the full modular group  $\mathrm{PSL}(2, \mathbb{Z})$ .

To understand the significance of the modularity of the characters  $F_{\mathcal{R}}(q)$ , let us recall that the group  $\mathrm{SL}(2, \mathbb{Z})$  has a natural interpretation as the (orientation preserving) mapping class group of a two-dimensional torus  $\mathbb{T}^2$ . Thus, we interpret  $\mathbb{T}^2$  as the quotient of the  $x-y$  plane by the equivalence relations  $(x, y) \sim (x+1, y)$  and  $(x, y) \sim (x, y+1)$ . Clearly, if  $a, b, c$ , and  $d$  are integers such that  $ad-bc=1$ , the formula  $(x, y) \rightarrow (ax+by, cx+dy)$  gives a diffeomorphism of  $\mathbb{T}^2$  to itself, and every orientation preserving diffeomorphism of  $\mathbb{T}^2$  is isotopic to a unique one of these. Thus,  $\mathrm{SL}(2, \mathbb{Z})$  can be considered in this sense to arise as a group of diffeomorphisms of a two-dimensional surface.

Thus, while it is natural that the one-dimensional symmetry group  $\mathrm{diff}^1 \mathbf{S}^1$  plays a role in the representation theory of the loop group  $\mathcal{L}G$ , the appearance of  $\mathrm{SL}(2, \mathbb{Z})$  means that in fact a kind of *two-dimensional symmetry* appears in this theory. Our first problem—modular moonshine in the theory of affine Lie algebras—is the problem of explaining the origin of this two-dimensional symmetry.

Now we move on to our second problem. A *braid* is a time dependent history of  $n$  points  $\mathbb{R}^2$ , which are required, up to a permutation, to end where they begin (Figure 1(a)). Braids with  $n$  strands form a group, the Artin braid group  $\mathcal{B}_n$ , with an evident law of composition, sketched in Figure 1(b). From a braid one can make a knot (or in general a link) by gluing together the top and bottom as in Figure 1(c). Although every braid gives in this way a unique link, the converse is not so; the same link may arise from many different braids. The crucial difference between braids and links is the following. Braids are classified up to time dependent diffeomorphisms of  $\mathbb{R}^2$  (that is, up to diffeomorphisms of  $\mathbb{R}^3$  that leave fixed one of the coordinates, the "time"  $t$ ), while links are classified up to full three-dimensional diffeomorphisms.

If one is given a representation  $\mathcal{S}$  of the braid group  $\mathcal{B}_n$ , and a braid  $B \in \mathcal{B}_n$ , then  $\mathrm{Tr}_{\mathcal{S}} B$  (the trace of the matrix that represents  $B$  in the representation  $\mathcal{S}$ ) is an invariant of the *braid*  $B$  (and depends in fact only on its conjugacy class in  $\mathcal{B}_n$ ), but there is no reason for it to be an invariant of the *link* that is obtained by joining the ends of the braid  $B$  according to the recipe in the figure.

Nevertheless, Vaughn Jones found a special class of representations of the braid group with the magic property that suitable linear combinations of the braid traces are in fact knot invariants and not just braid invariants. These knot invariants can be combined into the Jones polynomial, some of whose remarkable properties were described in Jones's lecture the other

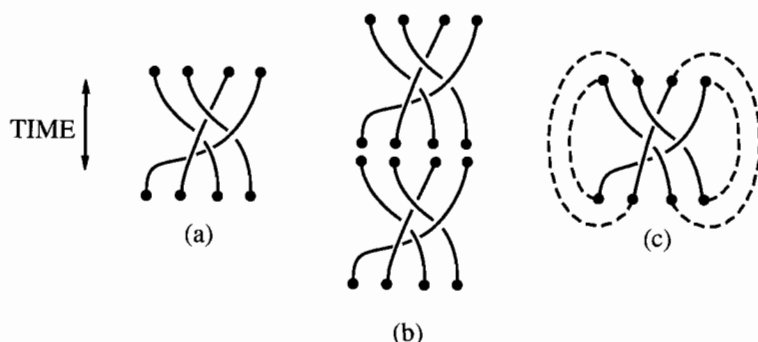


FIGURE 1. (a) A braid; (b) composition of two braids; (c) making a braid into a link.

day. The discovery of the Jones polynomial stimulated in short order the discovery of some related knot polynomials—the HOMFLY and Kauffman polynomials—whose logical status is rather similar. The challenge of understanding the Jones polynomial is to explain why the Jones braid representations, which obviously have two-dimensional symmetry, should really have three-dimensional symmetry.

Thus, we have two examples where one studies a *group representation* that obviously has  $d$ -dimensional symmetry, for some  $d$ , but turns out to have  $(d + 1) = D$ -dimensional symmetry, for reasons that might look mysterious. In our first example, the group is  $\mathcal{L}G$ ,  $d = 1$ , and  $D = 2$ . In the second example, the group is the braid group,  $d = 2$ , and  $D = 3$ .

Our third example, Donaldson/Floer theory, is of a somewhat different nature. In this case,  $d = 3$  and  $D = 4$ , but unlike our previous examples, Donaldson/Floer theory began historically not in the lower dimension but in the upper dimension. The mathematical theory begins in this case with Donaldson's invariants of a closed, oriented four-manifold  $M$ . In Donaldson's original considerations, it was important that the boundary of  $M$  should vanish. The attempt to generalize the Donaldson invariants to the case that  $\partial M = Y \neq \emptyset$  led to the introduction of the "Floer homology groups" which are vector spaces  $\text{HF}^*(Y)$  canonically associated with an oriented three-manifold  $Y$ . Though these vector spaces did not originate as group representations, their formal role is just like that of the group representations that entered in our first two examples.

In our first two problems of understanding modular moonshine and the Jones polynomial, the crucial question is to explain why  $(d + 1)$ -dimensional symmetry is present in a construction that appears to only have  $d$ -dimensional symmetry. At least from a historical point of view, Donaldson theory is of a completely different nature, since the four-dimensional symmetry has been built in from the beginning. Nevertheless, the logical structure of Donaldson/Floer theory is of a similar nature to that of the first two examples.

In each of our three examples, a pair of dimensions,  $d$  and  $D = d + 1$ ,

plays a key role. With the lower dimension we associate a vector space (the representations of  $\mathcal{L}G$  or  $\mathcal{B}_n$  or the Floer groups) and with the upper dimension we associate an invariant (the characters  $\text{Tr}_{\mathcal{R}} q^H$ , the knot polynomials, or the Donaldson invariants of four-manifolds). The facts are summarized in Table 1.

TABLE 1

Theory	Dimensions	Vector space in lower dimension	Invariant in upper dimensions
Modular moonshine	1, 2	Representations of loop groups	Modular forms $\text{Tr } q^H$
Jones polynomial	2, 3	Jones representations of braid group	Invariants of knots and three-manifolds
Donaldson/Floer theory	3, 4	Floer Homology groups	Donaldson invariants of four-manifolds

**3. Axioms of quantum field theory.** Let us now formalize the precise relationship between the vector spaces that appear in dimension  $d$  and the invariants in dimension  $d + 1$ . (In the physical context,  $d$  is called the dimension of space, and  $d + 1$  is the dimension of space-time.) In formalizing this relationship, we will follow axioms originally proposed (in the context of conformal field theory, essentially our first example) by Graeme Segal [10]. (In addition, Michael Atiyah has adapted those axioms for the topological context that is relevant to our second and third examples [11], with considerably more precision than I will attempt here.)

So we will consider quantum field theory in space-time dimension  $D = d + 1$ . The manifolds that we consider will be smooth manifolds possibly endowed with some additional structure. The type of additional structure considered will be characteristic of the theory. For instance, in the case of modular moonshine, this additional structure is a conformal structure; quantum field theories requiring such a structure (but not requiring a choice of Riemannian metric) are called conformal field theories. In the case of Donaldson/Floer theory, the extra structure consists of an orientation; in the case of the Jones polynomial, one requires an orientation and “framing” of tangent bundles (in a suitable stable sense). Theories that require structure of such a purely topological kind may be called topological quantum field theories. The “ordinary” quantum field theories most extensively studied by physicists require metrics on all manifolds considered.

The first notion is that to every  $d$ -dimensional manifold  $X$ , without boundary, and perhaps with some additional structure characteristic of the particular theory, one associates a vector space  $\mathcal{H}_X$ . A quantum field theory is said to be “unitary” if these vector spaces actually carry a Hilbert space structure; this is so in the theories of modular moonshine and the Jones poly-

nomial, but not in the case of Donaldson/Floer theory. In the case of the Jones polynomial and Donaldson/Floer theory, the vector spaces  $\mathcal{H}_X$  are finite dimensional, and a morphism of vector spaces is taken to mean an arbitrary linear transformation (preserving the unitary structure in the case of the Jones polynomial); in the theory of modular moonshine, the  $\mathcal{H}_X$  are infinite dimensional, and it is necessary to be more precise about what is meant by a morphism among these spaces.

In Segal's language, the association  $X \rightarrow \mathcal{H}_X$  is to be a functor from the category of  $d$ -dimensional manifolds with additional structure (and diffeomorphisms preserving the specified structures) to the category of vector spaces (and linear transformations of the appropriate kind).

Certain additional restrictions are imposed. The empty  $d$ -manifold  $\emptyset$  is permitted, and one requires that  $\mathcal{H}_\emptyset = \mathbb{C}$  ( $\mathbb{C}$  here being a one-dimensional vector space with a preferred generator which we call "1"). If  $X \amalg Y$  denotes the disjoint union of two  $d$ -dimensional manifolds  $X$  and  $Y$ , then one requires  $\mathcal{H}_{X \amalg Y} = \mathcal{H}_X \otimes \mathcal{H}_Y$ . If  $-X$  is  $X$  with opposite orientation, and  $*$  denotes the dual of a vector space, one requires  $\mathcal{H}_{-X} = \mathcal{H}_X^*$ .

Since the late 1920s, the spaces  $\mathcal{H}_X$  have been known to physicists as the "physical Hilbert spaces" (of the particular quantum field theory under consideration). The association  $X \rightarrow \mathcal{H}_X$  is roughly half of the basic structure considered in quantum field theory. The second half corresponds in physical terminology to the "transition amplitudes."

To introduce the transition amplitudes, we consider (Figure 2(a) on the next page) a cobordism of oriented (and possibly disconnected or empty)  $d$ -dimensional manifolds. Such a cobordism is defined by an oriented  $(d+1)$ -dimensional manifold  $W$  whose boundary is, say,  $\partial W = X \cup (-Y)$ , where  $X$  and  $Y$  are oriented  $d$ -dimensional manifolds (whose orientations respectively agree or disagree with that induced from  $W$ ). It is required that whatever structure (conformal structure, framing, metric, etc.) has been introduced on  $X$  and  $Y$  is extended over  $W$ . Such a cobordism is regarded as a morphism from  $X$  to  $Y$ . To every such morphism of manifolds, a quantum field theory associates a morphism of vector spaces

$$(3.1) \quad \Phi_W: \mathcal{H}_X \rightarrow \mathcal{H}_Y.$$

Of course, this association  $W \rightarrow \Phi_W$  should be natural, invariant under any diffeomorphism of  $W$  that preserves the relevant structures. Regarding  $-W$  as a morphism from  $-Y$  to  $-X$ , one requires that  $\Phi_{(-W)}: \mathcal{H}_{(-Y)} \rightarrow \mathcal{H}_{(-X)}$  should be the dual linear transformation to  $\Phi_W$ . And if  $W = W_1 \cup W_2$  is a composition of cobordisms (Figure 2(b) on the next page), one requires that

$$(3.2) \quad \Phi_W = \Phi_{W_2} \circ \Phi_{W_1}.$$

These requirements correspond physically to relativity, locality, and causality.

A very important special case of this is the case in which  $W$  is a closed  $D = (d+1)$ -dimensional manifold without boundary. Such a  $W$  can be

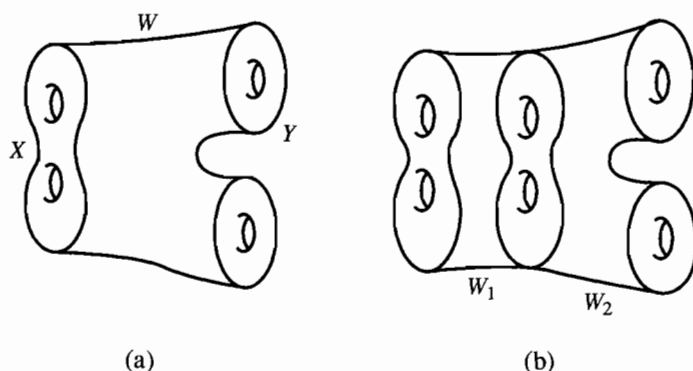


FIGURE 2. (a) A cobordism of oriented  $d$ -dimensional manifolds; (b) a composition of such cobordisms.

regarded as a morphism from the empty  $d$ -dimensional manifold  $\emptyset$  to itself. Since  $\mathcal{H}_{\emptyset} = \mathbb{C}$ , the associated morphism  $\Phi_W: \mathcal{H}_{\emptyset} \rightarrow \mathcal{H}_{\emptyset}$  is simply a number, which for physicists is often called the partition function of  $W$  and denoted  $Z(W)$ . This partition function is the fundamental invariant in quantum field theory; for different choices of theory, one gets the invariants of  $D$ -dimensional manifolds indicated in the last column in Table 1.

For  $Z(W)$  to be defined in a given quantum field theory,  $W$  must of course be endowed with the structure appropriate to the particular quantum field theory in question. For instance, in the case of modular moonshine,  $W$  must be a Riemann surface with a conformal structure. In genus one, this means that  $W$  is an elliptic curve, which can be represented by a point in the upper half-plane subject to the action of the mapping class group. The naturality of the association  $W \rightarrow Z(W)$  means that  $Z(W)$  can depend only on the equivalence class of the conformal structure of  $W$ , and it is this which leads to modular forms. In our other two examples, no metric or conformal structure is present, so we are dealing with topological invariants. In our second example of the Jones polynomial, the invariant  $Z(W)$  is an invariant of oriented three-manifolds which is an analog for three-manifolds of the Jones polynomials for knots in  $S^3$ . (The actual knot invariants can be obtained by an elaboration of the quantum field theory structure.) In our third example of Donaldson theory, the invariant  $Z(W)$  is the prototype of the invariants that appear in the celebrated Donaldson polynomials of oriented four-manifolds.

It is built into the axioms of quantum field theory that the fundamental invariants  $Z(W)$  can be computed from a decomposition of the type that is known in the case of three-manifolds as a Heegaard splitting. This means a realization of  $W$  as  $W = W_1 \cup W_2$ , where  $W_1$  and  $W_2$  are  $D$ -manifolds joined together along their common boundary  $\Sigma$ . In this case the morphism



$W$  from the empty manifold  $\emptyset$  to itself factorizes as morphism  $W_2$  from  $\emptyset$  to  $\Sigma$  composed with a morphism  $W_1$  from  $\Sigma$  to  $\emptyset$ , i.e.,

$$(3.3) \quad \Phi_W = \Phi_{W_1} \circ \Phi_{W_2}.$$

If  $1$  is the canonical generator of  $\mathcal{H}_\emptyset$ , we then have

$$(3.4) \quad Z(W) = (1, \Phi_W \cdot 1) = (1, \Phi_{W_1} \circ \Phi_{W_2} \cdot 1).$$

Let  $v \in \mathcal{H}_\Sigma$  be the vector  $v = \Phi_{W_2}(1)$ . Also, think of  $-W_1$  as a morphism from  $\emptyset$  to  $-\Sigma$ , and let  $w \in \mathcal{H}_{-\Sigma}$  be the vector  $w = \Phi_{-W_1}(1)$ . Then (3.4) amounts to

$$(3.5) \quad Z(W) = (w, v).$$

This ability to calculate via Heegaard splittings is part of the conventional *definition* of the Casson invariant (which has a quantum field theory interpretation analogous to that of Donaldson theory), and is essential in the calculability of the three-manifold invariants that are related to the Jones polynomial. Likewise, in the case of modular moonshine, the decomposition (3.5) is the key to the fact that the partition function  $Z(W)$  can be written as the character  $\text{Tr}_{\mathcal{H}} q^H$  of equation (2.1).

**4. Construction of quantum field theories.** The question arises, of course, of how these quantum field theories are to be constructed. About this enormous subject it is possible only to say a few words here.

The starting point is always the choice of an appropriate *Lagrangian*, which is the integral of a local functional of appropriate fields. For instance, if one is interested in understanding the Jones polynomial, one picks a finite-dimensional compact simple group  $G$  and one considers a connection  $A$  on a  $G$ -bundle  $E$  over a three-manifold  $M$ . Let  $F = dA + A \wedge A$  denote the curvature of this connection. On the Lie algebra  $\mathcal{G}$  of a compact group  $G$ , there is an invariant quadratic form which we denote by the symbol  $\text{Tr}$  (that is, we write  $(a, b) = \text{Tr}(ab)$ ).<sup>1</sup> For the Lagrangian, we take the Chern-Simons invariant of the connection  $A$ :

$$(4.1) \quad \mathcal{L} = \frac{k}{4\pi} \int_M \text{Tr} \left( A \wedge dA + \frac{2}{3} A \wedge A \wedge A \right).$$

(Here  $k$  is a positive integer, a fact that is required so that the argument  $e^{i\mathcal{L}}$  in the Feynman path integral is gauge invariant.) To construct a quantum field theory from this Lagrangian, there are two basic requirements. First, we must construct a functor from Riemann surfaces  $\Sigma$  to Hilbert spaces  $\mathcal{H}_\Sigma$ ; and second, for every cobordism  $W$  from  $\Sigma$  to  $\Sigma'$ , we must construct a morphism  $\Phi_W: \mathcal{H}_\Sigma \rightarrow \mathcal{H}_{\Sigma'}$ .

<sup>1</sup> The quadratic form is to be normalized so that the characteristic class  $\frac{1}{4\pi} \text{Tr} F \wedge F$  has periods that are multiples of  $2\pi$ .

For the first step, one proceeds as follows. Given the surface  $\Sigma$ , we consider the Lagrangian (4.1) on the three-manifold  $\Sigma \times \mathbb{R}$ . The space of critical points of the Lagrangian, up to gauge transformations, is known in classical mechanics as the "phase space" of the system under investigation. Let us call this phase space  $\mathcal{M}_\Sigma$ . In the case at hand, the Euler-Lagrange equation for a critical point of the Lagrangian (4.1) is the equation  $F = 0$ , where  $F = dA + A \wedge A$  is the curvature of the connection  $A$ . (That is, (4.1) is invariant to first order under variations of the connection of compact support if and only if  $F = 0$ .) A flat connection on  $\Sigma \times \mathbb{R}$  defines a homomorphism of the fundamental group  $\pi_1(\Sigma \times \mathbb{R})$  into  $G$ . Of course, this is the same as a homomorphism of  $\pi_1(\Sigma)$  into  $G$ . The classical phase space  $\mathcal{M}_\Sigma$  associated with the Lagrangian (4.1) is simply the moduli space of homomorphisms of  $\pi_1(\Sigma) \rightarrow G$ , up to conjugation by  $G$ .

Now, it is a general fact in the calculus of variations that the phase space associated with a Lagrangian such as (4.1) is always endowed with a canonical symplectic structure  $\omega$ . Indeed, this is how symplectic structures originally appeared in classical mechanics, and as such it was the starting point of symplectic geometry as a mathematical subject also. In the case at hand, the symplectic structure thus obtained in  $\mathcal{M}_\Sigma$  is known [13], and has been studied very fruitfully from the point of view of two-dimensional gauge theory [14], but my point is that this symplectic structure on  $\mathcal{M}_\Sigma$  can be considered to arise from a *three-dimensional* variational problem. This elementary fact is an important starting point for understanding the mysterious three-dimensionality of the Jones polynomial.

Once the appropriate phase space  $\mathcal{M}_\Sigma$  is identified, the association  $\Sigma \rightarrow \mathcal{H}_\Sigma$  is made by "quantizing"  $\mathcal{M}_\Sigma$  to obtain a Hilbert space  $\mathcal{H}_\Sigma$ . Geometric quantization is not sufficiently well developed to make quantization straightforward in general (or perhaps this is actually impossible in general), but in the case at hand quantization can be carried out by choosing a complex polarization of  $\mathcal{M}$ . This is accomplished by picking a complex structure  $J$  on  $\Sigma$  and using the Narasimhan-Seshadri theorem to identify  $\mathcal{M}$  with the moduli space of stable holomorphic  $G_\mathbb{C}$  bundles over  $\Sigma$ . This moduli space is then quantized by defining  $\mathcal{H}_\Sigma$  to be the space of holomorphic sections of a certain line bundle over  $\mathcal{M}$ . This space is independent of  $J$  (up to a projective factor) because of its interpretation in terms of quantization of the underlying classical phase space  $\mathcal{M}$ . The association  $\Sigma \rightarrow \mathcal{H}_\Sigma$  is the geometric origin of the Jones braid representations (or rather their analog for the mapping class group in genus  $g$ ).

Once the association  $\Sigma \rightarrow \mathcal{H}_\Sigma$  is understood, it remains to define for every cobordism  $W$  from  $\Sigma$  to  $\Sigma'$ , a corresponding morphism  $\Phi_W: \mathcal{H}_\Sigma \rightarrow \mathcal{H}_{\Sigma'}$ . The key notion here is the "Feynman path integral," that is, Feynman's concept of integration over the whole function space  $\mathcal{V}$  of connections on (the given  $G$ -bundle  $E$  over)  $W$ . Roughly speaking, the function space integral with prescribed boundary conditions gives the kernel of the morphism  $\Phi_W$ .

I have tried to explain heuristically the role of functional integrals in [12], and I will not repeat those observations here.

In conclusion, let me point out that if  $G$  is a *compact* group, then, as I have argued in [15], the quantum field theory associated with the Lagrangian (4.1) is related to the Jones polynomial and its generalizations. However, (4.1) makes sense for any gauge group  $G$  with an invariant quadratic form “Tr” on the Lie algebra. It is natural to ask what mathematical constructions are related to the quantum field theories so obtained. One case that can be conveniently analyzed is the case in which one replaces the compact group  $G$  by a group  $TG = \mathcal{G} \ltimes G$ ; here  $\mathcal{G} \ltimes G$  denotes the semidirect product of  $G$  with its own Lie algebra  $\mathcal{G}$ , the latter regarded as an abelian group acted on by  $G$ . It turns out [16] that with this choice of gauge group, the quantum field theory derived from (4.1) (with a certain choice of “Tr”) is related to recent work of D. Johnson on Reidemeister torsion [17], while if instead one considers a certain super-group whose bosonic part is  $TG$  then (4.1) (again with a certain choice of “Tr”) is related to the Casson invariant of three-manifolds. It is also very interesting to take  $G$  to be a semisimple but noncompact Lie group. The corresponding quantum field theories are very little understood, but there are indications that they should be very rich. In fact, it appears [18] that the theories based on  $SL(2, \mathbb{R})$  and especially  $SL(2, \mathbb{C})$  must be intimately connected with the theory of hyperbolic structures on three-manifolds, as surveyed the other day by Thurston.

**5. Conclusion.** In attending this meeting, I have found it striking how many of the lectures were concerned with questions that are associated with quantum field theory—and in many cases questions that might be characterized as questions about quantum field theory. In time we will hopefully gain a clearer picture of the scope of some of these newly emerging relations between geometry and physics. It is not too much to anticipate that many important constructions relating quantum field theory to topology and differential geometry remain to be discovered. Harder to foresee is whether—by the time of the one hundred fiftieth anniversary of the American Mathematical Society—the influence of quantum field theory will also extend to other areas of mathematics, such as algebraic geometry and number theory, which superficially might appear to be comparatively immune. Let me recall that in his lecture earlier this week, Dick Gross concluded by urging physicists and mathematicians to find a quantum field theory explanation for the appearance of modular forms in the study of the  $L$ -functions of elliptic curves. (This question was, of course, motivated by the relation of quantum field theory to the different kinds of modular moonshine.) Perhaps this challenge, or analogous ones about which one might speculate, will be met. Hints today concerning quantum field theoretic insights about number theory are probably no more compelling than hints of quantum field theoretic insight about differential geometry were ten years ago.

What significance might the emerging links between quantum field theory and geometry have for physics? It is very noticeable that the aspects of quantum field theory that are most useful in understanding the geometrical problems that I have been talking about are pretty close to the slightly specialized aspects of quantum field theory that appear in string theory. Modular invariance in the theory of affine Lie algebras is certainly a familiar story to string theorists. The Jones polynomial and its generalizations are related to the "rational conformal field theories" which are one of our main tools for finding exact classical solutions in string theory. The constructions that enter in formulating Donaldson theory as a quantum field theory are also very similar to what string theorists are accustomed to (in the use of world-sheet BRST operators).

Apart from being at least loosely connected with all of the geometrical problems that we have been discussing, string theory seems to be the center of some geometrical questions of central physical interest. The towering puzzle in contemporary theoretical physics is—at least from my standpoint—the puzzle of finding the geometrical context in which string theory should be properly formulated and understood. I am sure many physicists would share this judgment. With our present concepts, this problem (to which I attempted a thumbnail introduction in [12]) seems well out of reach. Perhaps it is not too far-fetched to hope that some insight in this mystery can be obtained from the further study of geometrical questions arising in quantum field theory.

## REFERENCES

1. Gerard 't Hooft, *Computation of the quantum effects due to a two dimensional pseudoparticle*, Phys. Rev. D **14** (1976), 3432.
2. A. Belavin, A. M. Polyakov, A. Schwarz, and Yu. S. Tjupkin, Phys. Lett. B **59** (1975), 85.
3. S. Donaldson, *An application of gauge theory to the topology of four manifolds*, J. Differential Geom. **18** (1983), 269; *Polynomial invariants for smooth four manifolds*, preprint, Oxford University.
4. M. F. Atiyah, *New invariants of three and four dimensional manifolds*, The Mathematical Heritage of Hermann Weyl (R. Wells, ed.), Amer. Math. Soc., Providence, RI, 1988.
5. J. H. Conway and S. P. Norton, *Monstrous moonshine*, Bull. London Math. Soc. **11** (1979), 308.
6. I. B. Frenkel, A. Meurman, and J. Lepowsky, *A moonshine module for the monster*, Vertex Operators in Mathematics and Physics (J. Lepowsky, S. Mandelstam, and I. M. Singer, eds.), Springer-Verlag.
7. P. Landweber (ed.), *Elliptic curves and modular forms in algebraic topology*, Lecture Notes in Math., vol. 1326, Springer-Verlag, 1988.
8. V. Kac, *Infinite dimensional Lie algebras*, Cambridge Univ. Press, 1985.
9. A. Pressley and G. Segal, *Loop groups*, Oxford Univ. Press, 1987.
10. G. Segal, preprint, Oxford University (to appear).
11. M. F. Atiyah, *Topological quantum field theories*, René Thom Festschrift (to appear).
12. E. Witten, *Geometry and physics*, Proc. Internat. Congr. Math., Berkeley, California, 1986.
13. W. Goldman, *Invariant functions on Lie groups and Hamiltonian flows of surface group representations*, Invent. Math. **85** (1986), 263.

14. M. F. Atiyah and R. Bott, *Yang-Mills equations over Riemann surfaces*, Philos. Trans. Roy. Soc. London. Ser. A **308** (1982), 523.
15. E. Witten, *Quantum field theory and the Jones polynomial*, Comm. Math. Phys. **121** (1989), 351.
16. —, *Topology-changing amplitudes in  $2 + 1$  dimensional gravity*, Nuclear Phys. **B323** (1989), 113.
17. D. Johnson, *A geometric form of Casson's invariant and its connection to Reidemeister torsion*, unpublished lecture notes.
18. E. Witten,  *$2 + 1$  dimensional gravity as an exactly soluble system*, Nuclear Phys. **B311** (1988/9), 46.

SCHOOL OF NATURAL SCIENCES, INSTITUTE FOR ADVANCED STUDY, PRINCETON, NEW JERSEY  
08540