BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 3, Number 1, July 1980

HARMONIC ANALYSIS AS THE EXPLOITATION OF SYMMETRY–A HISTORICAL SURVEY

BY GEORGE W. MACKEY

CONTENTS

Reprinted from *Rice University Studies* (Volume 64, Numbers 2 and 3, Spring–Summer 1978, pages 73 to 228), with the permission of the publisher.

1980 *Mathematics Subject Classification.* Primary 01, 10, 12, 20, 22, 26, 28, 30, 35, 40, 42, 43, 45, 46, 47, 60, 62, 70, 76, 78, 80, 81, 82.

PREFACE

This paper is an expansion (by a factor of twelve) of two talks that I gave in the spring of 1977 at the Rice University Conference on the history of analysis. I am not a historian in the usual professional sense of the word and I did not pretend to be reporting on the results of a careful scholarly investigation. My talks consisted rather of an informal account of the knowledge and impressions I have gained over the years as I have tried to satisfy my curiosity about the origins and interrelationships of the parts of mathematics and science which most interest me. Moreover, in contradistinction to most (if not all) professional historians of science and mathematics, I was much more interested in getting an overall approximate idea of how our present understanding unfolded than in studying the fine structure of particular discoveries. I totally ignored such questions as false starts, the thought processes of individual scientists, and the intellectual climate of the times. The question constantly in my mind was this: How much of what we understand now had they grasped by then?

When I began in early August 1977 to concentrate on the job of writing up my talks for publication, I found that I could not say what I wanted to say without making a number of assertions about matters of fact that were not always easy to verify but whose exactitude was at most marginally relevant to the story I was trying to tell. I made some attempt to make only correct assertions (sometimes by being deliberately vague) but my time and patience were limited and I am sure I did not always succeed.

The length of the finished product is not all due to striving for local historical accuracy (this would require at least fifty books), but rather to my desire to make clear the relationship of a large part of mathematics to my

main theme. To this end I composed a large number of brief introductory expositions, which I hope the average mathematician will find intelligible. It is these expositions and their connectedness in time and intellectual content which is the real point of the paper.[1]

1. INTRODUCTION

In this article I shall sketch the history, applications, and ramifications of a certain method. For want of a better word I shall call it the method of harmonic analysis, although I am aware that many people use these words to denote a different class of generalizations of the classical harmonic analysis of Fourier. In crude terms the method may be described as follows: Let S be a "space" or "set" and let G be a group of one-to-one transformations of S onto itself. Let $[s]x$ denote the transform of s in S by x in G. Ordinarily S will have further structure which will be preserved by the transformations of G so that the transformations $s \to [s]x$ are symmetries of S. It will be convenient to allow members of G other than the identity e to define the identity map so that some quotient group G/N is the actual transformation group. Now let $\mathfrak{F}$ be some vector space of complex valued functions on S which is G invariant in the sense that $s \to f([s]x)$, the translate of f by x, is in $\mathfrak{F}$ whenever f is in $\mathfrak{F}$. Then for each x in G, the mapping $f \to g$ where $g(s) = f([s]x)$ is a linear transformation V_x of $\mathfrak{F}$ onto $\mathfrak{F}$ and $V_{xy} = V_xV_y$ for all x and y in G. The mapping $x \to V_x$ is thus an example of what is called a (linear) representation of the group G. More generally, a (linear) representation of a group G is by definition any homomorphism $x \to W_x$ of G into the group of all bijective linear transformations of some vector space $\mathfrak{H}(W)$. The method I propose to discuss in this article consists (in its simplest form) in attempting to find subspaces M_λ of the space $\mathfrak{F}$ such that

(1) $V_x(M_\lambda) = M_\lambda$ for all x and λ.
(2) Every element f in $\mathfrak{F}$ is uniquely a finite or infinite sum $f = \Sigma f_\lambda$ where each $f_\lambda \in M_\lambda$.
(3) The subspaces M_λ are either not susceptible of further decomposition or are somehow much simpler in structure than $\mathfrak{F}$. Of course, one must have a topology in $\mathfrak{F}$ in order to make sense of infinite sums. More generally one also considers "continuous direct sums" or direct integrals and vector-valued as well as complex-valued functions. Of course, each M_λ is the space of a new representation V^λ, which is a so-called subrepresentation, and one speaks of the direct sum or direct integral decomposition of V. It turns out that the decomposition of functions in $\mathfrak{F}$ into sums and integrals of functions associated with the components of V is a decomposition that greatly simplifies many problems.

2. THE CHARACTERS OF FINITE COMMUTATIVE GROUPS AND THE CONNECTION WITH FOURIER ANALYSIS

Let W be a (linear) representation of the group G. If $\mathfrak{H}(W)$ is one-dimensional then each W_x is some complex number $\chi(x)$ times the identity, and $x \to \chi(x)I$ is a representation of G if and only if $\chi(xy) = \chi(x)\chi(y)$ for all x and y in G. When G is a finite commutative group, such functions χ are called characters. It is easy to see that the representations $x \to \chi(x)$ are the only representations of G that are *irreducible* in the sense that no proper subrepresentations exist. Let $\mathfrak{F}(G)$ be the vector space of all complex-valued functions on G and let $V_x f(y) = f(yx)$. Then V is a representation of G such that $\mathfrak{H}(V) = \mathfrak{F}(G)$ and the one-dimensional subspace generated by each character is clearly an invariant subspace. Indeed every one-dimensional invariant subspace is of this form, and one shows easily that V is a direct sum of one-dimensional representations each associated with a distinct character χ. Thus every member f of $\mathfrak{F}$ is uniquely of the form

$$\sum_{\chi \in \hat{G}} c_\chi \chi(x)$$

where $\hat{G}$ denotes the set of all characters on G. Evidently the product of two characters is again such, and the set $\hat{G}$ of all characters is itself a finite commutative group. Since $\chi(x^n) = \chi(x)^n = 1$ when n is the order of x, it follows that $|\chi(x)| \equiv 1$ for all χ and all x, so $\chi^{-1} = \overline{\chi}$ for all χ. Consider $\sum_{x \in G} \chi(x)$. This sum is clearly equal to $\sum_{x \in G} \chi(xy) = \chi(y) \sum_{x \in G} \chi(x)$. Thus whenever $\chi(y) \neq 1$, we have $\sum_{x \in G} \chi(x) = 0$. It follows that $\sum_{x \in G} \chi_1(x)\overline{\chi}_2(x) = 0$ whenever $\chi_1 \neq \chi_2$, so that the characters of G are *orthogonal* with respect to the inner product $f \cdot g = \sum_{x \in G} f(x)\,\overline{g(x)}$.

Now consider the expansion formula

$$f(x) = \sum_{\chi \in \hat{G}} c_\chi \chi(x) \tag{2.1}$$

Multiplying each side by $\overline{\chi'(x)}$ summing over G and using the orthogonality of the characters leads at once to a formula for $c_{\chi'}$.

$$c_{\chi'} = \frac{1}{o(G)} \sum_{x \in G} f(x)\overline{\chi}(x) \tag{2.2}$$

where $o(G)$ is the number of elements in G. The formulae (2.1) and (2.2) are strikingly similar to the formulae (2.3) and (2.4) below, which occur in the theory of Fourier series when sines and cosines are replaced by complex exponentials.

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{inx} \tag{2.3}$$

$$c_n = \frac{1}{2\pi} \int_0^{2\pi} f(x)e^{-inx}\,dx \tag{2.4}$$

and the analogy becomes closer when one notices 1) that functions on the real line with period 2π may be identified with functions on the compact topological group T which the additive real line becomes when the subgroup of all integer multiples of 2π is factored out; 2) that the functions $x \to e^{inx}$ are precisely the continuous characters on T; and 3) that $c_\chi = \frac{1}{o(G)} \sum_{x \in G} f(x)\overline{\chi}(x)$ may be written as $c_\chi = \frac{1}{\mu(G)} \int f(x)\overline{\chi}(x)d\mu(x)$, where μ is the measure on G such that the measure of every subset is the number of elements it contains.

What we have called the method of harmonic analysis thus includes classical harmonic analysis (in the sense of expansion in Fourier series) as a very special case.

The formula (2.2) may be looked upon as defining a linear transformation of the vector space of all complex-valued functions on G onto the vector space of all complex-valued functions on $\hat{G}$. Similarly, formula (2.1) may be looked upon as defining the inverse of this transformation. A key property of the resulting one-to-one correspondence between functions on G and functions on $\hat{G}$ is that the operation of translation of functions on G is carried over into the operation of multiplication by a fixed function for the functions on $\hat{G}$. Similarly, the formulae (2.3) and (2.4) define a one-to-one linear correspondence between certain (not all) periodic functions on the line and certain functions on the additive group of all integers; and this correspondence also carries translation into multiplication. More significantly and for the same formal reasons it converts differentiation into multiplication by a function and so converts differential equations into algebraic equations.

A considerable part of the utility of classical harmonic analysis may be traced to this simple fact.

Looking at the functions e^{inx} as group characters and Fourier analysis as a special case of the decomposition of group representations are of course twentieth-century viewpoints. Indeed the very concept of a group representation was not formulated until the closing years of the nineteenth century. On the other hand, characters of finite commutative groups go back to the work of Gauss, and the analogue of Fourier expansions for functions on such groups has played a key role in number theory since the beginning of the nineteenth century. Moreover, the use of functional transforms involving characters to convert translation into multiplication may be traced back to early eighteenth-century work in probability. Thus the method of har-

monic analysis has at least three independent origins; in probability, in number theory, and in mathematical physics. In the immediately following sections I shall sketch the history of these subjects with emphasis on the rise of the method of harmonic analysis.

3. PROBABILITY THEORY BEFORE THE TWENTIETH CENTURY

The basic facts of elementary probability theory were clarified, systematized, and to some extent discovered in a celebrated correspondence between Fermat (1601-1665) and Pascal (1623-1662), which began in 1654. The first book on the subject appeared in 1657 and was a short pamphlet by Huygens (1629-1695) entitled "De ratiocinio in ludo aleve." It applied the principles discovered by Fermat and Pascal to various gambling problems. The first major treatise on the subject was J. Bernoulli's (1654-1705) *Ars Conjectandi,* published posthumously in 1713. In addition to a commentary on Huygens's pamphlet (which was reprinted in full), it contained a statement and proof of the weak law of large numbers. This work was soon followed by another: the publication in 1718 of *Doctrine of Chances* by A. de Moivre (1667-1754). De Moivre's book is noteworthy for three things: In the form of an approximation to a formula of Bernoulli it contains an early intimation of the central limit theorem; it introduced the technique of solving problems in probability by reducing them to difference equations; and (at least implicitly) it introduced the technique of using "generating functions" to solve difference equations. No other major book on probability appeared until 1812, when Laplace (1749-1827) published his great treatise *Théorie Analytique des probabilités.* Based on a series of nine memoirs published between 1771 and 1786, Laplace's treatise developed and synthesized the work of his predecessors and put probability into a form which was to be more or less unchanged until the twentieth century. Two important new ideas were introduced between the appearance of the book by de Moivre and the first of Laplace's memoirs. In 1756 Thomas Simpson (1710-1761) began the application of probability theory to the study of errors of measurement, and in 1763 Bayes introduced the concept of inverse probability or probability of causes. Laplace's treatise included a development of both these ideas as well as a very extensive development of the ideas of de Moivre concerning difference equations, generating functions, and the central limit theorem. Indeed among Laplace's chief original contributions are 1) a formulation and heuristic proof of the central limit theorem; 2) an extension of the theory of difference equations to equations in several variables; and 3) a systematic use of generating functions in dealing with difference equations in one and several variables.

The nineteenth century was far richer in new applications of probability theory than in the development of new methods and principles. It was the

century in which Quetelet (1796-1874), Galton (1822-1911), and Pearson (1857-1926) began the probabilistic study of human variation, in which Mendel (1822-1884) applied probability to genetics, and in which Maxwell (1831-1879), Boltzmann (1844-1906), and Gibbs (1839-1903) developed a statistical theory of heat and thermodynamics. As far as purer aspects are concerned, there were two chief contributors after the early and independent work of Gauss (1777-1855) and Legendre (1752-1833) on the theory of errors and the method of least squares. These were Poisson (1781-1840) and Tchebycheff (1821-1894). Poisson recognized the importance of the distribution which bears his name, generalized Bernoulli's work to the case of probabilities that vary from trial to trial, and published a book on probability in 1837. Tchebycheff was the first to think systematically in terms of "random variables" and their "expectations" and "moments." Using these concepts he discovered a simple inequality in 1867 that led to a remarkably simple proof of Bernoulli's law of large numbers. Moreover, he inaugurated a program for using the moment concept to give a rigorous proof of the central limit theorem. This program was completed by his student Markov (1856-1922), who became one of the leading probabilists of the early twentieth century. Shortly thereafter Liapunov (1858-1918), another pupil of Tchebycheff, found a simpler and better proof of the central limit theorem using "characteristic functions" instead of moment sequences.

4. THE METHOD OF GENERATING FUNCTIONS IN PROBABILITY THEORY

During the first third of the twentieth century, it became clear that both the central limit theorem and the law of large numbers are essentially corollaries of theorems in harmonic analysis. In particular, the so-called "characteristic function" of a probability distribution is just its Fourier transform, and Liapunov's proof of the central limit theorem essentially exhibits the theorem as a corollary of the (nonobvious) fact that the Fourier transform is a homeomorphism between appropriately topologized function spaces. Moreover (as I shall explain in some detail in a later section), the pioneering work of Norbert Wiener in the 1920s led over the next few decades to a rather profound development and intermingling of concepts from probability theory with those of harmonic analysis on the line. Thus it is interesting that harmonic analysis as a method seems to have first been used to deal with problems in probability theory.

Let r be a positive integer and for each $n = 1,2,3,\cdots$, let p_n^r denote the probability that at least one "run" of r heads will occur during n coin tosses. We assume that the tosses are independent and that the probability of getting heads on any given toss is q where $0 < q < 1$. The problem is to

compute p_n^r for each n. This problem was first posed and solved by de Moivre. His method was to consider p_n^r as a function of n and show that this function satisfies a certain "difference equation." Suppressing the r and writing $p_n^r = P(n)$, application of elementary principles of probability theory leads to the conclusion that

$$P(n+1) = P(n) + (1 - P(n-r))\,q^r(1-q)$$

for all $n \geq r$. Setting $\tilde{P}(n) = 1 - P(n)$ thus yields the homogeneous difference equation

$$\tilde{P}(n+1) = \tilde{P}(n) - q^r(1-q)\,\tilde{P}(n-r)\,.$$

Our unknown function $\tilde{P}$ is clearly the unique solution of this equation which satisfies the "initial conditions" $\tilde{P}(1) = \tilde{P}(2) = \cdots = \tilde{P}(r-1) = 1$, $\tilde{P}(r) = 1 - q^r$.

Many problems in probability theory may be thus reduced to the solution of difference equations. De Moivre's book *Doctrine of Chances* is rich in examples and in methods for solving such equations. One important method—the method of generating functions—occurs at least implicitly in de Moivre's work, but was first developed and used systematically by Laplace in a paper published in 1782. It was Laplace who coined the term "generating function." The first chapter of his treatise on probability is largely a reprint of the contents of the 1782 paper.

The idea of the method is very simple. Let $f(n)$ be an unknown function of the non-negative integer variable n, and suppose that $f(n+2) = af(n+1) + bf(n)$ for all $n \geq 0$ where a and b are known constants. Laplace calls the power series

$$\tilde{f}(t) = f(0) + tf(1) + t^2f(2) + \cdots$$

the generating function of f and shows that determining $\tilde{f}$ can be reduced to solving an algebraic equation. The point is that if $g(n) = f(n+1)$ then $\tilde{g}(t) = f(1) + tf(2) + t^2f(3) + \cdots = \dfrac{\tilde{f}(t) - f(0)}{t}$, and if $h(n) = f(n+2)$ then $\tilde{h}(t) = f(2) + tf(3) + \cdots = \dfrac{\tilde{f}(t) - f(0) - tf(1)}{t^2}$. Thus f satisfies the given difference equation if and only if $\dfrac{\tilde{f}(t) - f(0) - tf(1)}{t^2} = a\left(\dfrac{\tilde{f}(t) - f(0)}{t}\right) + b\tilde{f}(t)$, and solving a simple algebraic equation gives us a formula for $\tilde{f}$ as a rational function of t whose coefficients depend on $f(0)$ and $f(1)$. To find f for other values of n, we need only expand this rational function in a power series.

The method can obviously be applied to any kth order difference equation with one variable and constant coefficients, but, as Laplace em-

phasized, it can also be used for the difference equations in several variables to which one is led by more complicated problems in probability.

The relationship to what we have called the method of harmonic analysis is almost evident. Since $t^{n+m} = t^n t^m$, the functions $n \to t^n$ are characters on the group Z of all integers, and the generating function $\tilde{f}(t) = f(0) + tf(1) + \cdots$ bears just the same relationship to the function f (extended to be defined on all of G by letting it be zero for negative n) that f and c bear to one another when G is a finite commutative group. The difference equation becomes an algebraic equation because translation becomes multiplication by a function.

5. NUMBER THEORY BEFORE 1801

The theory of numbers is one of the very oldest branches of mathematics, going back at least to work of Euclid around 300 B.C. Euclid already knew a proof that there are an infinite number of primes, and by A.D. 300 Diophantus had written a treatise on methods for finding the integral solutions of indeterminate equations. The six surviving "books" of this treatise were translated into Latin in 1621, and modern number theory is considered to have begun when the same Fermat who helped found probability theory read this translation and began to study the subject for himself. He announced his results in letters to others and made marginal notes in his copy of the works of Diophantus, all too often neglecting to explain how he had arrived at them or how they might be proved. A century later Leonard Euler (1707-1783) became interested in the challenge presented by the unproved assertions of Fermat and produced the first published proofs of a number of them—often with great difficulty. Consider for example the problem of finding the integer solutions of $x^2 + y^2 = n$ where n is a given positive integer. It is not difficult to reduce the problem to the case in which n is a prime p, and it is also quite easy to show that there can be no solution when p is of the form $4l + 3$. Every other prime is either 2 (where there clearly is a solution) or of the form $4l + 1$. Fermat asserted that every prime of the form $4l + 1$ is a sum of two squares, and Euler managed to demonstrate this in 1754, but reportedly only after many years of effort. Later he was able to prove a number of analogous assertions of Fermat, such as the solvability of $x^2 + 3y^2 = p$ where p is a prime of the form $6n + 1$ and of $x^2 + 2y^2 = p$ when p is a prime of the form $8n + 1$. Rather earlier Euler made a very original and important contribution to number theory by noticing that an infinite series of the form $\sum_{n=1}^{\infty} \frac{1}{n^s}$ (with s a real number greater than one) can be written as an infinite product $\prod_p \left(\frac{1}{1 - \frac{1}{p^s}}\right)$ where p runs over all the primes. Using this and the fact that $\sum \frac{1}{n}$ diverges, he was able to show that $\sum \frac{1}{p}$ di-

verges, thus strengthening and re-proving Euclid's discovery that there is an infinite number of primes. Similar arguments enabled him to show that there is an infinite number of primes of the form $4n + 1$ as well as of the form $4n + 3$.

The next major advances were made by J.L. Lagrange (1736-1813), the first mathematician after Euler to reach comparable stature. If one starts to study Diophantine equations systematically, one finds that the theory of linear equations can be completely worked out rather easily. The simplest case presenting a genuine challenge is that of quadratic equations in two unknowns—the most general being the equation

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

where $A, B, C, D, E,$ and F are given integers. The linear terms can be eliminated by simple transformations and one is confronted with equations of the form $Ax^2 + Bxy + Cy^2 = -F$, that is with the problem of deciding whether and in how many ways a given integer $-F$ may be represented by the "binary quadratic form" $Ax^2 + Bxy + Cy^2$. Special cases of the problem were studied and solved or partially solved by Fermat and Euler as described above. Lagrange's contribution was to attack the general case and to discover a number of important theorems about it. His main results appear in two long memoirs in publications dated 1773 and 1775 respectively.

Of key importance in Lagrange's work is a natural notion of equivalence between quadratic forms. If $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ is a matrix of integers such that $ad - bc = \pm 1$, then substituting $x = ax' + by', y = cx' + dy'$ converts $Ax^2 + Bxy + Cy^2$ into another form $A'x'^2 + B'x'y' + C'y'^2$, which evidently represents precisely the same integers that $Ax^2 + Bxy + Cy^2$ does. Calling two forms equivalent when they may be obtained from one another in this way, one sees that in the representation problem one need only consider a single form in each class. The integer $D = B^2 - 4AC$ is called the discriminant of the form and is easily verified to depend only on the class of the form. It is possible for inequivalent forms to have the same discriminant, however, and two of Lagrange's more important discoveries may be formulated as follows: 1) For a given value of D there can be only a finite number of distinct classes of forms having D as a discriminant; 2) when the number of classes of forms with discriminant D is greater than one they must all be considered together if one wants to reduce the problem of solving $Ax^2 + Bxy + Cy^2 = n$ to the case in which n is a prime. For example if $Ax^2 + Bxy + Cy^2 = p_1p_2$ where p_1 and p_2 are distinct primes and x and y are integers, one cannot conclude in general that either of the equations $Ax^2 + Bxy + Cy^2 = p_1$ or $Ax^2 + Bxy + Cy^2 = p_2$ has a solution in integers. One can only conclude that there exist forms $A'x^2 + B'xy + C'y^2$ and $A''x^2 + B''xy + C''y^2$ where $(B'')^2 - 4A''C'' = (B')^2 - 4A'C' = B^2 - 4AC$ such that

$$A'x^2 + B'xy + C'y^2 = p_1 \text{ and } A''x^2 + B''xy + C''y^2 = p_2$$

both have integer solutions. To get an elegant theory one has to be less ambitious and seek to evaluate the sum $\phi_{Q_1}(n) + \phi_{Q_2}(n) + \cdots + \phi_{Q_h}(n)$ where $\phi_Q(n)$ is the (suitably normalized) number of solutions of $Q(x,y) = n$ in integers and $Q_1, \cdots, Q_h$ constitutes a complete set of mutually inequivalent forms having the common discriminant D. (With suitable normalization even indefinite forms ($D > 0$) lead to equations with a finite number of solutions.) One can reduce to the case in which D is square free and then the key facts may be formulated as follows (Lagrange presented them differently):

(1) $\phi_D = \phi_{Q_1} + \phi_{Q_2} + \cdots + \phi_{Q_h}$ is "multiplicative" in the sense that $\phi_D(n_1n_2) = \phi_D(n_1)\,\phi_D(n_2)$ whenever n_1 and n_2 are relatively prime.

(2) If p is a prime and $k = 1,2\cdots$ then $\phi_D(p^k) = 1 + \chi_D(p) + \cdots + (\chi_D(p))^k$ where $\chi_D(p) = \phi_D(p) - 1$. It follows from (1) and (2) that $\phi_D(n)$ can be computed from the factorization of n when $\phi_D(p)$ is known for all primes p.

(3) $\phi_D(p)$ is 0,2, or 1, and which it is depends only on the solvability of the equations $Q_j(x,y) = 0 \bmod p$ and hence the value of $D \bmod p$.

To get an overview of the values of $\phi_D(p)$ for fixed D, one needs to apply the celebrated quadratic reciprocity law, which together with its two supplements allows one to compute the behavior of $D \bmod p$ from that of $p \bmod D$. The probable truth of this law was known to both Euler and Lagrange, but it was Legendre (1752-1833) who first clearly stated it in a paper published in 1785. He also offered a proof, but it relied on a lemma that was first properly proved by Dirichlet over half a century later. Let p be an odd prime and for each nonzero integer n let us define the "Legendre symbol $\left(\frac{n}{p}\right)$ to be 1 or -1 according as the nonzero integer n is or is not a square mod p. It is trivial that $\left(\frac{nm}{p}\right) = \left(\frac{n}{p}\right)\left(\frac{m}{p}\right)$ so that it suffices to know $\left(\frac{n}{p}\right)$ when n is 2, -1, or an odd prime to know it for all n. The quadratic reciprocity law asserts that $\left(\frac{q}{p}\right)\left(\frac{p}{q}\right) = (-1)^{(p-1)(q-1)/4}$ whenever q is an odd prime. Its two supplements state that $\left(\frac{-1}{p}\right) = (-1)^{(p-1)/2}$ and $\left(\frac{2}{p}\right) = (-1)^{(p^2-1)/8}$ and are easier to prove.

Legendre published an important book on number theory in 1798. He presented Lagrange's theory with various improvements, including the quadratic reciprocity law, and also made a start on a theory of ternary quadratic forms in a celebrated treatise published in 1798.

6. THE WORK OF GAUSS AND DIRICHLET AND THE INTRODUCTION OF CHARACTERS AND HARMONIC ANALYSIS INTO NUMBER THEORY

The first person to carry on the work begun by Lagrange and Legendre on the general theory of binary quadratic forms was C.F. Gauss (1777-1855). By his own account Gauss did not become aware of the work of his predecessors until he entered the university at Göttingen in the fall of 1795. Earlier the same year he accidentally discovered that if p is an odd prime, then -1 is a square mod p if and only if p is of the form $4n + 1$. This excited him tremendously and, determined to get to the bottom of such phenomena, he had found and proved the quadratic reciprocity law by the end of March 1795. One thing led to another, and before arriving in Göttingen and beginning to study the works of Euler, Lagrange, and Legendre, Gauss had rediscovered many of their results. All this is explained in the introduction to his classic treatise *Disquisitiones Arithmeticae* published in 1801, which set the course for the future development of number theory. He claims that most of the material in the first four of the seven sections of his treatise was known to him before he arrived in Göttingen and that a large part of the book had been set up in type before Legendre's book of 1798 appeared. While he acknowledges that he was inspired to study quadratic forms by the work of Lagrange and Legendre, he reworked the whole subject in his own way, giving new proofs and introducing important new ideas and concepts. The seventh section contains Gauss's famous proof that for any prime p of the form $2^n + 1$ (e.g., 17) one can give a ruler and compass construction of a regular polygon with p sides. He is reported to have definitely made up his mind to be a mathematician when he found this beautiful result in the spring of 1796. At this point the celebrated Galois theory of equations was thirty-five years in the future, but Gauss's proof is imbedded in what amounts to a complete development of that theory for the equation $x^p - 1 = 0$. ($x^p - 1$ factors into $x - 1$ and an irreducible polynomial of degree $p - 1$. If $p - 1 = 2^k$, it follows from Galois theory that the equation can be solved by rational operations and the taking of square roots.)

In his work on the theory of binary quadratic forms, Gauss confined himself to the case in which the middle coefficient is even, writing $Ax^2 + 2Bxy + Cy^2$ and defining the discriminant to be $B^2 - AC$ instead of $B^2 - 4AC$. Also he pointed out that simplifications ensue if one makes a distinction between "proper" and "improper" equivalence of forms; two forms being said to be properly equivalent only when the transformation matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ has determinant 1 rather than -1. By far his most important contribution, however, was his observation that there is a natural (but not obvious) composition law for proper equivalence classes of forms having a fixed discriminant and that this composition law converts the finite set of classes with square free discriminant D into a finite commutative group. Of

course, group theory did not then exist (cf. section 11) and Gauss did not use this terminology, but this is in effect what he proved about his composition law. Legendre had already divided the classes with a given square free discriminant into subsets that Gauss called genera and related in an interesting way to his composition law.

In modern terminology, the genera are just the cosets of the subgroup of squares in the group of all classes with discriminant D. As defined by Gauss, Q_1 and Q_2 are in the same genus if they have the same "total character" where the "total character" is a system of ones and minus ones that is canonically associated to each form. In effect Gauss defined a finite set $\chi_1, \ldots, \chi_l$ of functions from classes of forms to ± 1 and called them characters. He then put two forms Q_1 and Q_2 in the same genus when $\chi_j(Q_1) = \chi_j(Q_2)$ for all j. It turns out that Gauss's characters are characters in the modern sense for the finite commutative group in question. The fact that the genera are cosets of some subgroup which contains all squares follows at once from simple group theoretical considerations. To say that it contains only squares is equivalent to saying that every character of order two is a product of Gauss characters. Gauss proved this by a difficult argument using ternary forms. As he expressed it, every form in the principal genus can be obtained by composing some form with itself. This is known as Gauss's theorem on duplication. That all genera have the same number of elements and that the genera inherit a composition law are evident from the group theoretical interpretation.

To understand the significance of the division of classes into genera from the point of view of solving $Q(x,y) = n$ in integers and at the same time to appreciate the usefulness of finite Fourier analysis in number theory, it will be convenient to anticipate the future and introduce the characters of the group of classes that take on values other than ± 1. Let $Q_1, Q_2, \cdots, Q_h$ be a complete set of inequivalent binary forms of square free discriminant D and let $\phi_j(n)$ denote the (suitably normalized) number of representations of n by Q_j. We have already remarked in section 5 that if $\phi = \phi_1 + \phi_2 + \cdots + \phi_h$, then ϕ is multiplicative and there is a simple formula expressing $\phi(p^k)$ in terms of $\phi(p)$. More generally, one can show that for each character χ of the group C_D defined by composition of classes, the function $n \to \phi_\chi(n) = \Sigma \chi(Q_j)\, \phi_j(n)$ has these same properties. That is, $\phi_\chi(n_1\, n_2) = \phi_\chi(n_1)\, \phi_\chi(n_2)$ when n_1 and n_2 are relatively prime, and $\phi_\chi(p^k) = a(p)^k + b(p)a(p)^{k-1} + b(p)^2\, a(p)^{k-2} + \cdots + b(p)^k$ where $a(p) + b(p) = \phi_\chi(p)$ and $a(p)b(p) = \phi(p) - 1$. Thus if $\phi_j(p)$ is known for all primes p and all $j = 1,2, \cdots, h$, then $\phi_\chi(p)$ can be computed for all χ and p and hence $\phi_\chi(n)$ for all χ and n. But by finite Fourier analysis $\phi_j(n) = h\, \Sigma\, \overline{\chi(Q_j)}\, \phi_\chi(n)$. Thus one has an explicit formula for each $\phi_j(n)$ as a finite linear combination of the multiplicative functions $\phi_\chi(n)$, each of which can be computed once one knows the ϕ_j at the primes.

Unfortunately it seems to be quite difficult to find an analogue for general χ of the fact that $p \to \phi(p)$ is the restriction to the primes of a periodic function on the integers and correspondingly difficult to get an explicit expression for $\phi_j(n)$ in the general case. But this difficulty disappears when χ is of order two. Thus in the special case in which all characters are of order two (i.e., when there is just one class in each genus), one can get a simple explicit formula for each ϕ_j. In general one needs to sum the ϕ_j only over a genus rather than a class in order to obtain such formulae.

The word *character* as used today stems directly from Gauss's use of the term in his theory of binary quadratic forms. As a homomorphism of an *abstract* finite commutative group into the group of roots of unity it was first defined in 1882 by Weber, who refers to the version of Dedekind's ideal theory for algebraic number fields published in 1879. In that reference Dedekind makes the same definition for the special case of the ideal class group of an algebraic number field. As Dedekind himself had pointed out earlier, the ideal class group is a generalization of Gauss's group of equivalence classes of binary quadratic forms. Dedekind's mode of expression is such as to make it clear that he regards his definition as a generalization of that of Gauss.

Characters and Fourier analysis on finite commutative groups occur implicitly in other parts of Gauss's work. For example, let R_N denote the ring of integers mod N for $N = 2,3,4 \ldots$ and let $\phi_N(k)$ be the number of members l of R_N such that $l^2 = k$ in R_N. The characters on the additive group of R_N are the functions $k \to \chi_q(k) = e^{2\pi ikq/N}$ where $q = 0,1, \cdots, N-1$. Thus the Fourier transform $\hat{\phi}_N$ of ϕ_N is the function $\chi_q \to \sum_{k=0}^{N-1} \chi_q(k)\, \phi_N(k) = \sum_{s=0}^{N-1} e^{2\pi iqs^2/N}$. When N is an odd prime or a product of two such and q does not divide N, it is easy to see that $\hat{\phi}_N(\chi_1) = \left(\frac{q}{N}\right) \hat{\phi}_N(\chi_1)$ where $\left(\frac{q}{N}\right)$ is the Legendre symbol, so that it suffices to know $\hat{\phi}_N(\chi_1)$ to know $\hat{\phi}_N$. It is also easy to see that $\hat{\phi}_N(\chi_1)^2 = \pm N$. But Gauss found it quite difficult to determine the unknown signs and show that $\hat{\phi}_N(\chi_1) = \sqrt{N}$ or $i\sqrt{N}$ depending on whether N is of the form $4n + 1$ or $4n + 3$. Indeed, Gauss's fourth proof of the quadratic reciprocity law consists in showing it to be an easy corollary of the precise determination of the argument of the "Gauss sum" $\sum_{s=0}^{N-1} e^{2\pi is^2/N} = \hat{\phi}_N(\chi_1)$. According to Davenport ([4], p. 14), the most satisfactory evaluation of the Gauss sums is one given by Dirichlet in 1835 and based on a straightforward application of the so-called Poisson summation formula in the theory of Fourier transforms on the line. Thus the quadratic reciprocity law itself may be regarded as resulting from an application of harmonic analysis to numbeı theory.

Dirichlet (1805-1859) was an assiduous student of Gauss's *Disquisitiones* and apparently the first to understand some of its more obscure parts. He simplified and amplified many arguments, but above all managed to complete and extend Gauss's work in two important respects. First of all he gave the first valid proof of the fact that whenever a and m are relatively prime positive integers, then the arithmetic progression $a, a + m, a + 2m \cdots$ contains infinitely many primes. Second, he found an explicit formula for computing the number $h(D)$ of equivalence classes of binary quadratic forms of given discriminant D. The two results are closely related and depend on using limits and other concepts from analysis—in particular infinite series of the form $\sum_{n=1}^{\infty} \frac{a_n}{n^s}$ where s is > 1. Such series are now known as Dirichlet series and Dirichlet is often credited with being the founder of analytic number theory. His results were published in several installments between 1837 and 1840.

Dirichlet's proof of the existence of infinitely many primes in an arithmetic progression was inspired by Euler's proof of the existence of infinitely many primes and is at the same time a beautiful example of an application of Fourier analysis on finite commutative groups. Recall that Euler based his proof on the factorization $\sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{p} \frac{1}{1 - \frac{1}{p^s}}$ for $s > 1$. It is natural to try to copy Euler by replacing $\sum_{n=1}^{\infty} \frac{1}{n^s}$ by $\sum_{n=1}^{\infty} \frac{\theta(n)}{n^s}$ where $\theta(n)$ is 1 when n is in the given progression and zero otherwise. But $\sum_{n=1}^{\infty} \frac{\theta(n)}{n^s}$ does not factor. On the other hand it is easy to see that $\sum_{n=1}^{\infty} \frac{\chi(n)}{n^s} = \prod_{p} \frac{1}{1 - \frac{\chi(p)}{p^s}}$ whenever χ is a complex valued function on the positive integers which is strongly multiplicative in the sense that $\chi(n_1 n_2) = \chi(n_1)\chi(n_2)$ and is also such that $|\chi(n)| \leq 1$. Consider then the ring R_m of integers mod m. The elements of R_m which have multiplicative inverses form a finite commutative group under multiplication. If χ is any character on this group, we may extend it to be defined on all of R_m by making it zero where it is not already defined, and then regard it as a periodic function on the integers. Such functions are strongly multiplicative and are called *Dirichlet characters* mod m. It follows at once from finite Fourier analysis (and the fact that $\theta(n)$ is zero when n fails to have an inverse in R_m) that θ is a unique linear combination of Dirichlet characters mod m and hence that $\sum_{n=1}^{\infty} \frac{\theta(n)}{n^s}$ is a linear combination of Dirichlet series $\sum_{n=1}^{\infty} \frac{\chi(n)}{n^s}$ which factor into so-called "Euler products"

$\prod_p \frac{1}{1 - \frac{\chi(p)}{p^s}}$. Exploiting this fact, Dirichlet was able to adapt Euler's argument.

This adaptation, however, was not entirely straightforward. It depended on being able to prove that $L(s,\chi) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s}$ tends to a finite nonzero value as $s \to 1$ whenever $\chi \not\equiv 1$. This can be done without great difficulty except when χ is a character of order 2. To take care of this case Dirichlet used a very ingenious argument. Let $Q_1, \cdots, Q_h$ be a complete set of inequivalent binary quadratic forms with a fixed square free discriminant D, and for each $n = 1, 2, \cdots$ let $\phi_j(n)$ denote the (suitably normalized) number of representations of n by Q_j. Then as indicated above, the function $n \to \phi(n) = \phi_1(n) + \cdots + \phi_h(n)$ is multiplicative and $\phi(p^k) = 1 + \chi(p) + \chi(p)^2 + \cdots + \chi(p)^k$ where $\chi(p) = \phi(p) - 1$. Consider now the Dirichlet series $\sum_{n=1}^{\infty} \frac{\phi(n)}{n^s}$. The multiplicativity of ϕ shows that it factors as $\prod_p \left(\sum_{k=0}^{\infty} \frac{\phi(p^k)}{p^{ks}} \right)$ and the formula for $\phi(p^k)$ lets one replace this by

$$\prod_p \frac{1}{\left(1 - \frac{1}{p^s}\right)\left(1 - \frac{\chi(p)}{p^s}\right)}.$$

Hence

$$\sum_{n=1}^{\infty} \frac{\phi(n)}{n^s} = \left(\sum_{n=1}^{\infty} \frac{1}{n^s} \right) \left(\sum_{n=1}^{\infty} \frac{\phi_\chi(n)}{n^s} \right)$$

where $n \to \phi_\chi(n)$ is the unique strongly multiplicative function that coincides on the primes with $\chi(p) = \phi(p) - 1$. But the quadratic reciprocity law tells us that each ϕ_χ is a Dirichlet character mod m for some m, and it is easy to verify that every Dirichlet character of order 2 occurs. In other words, the function $L(s,\chi)$ whose behavior at $s = 1$ is to be investigated may be written as a quotient $\sum_{n=1}^{\infty} \frac{\phi(n)}{n^s} \Big/ \sum_{n=1}^{\infty} \frac{1}{n^s}$ where ϕ is as indicated above for some square free discriminant D. The behavior of $\sum_{n=1}^{\infty} \frac{a_n}{n^s}$ at $s = 1$ is determined by the asymptotic behavior of $a_1 + a_2 + \cdots + a_n$ as $n \to \infty$. Using simple geometric arguments, Dirichlet showed that

$$\frac{\phi_j(1) + \phi_j(2) + \cdots + \phi_j(n)}{n}$$

has a limit τ as $n \to \infty$, which is the same for all j and can be computed when D is known. Thus $\frac{\phi(1) + \cdots + \phi(n)}{n}$ has $h\tau$ as a limit when $n \to \infty$. Putting all of this together, he was able to deduce that $\sum_{n=1}^{\infty} \frac{\phi(n)}{n^s} \Big/ \sum_{n=1}^{\infty} \frac{1}{n^s}$ has a

finite nonzero limit as $s \to 1$. His formula for h in terms of D was a byproduct of these constructions.

In connection with these applications of analysis to number theory, it is interesting to recall that Dirichlet was also the first to prove (1829) that a Fourier series of a function actually converges when the function satisfies suitable weak conditions. It should also be noted that passing from $n \to \phi(n)$ to the Dirichlet series $\Sigma \frac{\phi(n)}{n^s}$ is a form of harmonic analysis in itself, since the functions $n \to n^s$ are the restrictions to the integers of characters on the multiplicative group of all rationals. In particular the explicit formula for $\phi(n)$ that follows from its multiplicativity and the fact that $\phi(p^k) = 1 + \chi(p) + \cdots + \chi(p)^k$ is equivalent to the rather simple statement that $\Sigma \frac{\phi(n)}{n^s} = \prod_p \left(\frac{1}{1 - \frac{1}{p^s}}\right) \prod_p \left(\frac{1}{1 - \frac{\chi(p)}{p^s}}\right)$.

7. MATHEMATICAL PHYSICS BEFORE 1807

Mathematical physics in simple form goes back at least to Archimedes (287–212 B.C.), who formulated the laws governing the magnification of forces by levers and pulleys and the magnitudes of the forces that fluids exert on bodies immersed in them. However, except for the ideas of Copernicus (1473–1503) concerning the central position of the sun among the planets, little further progress seems to have been made until near the end of the sixteenth century when Tycho Brahe (1546–1601) made extremely accurate naked-eye observations of planetary motions, and Stevinus (1548–1626) and Galileo (1564–1642) began to study swinging pendulums, falling bodies, etc. The observations of Tycho Brahe led to Kepler's (1571–1630) empirical laws of planetary motion in 1609, and one may think of modern mathematical physics as being formally inaugurated in 1637 and 1638 with the respective publications of Descartes's (1596–1650) "Discourse on Method" and Galileo's "Two new Sciences." The first introduced analytic geometry (independently invented by Fermat) to the world, and the second clarified the foundations of mechanics. A good idea of the state of knowledge at the time can be had from the following words of Galileo: "Some superficial observations have been made, as, for instance, that the free motion of a heavy falling body is continuously accelerated but to just what extent this acceleration occurs has not yet been announced"; and "It has been observed that missiles and projectiles describe a curved path of some sort; however no one has pointed out the fact that this path is a parabola."

The work begun by Galileo was enormously advanced in the 1660s by the work of Isaac Newton (1642–1727), who showed that Kepler's laws of plan-

etary motion and Galileo's laws of falling bodies are both consequences of a set of simple laws concerning a) motion in general, and b) the magnitude of the force attracting any two masses in the universe toward one another. To deal with the variable accelerations predicted by these laws, Newton invented the differential and integral calculus. This investion was also made (apparently independently) by Leibniz (1646–1716), and Newton acknowledges getting the idea from a method of Fermat for finding tangents to curves. Newton's *Principia*, containing a systematic account of the inventions and discoveries just alluded to, was published in 1687, half a century after the books of Descartes and Galileo. Newton also concerned himself with the theory of light and published a book on the subject in 1704, based on the hypothesis that light consists of rapidly moving particles. The opposing view, that light is a form of wave motion, was defended and developed in a book published in 1690 by Newton's slightly older contemporary Huygens (1629–1695). Newton's view was shown to be untenable in the early nineteenth century, but until then it was the view accepted by a majority of scientists.

Working out the full consequences of Newton's laws for planetary motion presents enormous mathematical difficulties, and celestial mechanics has been a source of profound and challenging mathematical problems since the appearance of the *Principia*. It will probably continue to be so for the foreseeable future. However, there is also the problem of applying similar ideas to other kinds of motion—in particular the relative motion of the parts of elastic bodies and fluids. The main steps in laying the foundations for such a continuum mechanics were taken in the mid-eighteenth century by Daniel Bernoulli (1700-1782), Euler (1707-1783), and D'Alembert (1717-1783). Bernoulli is usually considered to be the founder of fluid mechanics. His book *Hydrodynamica* appeared in 1738, originating the term *hydrodynamics*. On the other hand it is D'Alembert who is considered to be the originator of the idea of reducing problems in continuum mechanics to the study of partial differential equations, and Euler who first wrote down the system of partial differential equations governing the flow of a nonviscous (but possibly compressible) fluid. D'Alembert's study of the partial differential equation governing the motion of a vibrating string $\frac{\partial^2 f}{\partial t^2} = \mu^2 \frac{\partial^2 f}{\partial x^2}$ was published in 1747 and Euler's equations for nonviscous fluid flow came out in 1755.

If one formulates Newton's laws for gravitating "particles" in a suitable manner, they have a straightforward generalization that encompasses continuum mechanics as well. Let $x_1, y_1, z_1, \cdots, x_n, y_n, z_n$ be the coordinates in some rectangular coordinate system of n "particles." Newton's laws then assert the existence of $n + 1$ positive constants $m_1, \cdots, m_n, G$ such that when the particles move under their mutual attractions the coordinates as

functions of the time satisfy the differential equations

$$m_j \frac{d^2x_j}{dt^2} = -\frac{\partial V}{\partial x_j}$$

$$m_j \frac{d^2y_j}{dt^2} = -\frac{\partial V}{\partial y_j}$$

$$m_j \frac{d^2z_j}{dt^2} = -\frac{\partial V}{\partial z_j}$$

where V is the function of $x_1, \cdots, z_n$ given by the formula

$$V(x_1, \cdots, z_n) = \sum_{\substack{i,j=1\cdots n \\ i \neq j}} \frac{G\, m_i m_j}{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}} .$$

It is clear that replacing $m_1, \cdots, m_n, G$ *by* $\lambda m_1, \cdots, \lambda m_n, G/\lambda$ does not change the allowed trajectories, and that on the other hand the trajectories uniquely determine the products $m_1G, m_2G, \cdots, m_nG$. Thus the ratios of the m_j are uniquely determined and are by definition the *relative masses* of the particles. By assigning an arbitrary number as the mass of an arbitrarily chosen particle, all other particles acquire a well-defined positive mass. One says that one has chosen a unit of mass, and once this is chosen, G has a uniquely determined value called the *gravitational constant*. An important and easy consequence of the above differential equations is that the function

$$\sum_{j=1}^{n} \frac{m_j}{2}\left[\left(\frac{dx_j}{dt}\right)^2 + \left(\frac{dy_j}{dt}\right)^2 + \left(\frac{dz_j}{dt}\right)^2\right] + V(x_1, y_1, z_1, \cdots, x_n, y_n, z_n)$$

remains constant throughout time and is accordingly what is called an integral of the motion. Its value is called the *energy* of the motion on that particular trajectory. The term

$$\frac{m_j}{2}\left[\left(\frac{dx_j}{dt}\right)^2 + \left(\frac{dy_j}{dt}\right)^2 + \left(\frac{dz_j}{dt}\right)^2\right] = \frac{m_j v_j^2}{2},$$ where v_j is the absolute value of the velocity of the jth particle is called the *kinetic energy* of that particle. Any increase or decrease in the total kinetic energy T is exactly balanced by a decrease or increase of the function V, which is accordingly called the *potential energy*. Notice finally that the differential equations of motion can be expressed in terms of the two functions T and V. Let $q_1, \cdots, q_{3n} = x_1,$

$y_1, z_1, \cdots, x_n, y_n, z_n, q_1, \cdots, \dot{q}_{3n} = \frac{dx_1}{dt}, \frac{dy_1}{dt}, \cdots, \frac{dz_n}{dt}$. Then T is a function of the $\dot{q}_j$ above and $\frac{\partial T}{\partial \dot{q}_j} = m_j \dot{q}_j$. Thus the equations of motion may be written in the form $\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_j}\right) = -\frac{\partial V}{\partial q_j}$.

I shall not give details here, but it is not difficult to write down an analogue of these equations when the system is a continuum so that a configuration is described by a scalar or vector-valued function on a portion of space instead of a $3n$-tuple of real numbers. There is no difficulty in defining the kinetic energy of a moving continuous distribution of matter. One simply integrates the local kinetic energy defined by the velocity distribution and mass density function. The equations of motion are thus determined as soon as the analogue of V is known. This is a numerical function defined on all possible configurations of the continuous matter in question, and it must be determined by suitable experiments in each case. Fortunately the possibilities are not as various as one might think. A fluid (liquid or gas), for example, is characterized for mechanical purposes by the fact that V depends only on the function ϱ that describes the (possibly variable) mass per unit volume and may be computed from ϱ by integrating $g(\varrho)$ over the space occupied by the fluid. Here g is a real function defined on the positive real axis and characteristic of the fluid in question. One is usually not given g directly but rather the function $\varrho \to -\varrho^2 \frac{dg}{d\varrho}$, whose value at any given ϱ is called the *pressure* at that density. Actually, g (or equivalently the pressure function) depends not only on the nature of the fluid but on its so-called "temperature" as well, and the whole of classical continuum mechanics is valid only insofar as temperature changes can be ignored. The beautifully subtle theory known as thermodynamics, which deals with the interplay between continuum mechanics and temperature changes, was not developed until around 1850. Indeed, the work of Joseph Black (1728-1799) clarifying the distinction between temperature and quantity of heat and introducing the concept of specific heat did not begin until 1764.

The theory of the possible potential energy functions for a solid (elasticity theory) is much more complicated than for a fluid, and even the linear approximation (valid for small displacements from equilibrium) was not adequately worked out until the 1820s. Eighteenth-century work on elasticity was by and large confined to doing special problems by ad hoc methods. On the other hand, the theory of a flexible string is like that of a one-dimensional fluid in that the potential energy is determined by a single real function relating "tension" to linear density. In terms of this function and the generalization of Newton's laws indicated above, it is not difficult to write down the equation of motion—a non-linear partial differential equation

analogous to Euler's equations for nonviscous fluids. (A nonviscous fluid is one in which one can ignore the "dissipation of energy" as heat.) The equation studied by D'Alembert in 1747, $\frac{\partial^2 f}{\partial t^2} = \mu^2 \frac{\partial^2 f}{\partial x^2}$, is the approximation that results when one assumes f to be small and the tension linearly related to the density. D'Alembert discovered the rather easy argument leading to the conclusion that every solution may be written uniquely in the form

$$f(x,t) \equiv \phi(x - \mu t) + \psi(x + \mu t)$$

where ϕ and ψ are differentiable functions of one variable but are otherwise arbitrary. A year later Euler pointed out an important implication of D'Alembert's result. Since $f(x, 0) \equiv \phi(x) + \psi(x)$ and $\frac{\partial f}{\partial t}(x, 0) = \mu\psi(x) - \mu\phi(x)$, ϕ and ψ are uniquely determined by the configuration and its rate of change at $t = 0$. Thus the whole trajectory of the string is determined once its configuration and the rate of change of that configuration are known at $t = 0$. Five years later in 1753 Bernoulli considered the case of the string with fixed end points and length ℓ and for the first time emphasized the significance of the linearity of the equation in permitting the construction of solutions by "superposition," i.e., by taking arbitrary linear combinations of solutions already at hand. He saw in particular that $\sum_{n=0}^{\infty} a_n \sin\frac{\pi x n}{\ell} \cos\frac{n\pi\mu}{\ell}(t - b_n)$ must be a solution for "all" choices of the real constants a_n and b_n and gave heuristic arguments indicating that every solution can be written in this form. Combined with D'Alembert's general solution, the validity of Bernoulli's argument would imply the possibility of expanding a more or less arbitrary function in the form $\sum_{n=0}^{\infty} a_n \sin\frac{\pi x n}{\ell}$. This seemed paradoxical to many mathematicians of the time; a controversy arose and Bernoulli's conclusion was rejected. As a result the systematic application of harmonic analysis to the solution of linear partial differential equations came over half a century later than it might have.

Of the linear partial differential equations arising in continuum mechanics, the first to be studied after D'Alembert's equation of the vibrating string were "Laplace's equation" $\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} = 0$, and the "wave equation" $\frac{\partial^2\varrho}{\partial t^2} = \mu^2\left(\frac{\partial^2\varrho}{\partial x^2} + \frac{\partial^2\varrho}{\partial y^2} + \frac{\partial^2\varrho}{\partial z^2}\right)$. Both arise in studying special cases and approximate solutions of Euler's non-linear equations for nonviscous fluid flow and were written down by Euler in 1752 and 1759 respectively. When a fluid moves in such a way that the three components u, v, w of its

velocity can be written in the form $\frac{\partial\psi}{\partial x}, \frac{\partial\psi}{\partial y}, \frac{\partial\psi}{\partial z}$ where ψ is a single scalar function, one says that the flow is irrotational and that ψ is the *velocity potential.* The concept and even the term occur in Bernoulli's *Hydrodynamica* of 1738. When the fluid is also incompressible so that $\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} \equiv 0$, one sees (as noted by Euler in 1752) that the velocity potential satisfies the equation $\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} = 0$ and that many problems in fluid flow can be reduced to finding suitable solutions. It was well into the nineteenth century, however, before its theory began to be well understood. The same is true of the wave equation, which is a three-dimensional analogue of the vibrating string equation studied by D'Alembert in 1747. Euler used it to describe the small density oscillations of a fluid, such as occur for example in the propagation of sound.

In the later part of the eighteenth century, other aspects of the physical world began to be subjected to non-trivial mathematical analysis. We have already mentioned Black's work on heat and temperature. Black was also a pioneer in introducing quantitative methods into chemistry, making careful measurements of the masses involved when calcium carbonate decomposes into calcium oxide and carbon dioxide and preparing the way for the fundamental work of Priestley (1733-1804) and Lavoisier (1743-1794) between 1775 and 1785 on the nature of combustion and the law of conservation of matter. Modern chemistry is considered to have been founded with the publication of Lavoisier's book *Traité élémentaire de chimie* in 1789. In the 1750s and 1760s Michel (1724-1793), Priestley, and others began to study electrical and magnetic phenomena more quantitatively. They did experiments and made deductions that made it seem quite likely that magnetic poles and electric charges attract and repel one another with a force that is like gravitational attraction in varying inversely with the square of the distance. Then between 1785 and 1789 Coulomb published a series of memoirs reporting his own very careful measurements and convincing the scientific world of the validity of what is now known as Coulomb's law.

While Laplace's equation $\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} = 0$ first appears in work of Euler in fluid mechanics, it was destined to be important in dealing with all attractive and repulsive forces varying according to an inverse square law. Lagrange in 1773 pointed out that the field of force F_x, F_y, F_z produced by the gravitational attraction of an arbitrary distribution of matter is like an irrotational velocity distribution in being realizable as the three partial derivatives of a single "potential function" V. Twelve years later in 1785 Laplace observed that it was also "solenoidal" in the sense that $\frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} +$

$\frac{\partial F_z}{\partial z} = 0$, so that the potential V satisfies the equation that now bears his name. He overlooked the modification (later introduced by Poisson) that has to be made at points occupied by matter. Laplace (and his contemporary Legendre) were interested in computing the gravitational attraction due to bodies of various shapes and sizes and found it convenient to think in terms of the potential function V and its properties. It was in connection with this work that they introduced the functions on the surface of a sphere known as surface harmonics and expanded solutions of Laplace's equation (in polar coordinates) as a series of powers of r multiplied by surface harmonics. This expansion theory, when properly developed, permits one to show that an arbitrary continuous function on the surface of a sphere can be extended uniquely to satisfy Laplace's equation in the interior. Extending this theorem to closed surfaces of more general shape was to be a major theme of nineteenth-century mathematics and to lead to new developments of far-reaching importance. As I shall explain later, the use of expansions in surface harmonics to study Laplace's equation inside a sphere may be regarded as an early (and of course unconscious) application of non-commutative harmonic analysis.

Lagrange's observation of 1773 was only a small part of his contribution to the development of mathematical physics. He improved and extended the investigations of Bernoulli, Euler, and D'Alembert in many ways, and in competition with Laplace he made important inroads into celestial mechanics. He is credited above all, however, with completing the transition from a geometrical to an analytical point of view in dealing with mechanics and with basing the whole subject on simple general principles. His *Mécanique analytique*, published in 1787 exactly one hundred years after Newton's *Principia*, is a magnificent survey of the discoveries of Galileo, Newton, Bernoulli, Euler, D'Alembert, etc., all reworked into a coherent elegant scheme. His thoroughly analytical point of view is exemplified by his boast that the book contains no diagrams.

8. THE WORK OF FOURIER, POISSON, AND CAUCHY, AND EARLY APPLICATIONS OF HARMONIC ANALYSIS TO PHYSICS

The work of the eighteenth century on continuum mechanics was severely handicapped by a lack of systematic methods for solving or otherwise coping with the partial differential equations to which Euler and his contemporaries had been led. Only simple equations like $\frac{\partial^2 \psi}{\partial t^2} = \frac{1}{a^2} \frac{\partial^2 \psi}{\partial x^2}$ could be dealt with in a general way. The clue provided by the theory of this equation was missed — not only by Euler and D'Alembert as indicated in the last section, but again by Lagrange in a memoir on the theory of sound published in 1759.

The breakthrough came in 1807, when J. B. Fourier (1768-1830) submitted a long memoir on the conduction of heat to the French Academy of Sciences. Simple hypotheses and arguments led to the conclusion that the variable temperature T in a homogeneous body will satisfy a partial differential equation of the form $\frac{\partial T}{\partial t} = \mu\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2}\right)$ where μ is a constant depending on the material of which the body is made, x, y, and z are spatial coordinates, and t is the time. With this equation as a starting point Fourier made a profound analysis of a number of problems in heat flow. Once the temperatures at the boundaries are specified, it turns out that the distribution of temperature over the body at time $t = 0$ determines this distribution at all later times, and this later distribution can be calculated from a simple algorithm.

The simplest case to discuss is that in which the body is a thin bar, insulated so that heat flows in and out only through the ends and T depends only on the coordinate x. Then the equation becomes $\frac{\partial T}{\partial t} = \mu \frac{\partial^2 T}{\partial x^2}$. If the temperatures are held fixed at A and B at the ends and these occur at $x = 0$ and $x = \ell$, we may write $T = \left(\frac{B-A}{\ell}\right)x + A + \tilde{T}$ where $\tilde{T}$ satisfies the same equation but $\tilde{T}(0, t) = \tilde{T}(\ell, t) = 0$. Now $\sin\frac{\pi n x}{\ell}$ is zero at $x = 0$ and $x = \ell$ for all $n = 1, 2, \cdots$. Thus if $g_t(x) = \tilde{T}(x, t)$ can be written as a linear combination of the functions $x \to \sin\frac{\pi n x}{\ell}$, the coefficients will depend upon t and we will have $\tilde{T}(x, t) = \sum_{n=0}^{\infty} c_n(t) \sin\frac{\pi n x}{\ell}$ where the $c_n(t)$ remain unknown. However, at least at a formal level one verifies at once that $\tilde{T}$ satisfies the partial differential equation of heat conduction if and only if $\frac{d}{dt} c_n(t) = \frac{-\mu\pi^2 n^2}{\ell^2} c_n(t)$ so that $c_n(t) = e^{(-\mu\pi^2 n^2 t)/\ell^2} c_n(0)$. This implies that $\tilde{T}(x,t) = \sum_{n=0}^{\infty} c_n(0) e^{(-\mu\pi^2 n^2 t)/\ell^2} \sin\frac{\pi n x}{\ell}$, and we need only know the constants $c_n(0)$ to know $\tilde{T}(x, t)$ for all x and t. But these constants are just the expansion coefficients of $\tilde{T}(x, 0)$ in the Fourier series expansion $\tilde{T}(x, 0) = \Sigma c_n(0) \sin\frac{\pi n x}{\ell}$ and can be computed from Fourier's formula

$$\int_0^\ell \tilde{T}(x, 0) \sin\frac{\pi n x}{\ell}\, dx = c_n(0) \int_0^\ell \sin^2\frac{\pi n x}{\ell}\, dx .$$

This is not the way Fourier proceeded, but is perhaps the easiest way to understand the reason for the truth of Fourier's algorithm, which amounts simply to this: Expand the reduced temperature $\tilde{T}$ at $t = 0$ in a sine series, computing the coefficients as indicated. To get the temperature at time t,

simply multiply the nth coefficient by $e^{(-\mu\pi^2 n^2 t)/\ell^2}$ and add up the modified sine series.

The key point of course is the assumption that a more or less arbitrary function on a finite interval can be written as the sum of a linear function and a sine series $\sum_{n=1}^{\infty} c_n \sin \frac{n\pi x}{\ell}$. This is what Fourier insisted upon and what his eighteenth-century predecessors had refused to believe. As a matter of fact, Fourier's 1807 memoir was rejected by the French Academy as insufficiently rigorous — Lagrange was a member of the jury. On the other hand, he was encouraged to continue his admittedly very original researches and in 1812 won an academy prize for another version of the same memoir. This also was criticized for insufficient rigor and Fourier's work was not published in detail until the appearance in 1822 of his immensely influential classic *Théorie analytique de la chaleur.*

The importance of Fourier's treatise of course does not lie in its applications to solving problems in heat flow, but to the universality of the methods he employed (not to mention the influence on foundational questions and the development of set theory produced by the study of convergence and other points of rigor). There is no problem in extending the expansion technique to functions of several variables. One simply replaces $\sin\frac{n\pi x}{\ell}$ by $\sin\frac{n\pi x}{\ell_1}\sin\frac{m\pi y}{\ell_2}$ for two variables and the extension to more variables is obvious. More generally, the method is one that can be adapted to deal with the wave equation, with Laplace's equation, and in fact with any linear partial differential equation with constant coefficients. Not since Newton and Leibniz introduced the calculus well over a century earlier had mathematical physicists been provided with so powerful a tool. Now the partial differential equations that had accumulated (and were still accumulating) as the mathematical analysis of physical phenomena proceeded could be solved and their implications studied. It was an enormous advance.

Before saying more about what Fourier and his contemporaries accomplished, let us look briefly at how Fourier's method fits into the general scheme of harmonic analysis as outlined earlier. A function of x, y, and z which is periodic in each variable with periods ℓ_1, ℓ_2, and ℓ_3 respectively can be regarded as a function on the group obtained from the additive group of all triples of real numbers by factoring out the subgroup of all triples of the form $n_1\ell_1, n_2\ell_2, n_3\ell_3$ where n_1, n_2, and n_3 are integers. The continuous characters on this group are just the functions $x, y, z \to e^{(2\pi i n_1 x)/\ell_1} e^{(2\pi i n_2 y)/\ell_2} e^{(2\pi i n_3 z)/\ell_3}$ where n_1, n_2, and n_3 are integers. Since the real and imaginary parts of $e^{(2\pi i n x)/\ell}$ are just $\cos\frac{2\pi n x}{\ell}$ and $\sin\frac{2\pi n x}{\ell}$, Fourier's theorem on expanding in a sine series is easily deducible from a theorem permitting the expansion of a more or less arbitrary functions on our quotient group in terms

of characters. Functions on intervals, rectangles, etc., may be looked upon of course as restrictions of periodic functions on the line, the plane, etc., respectively. The expansion is useful in solving partial differential equations with constant coefficients for just the same reason that passing to the generating function is useful in solving the difference equations of probability theory. If one thinks of the set of coefficients as a function on the group of characters and thinks of this function as the primary unknown, the differential equation becomes an algebraic equation. This is because partial differentiation transforms into multiplication by a function. The chief difference between Fourier's application of this principle and the earlier applications in probability and number theory lie in the fact that Fourier was dealing with a continuous group.

In dealing with problems on all of space or on the whole real line, one can apply the same principles but not to a compact quotient group. One must deal with the full locally-compact group of all n-tuples of real numbers for various values of n. The most general continuous character on R^n is $x_1, \cdots, x_n \to e^{i(x_1 z_1 + \cdots + x_n z_n)}$ where the z_j are arbitrary complex numbers and these characters are bounded or equivalently of absolute value one when and only when the z_j are all real. In any event there is a whole continuum of possible characters, and infinite sums have to be replaced by integrals over this continuum. Instead of the formula $f(x) = \sum_{n=-\infty}^{\infty} c_n e^{inx}$ where $c_n = \frac{1}{2\pi}\int_0^{2\pi} f(x)\, e^{-inx}\, dx$, one has $f(x) = \int_{-\infty}^{\infty} c(y)\, e^{ixy}\, dy$ where $c(y) \equiv \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)\, e^{-ixy}\, dx$ and similarly for higher dimensions. Taking f to be real and even, these formulae reduce to $f(x) = \int_0^{\infty} 2c(y) \cos xy\, dy$ and $2c(y) = \frac{2}{\pi}\int_0^{\infty} f(x) \cos xy\, dx$, and in this form were known and used by Fourier in his 1812 prize paper.

The applications of Fourier's ideas to other branches of physics did not have to wait for the publication of his book in 1822. Poisson (1781-1840) and Cauchy (1789-1857) were twenty-six and eighteen years old respectively in 1807, when the thirty-nine-year-old Fourier submitted his memoir to the French Academy, and both began after a while to study the partial differential equations of physics and to apply the methods of harmonic analysis. Indeed a sort of three-cornered competition arose. Although Fourier had published no details, Poisson read the 1807 memoir in manuscript and published a five-page summary and review of it in 1808. Moreover, eight years later Fourier published his own summary — including his ideas on the Fourier integral. By 1816 both Poisson and Cauchy had written papers applying harmonic analysis to the solution of the wave equation in three dimensions, and in 1823 Cauchy published a paper explicitly pointing out how the Fourier transform made it possible to deal with an arbitrary linear partial differ-

ential equation with constant coefficients. In a paper published in 1817 Cauchy claims to have independently discovered the reciprocal formulae of the cosine transform

$$c(y) = \sqrt{\frac{2}{\pi}} \int_0^\infty f(x) \cos xy \, dx \quad f(x) = \sqrt{\frac{2}{\pi}} \int_0^\infty c(y) \cos xy \, dy,$$

but acknowledges the priority of Fourier. For a very full account of the whole story the reader is referred to the book *Joseph Fourier 1768-1830* [7] by I. Grattan-Guinness, which contains among other things the complete text of Fourier's 1807 memoir.

Simultaneously with this work on harmonic analysis and its applications to the wave equation, a new field of application was emerging in a revival of Huygens's ideas about the wave nature of light. Between 1801 and 1827 the work of Young (1773-1829) and Fresnel (1788-1827) with important contributions by Malus (1775-1812) and Arago (1786-1853) led to a complete overthrow of the corpuscular theory and the establishment of the wave theory as far superior in explaining known phenomena. Young's early work in explaining diffraction patterns and the colors of thin films as due to interfering waves was not well received in spite of the fact that he showed how to compute the wave length of the light from the diffraction pattern. However, in 1810 Malus accidentally passed light reflected from a window pane through a doubly refracting crystal and found that the two refracted rays were of radically different intensities. He had discovered the polarization of reflected light, and Malus investigated this phenomenon in detail. In 1816 Fresnel and Arago discovered that oppositely polarized light rays do not interfere, and Young offered an explanation based on the hypothesis that light waves are transverse. Transverse waves are waves in which the vibrations take place perpendicular to the direction of propagation. If one thinks of oppositely polarized waves as being waves in which the vibrations are perpendicular to one another as well as to the common direction of motion, one can understand the non-interference. Fresnel published an elaborate memoir based on these ideas in 1827. One great problem remained, however. If light is a wave, what is it that is waving and what are its properties? The only known examples of transverse waves occurred in the vibrations of elastic solids. Stimulated by the memoir of Fresnel, Poisson and Cauchy took up the study of the small vibrations of elastic solids. Cauchy was the first to write down the correct system of three linear second order differential equations. He did this in 1828, and in the same year Poisson analyzed this system and showed that there would be two kinds of waves, longitudinal and transverse, each with its own characteristic velocity. In the ensuing years Cauchy made three attempts to find a possible elastic solid whose transverse waves would behave like light, but none succeeded. Decades later the puzzle was solved by Maxwell's theory of the oscillations of an electromagnetic field.

9. HARMONIC ANALYSIS, SOLUTIONS BY DEFINITE INTEGRALS, AND THE THEORY OF FUNCTIONS OF A COMPLEX VARIABLE

The mapping set up by the Fourier transform $f(x_1, \cdots, x_n) \to \hat{f}(y_1, \cdots, y_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \cdots, x_n)\, e^{i(x_1 y_1 + \cdots + x_n y_n)}\, dx_1 \cdots dx_n$ has an important formal property which is an integrated counterpart of the fact that partial differentiation transforms into multiplication by $-i$ times the corresponding coordinate. It is the property that the product of two Fourier transforms $\hat{f}$ and $\hat{g}$ is the Fourier transform $\hat{h}$ of a function h which can be constructed from f and g by a simple integral formula and is called then *convolution*. $h(x_1, x_2, \cdots, x_n) = f * g(x_1, \cdots, x_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1 - t_1, x_2 - t, \cdots, x_n - t_n)\, g(t_1, t_2, \cdots, t_n) dt_1\, dt_2 \cdots dt_n$.

If one solves a linear partial differential equation with constant coefficients by using the Fourier transform to turn it into an algebraic equation, it will often turn out that one has an explicit expression for the Fourier transform of the unknown function as the product of the Fourier transform of a given function and some other explicitly known function which is determined by the partial differential equation in question. Taking the inverse Fourier transform, the solution to the problem is exhibited as the convolution of the given function with the inverse Fourier transform of the function determined by the differential equation. For example, consider the problem of solving $\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = 4\pi\varrho$ where ϱ is known and V is to be determined. Taking Fourier transforms one has $-(u^2 + v^2 + w^2)\hat{V}(u,v,w) = 4\pi\hat{\varrho}$ so that $\hat{V} = -\frac{1}{u^2 + v^2 + w^2}\, 4\pi\hat{\varrho}$. Taking inverse Fourier transforms one finds that $V = 4\pi\varrho * g$ where g is the inverse Fourier transform of $-\frac{1}{u^2 + v^2 + w^2}$ and this can be computed to be $\frac{1}{4\pi\sqrt{x^2 + y^2 + z^2}}$. Thus

$$V(x,y,z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{\varrho(x - x', y - y', z - z')}{\sqrt{(x')^2 + (y')^2 + (z')^2}} dx'\, dy'\, dz',$$

which may be recognized as the formula for computing the potential due to a charge or mass distribution of density ϱ. The equation it solves is Poisson's correction of Laplace's equation for points of space at which the density is not zero. Of course, these formal considerations are only valid when the functions satisfy suitable regularity and boundedness conditions. Other examples one can give include the general solution of the wave equation as a "superposition of plane waves" and the solution of the initial value problem for the heat equation as a convolution of the initial temperature distribution with the function $x,y,z \to \frac{1}{(2a\sqrt{\pi t})^3}\, e^{-(x^2 + y^2 + z^2)/4a^2 t}$.

Many of these explicit solutions can be (and originally were) obtained by other methods that do not involve the use of Fourier analysis. Poisson was

particularly active in developing such integral formulae. He seems to have been the first to recognize that the existence of conductors of electricity combined with Coulomb's law presented a challenging mathematical problem: How does a given charge distribute itself on a conductor? And, more generally, given a system of a finite number of conductors with a given charge on each, how does the charge distribute itself? It is easy to reduce this problem to a purely mathematical one involving Laplace's equation, and Poisson created (mathematical) electrostatics with a long memoir on the subject published in 1812. Twelve years later he published an important memoir on magnetism showing how the inverse square law for magnetic poles and the apparent non-existence of isolated magnetic poles both follow from a theory in which the fundamental entity is a continuous distribution in space of so called "magnetic dipoles." Let $\alpha^2 + \beta^2 + \gamma^2 = 1$ and consider two magnetic poles of strengths $\frac{-m}{\epsilon}$ and $\frac{m}{\epsilon}$ located at $x_0 \mp \frac{\epsilon\alpha}{2}, y_0 \mp \frac{\epsilon\beta}{2}, z_0 \mp \frac{\epsilon\gamma}{2}$. Then the net magnetic field they produce has a limit as ϵ tends to zero. It is called the field of a dipole at x_0, y_0, z_0 whose dipole moment is the vector $m\alpha, m\beta, m\gamma$. From a strictly mathematical point of view, Poisson's assumption (when the field behaves suitably at ∞) is equivalent to the astion that the magnetic field components H_x, H_y, H_z satisfy the partial differential equation $\frac{\partial H_x}{\partial x} + \frac{\partial H_y}{\partial y} + \frac{\partial H_z}{\partial z} = 0$ (in vector form div $\vec{H} = 0$). Indeed, introducing the vector notation $\text{curl}(\vec{A}) = \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}, \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}, \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right)$, one verifies easily that $\vec{A}$ is uniquely determined by div $\vec{A}$ and curl $\vec{A}$ provided that it goes to zero properly at ∞. Since div $\vec{H} = 0$ one can attempt to find a formula expressing H in terms of curl $\vec{H}$. This amounts to solving a system of partial differential equations of the form div $\vec{H} = 0$, curl $\vec{H} = \vec{I}$ and the method of Fourier transforms may be applied. It works and leads to a formula for $\vec{H}$ in terms of $\vec{I}$ of (vector) convolution type. This formula may be interpreted as asserting that $\vec{H}$ is the field due to a distribution of dipoles of density $\frac{\vec{I}}{4\pi}$. The relationship between $\vec{H}$ and $\vec{I}$ is quite analogous to that between V and ϱ in electrostatics. Indeed, just as any vector field vanishing suitably at ∞ and having zero divergence can be regarded as the magnetic field due to a continuous distribution of magnetic dipoles, any vector field with zero curl (and suitably vanishing at ∞) can be regarded as the electric field due to a continuous distribution of charge. We are dealing on the one hand with the problem of expressing $\vec{H}$ in terms of curl $\vec{H}$ given that div $H = 0$ and on the other with the problem of expressing $\vec{E}$ in terms of div $\vec{E}$ given that curl $\vec{E} = 0$.

The early nineteenth century was an exciting period in the development of physics and chemistry quite apart from the discovery of powerful mathe-

matical methods. We have already mentioned the renaissance and development of the wave theory of light. Another advance of tremendous importance was the discovery by Oersted (1777-1851) in 1819 of a direct relationship between electricity and magnetism — a discovery made possible by Volta's invention of the electric battery in 1800. The work of Volta (1745-1827), based on still earlier work of Galvani (1737-1798) and others on "animal electricity," was of equal importance in its own right. It made steady electric currents available for the first time and through the work of Nicholson (1753-1816) and Carlisle (1768-1840) in the same year forged a fundamental link between electricity and chemistry. Nicholson and Carlisle showed that (impure) water can be decomposed into its elements by passing an electric current through it. What Oersted did in 1819 was to observe that an electric current in a wire deflects nearby compass needles. Almost immediately the quantitative aspects of this unexpected new phenomenon were under intensive investigation by a number of scientists, the most important being Biot (1774-1862), Savart (1791-1841), and Ampère (1775-1836). Ampère investigated the magnetic effects of one current on another and published a long and important memoir on electromagnetism in 1825. The basic quantitative law is usually attributed to Biot and Savart and was formulated by them in infinitesimal physical terms. An equivalent formulation in the spirit of our discussion of Poisson's theory of ordinary magnetism is that the magnetic field produced by any finite system of moving charges satisfies the partial differential equation $\operatorname{div} \vec{H} = 0, \operatorname{curl} \vec{H} = \frac{4\pi \vec{i}}{c}$ where $\vec{i}$ is the current density and c is a fundamental constant. One can recover the Biot-Savart law by integrating these equations and expressing $\vec{H}$ as a (vector) convolution of the appropriate function with $\vec{i}$. The mathematics is identical with that of Poisson's theory. One simply treats an "infinitesimal" current element $\vec{i}$ as a point dipole of (vector) dipole moment $\frac{i}{c}$. A few years later, in 1830, Faraday (1791-1867) and Henry (1797-1878) independently discovered that there is a converse relation between electricity and magnetism. A changing magnetic field produces an electric field. When this effect was quantitatively understood, one saw that the equation $\operatorname{curl} \vec{E} = 0$ is valid only when $\vec{H}$ does not change with time, and more generally should read $\operatorname{curl} \vec{E} = -\frac{1}{c}\frac{\partial \vec{H}}{\partial t}$. That the equation $\operatorname{curl} \vec{H} = \frac{4\pi i}{c}$ has to be similarly corrected by adding $\frac{1}{c}\frac{\partial \vec{E}}{\partial t}$ to the right-hand side was recognized only a generation later (on theoretical grounds) by a man who was born just after Faraday and Henry made their discoveries. Maxwell (1831-1879) made this proposal in the 1860s, showed that it implied that both $\vec{E}$ and $\vec{H}$ satisfy the wave equation $\frac{1}{c^2}\frac{\partial^2 \vec{A}}{\partial t^2} = \left(\frac{\partial^2 \vec{A}}{\partial x^2} + \frac{\partial^2 \vec{A}}{\partial y^2} + \frac{\partial^2 \vec{A}}{\partial z^2}\right)$ at points of space free of charge, and

was led thereby to his celebrated electromagnetic theory of light. The constant c is of course just the velocity of light — a fact which had been noted some years earlier as a curious and possibly significant coincidence.

The main point I have been trying to make in this section so far is that the use of formulas of convolution type is thinly disguised harmonic analysis and that in this disguise harmonic analysis was a key factor in the development of electricity and magnetism in the early nineteenth century as well as in the theory of heat conduction and wave propagation.

Another major development of the early nineteenth century was the founding of the theory of functions of a complex variable by Gauss and Cauchy. The use of calculations involving complex numbers, which were more or less equivalent to the Cauchy integral formula, goes back well into the eighteenth century, but it was mistrusted and not well understood. Complex numbers themselves were still regarded as rather mysterious entities in the early nineteenth century. The exact history is rather complicated and I shall not attempt to trace it. I shall rather content myself with stating that Gauss and Cauchy founded the theory in the sense that they systematized and rigorized earlier uses of the basic idea. Its formal birthdate is often taken to be 1825, the year in which Cauchy published the integral formula that bears his name, although Cauchy and Gauss both knew the result a decade earlier. Our main purpose here is to indicate briefly the very close connections of this theory with harmonic analysis.

Consider a function f which depends analytically on the complex variable z inside the disk $|z| < R$. For each r with $0 < r < R$ let f_r denote the restriction of f to the circle $|z| = r$. Expanding in a Fourier series one finds that $f_r(re^{i\theta}) = \sum_{n=-\infty}^{\infty} c_n(r)\, e^{in\theta}$, and the analyticity condition implies that $c_n(r) = c_n r^n$. Finally the continuity at the origin implies that $c_n = 0$ for $n < 0$. Thus $f(re^{i\theta}) = \sum_{n=0}^{\infty} c_n\, r^n\, e^{in\theta}$ so if $z = re^{i\theta}$, $f(z) = \sum_{n=0}^{\infty} c_n\, z^n$ and the expansibility of an analytic function in a power series is established. Of course the coefficients c_n may be determined from any f_r. Thus if $r_1 < r_2$ we have $f(r_1 e^{i\theta}) = \sum_{n=0}^{\infty} c_n\, r_1^n\, e^{in\theta}$ where $c_n r_2^n = \frac{1}{2\pi}\int_0^{2\pi} f(r_2\, e^{i\phi})\, e^{-in\phi}\, d\phi$. Hence $f(r_1\, e^{i\theta}) = \frac{1}{2\pi}\sum_{n=0}^{\infty}\left(\frac{r_1}{r_2}\, e^{i\theta}\right)^n \int_0^{2\pi} f(r_2\, e^{i\phi})\, e^{-in\phi}\, d\phi = \frac{1}{2\pi}\int_0^{2\pi}\left(\sum_{n=0}^{\infty}\left(\frac{r_1}{r_2}\, e^{i(\theta-\phi)}\right)^n\right) f(r_2\, e^{i\phi})\, d\phi$. But

$$\sum_{n=0}^{\infty}\left(\frac{r_1}{r_2}\, e^{i(\theta-\phi)}\right)^n = \frac{1}{1-\dfrac{r_1 e^{i\theta}}{r_2 e^{i\phi}}},$$

so that $f(r_1\, e^{i\theta}) = \frac{1}{2\pi}\int_0^{2\pi} \dfrac{f(r_2 e^{i\phi})\, d\phi}{1-\dfrac{r_1\, e^{i\theta}}{r_2\, e^{i\phi}}}$, and writing $z = r_1 e^{i\phi}$, $\zeta = r_2 e^{i\phi}$ thus

becomes $f(z) = \frac{1}{2\pi i} \int_{|\zeta| = r_2} \frac{f(\zeta)\, d\zeta}{\zeta - z}$, which is Cauchy's integral formula for the circle $|z| = r_2$.

With the Cauchy integral theorem on the circle and the expansibility of an analytic function in a power series as a starting point, one can deduce the basic theorems in elementary complex variable theory quite quickly and easily. Thus there is a sense in which the theory of functions of a complex variable is an aspect of Fourier analysis. From another point of view the theory of functions of a complex variable is the theory of the solutions of the Cauchy-Riemann equations $\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}$, $\frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$, and the properties of this system can be deduced from Fourier analysis just as with the partial differential equations of electromagnetism. Indeed, thinking of v, u as a two vector, the Cauchy-Riemann equations are just the two-dimensional version of $\operatorname{curl} \vec{E} = \operatorname{div} \vec{E} = 0$.

The argument used to deduce the Cauchy integral theorem for the circle from the theorem on expansibility of periodic functions in a Fourier series can be used in almost exactly the same way to deduce Poisson's formula expressing a harmonic function in a disk in terms of its boundary values. There is a similar formula (also due to Poisson) expressing a function harmonic in a ball in terms of its boundary values, and it is natural to ask whether this formula can be deduced in an analogous fashion. This would seem to require giving the surface of the sphere the structure of a commutative group and using an expansion theorem in terms of its characters. While it can be proved that this surface S cannot be made into a group at all in a manner consistent with its topology, there is a compact non-commutative group which acts transitively on S—namely $SO(3)$, the rotation group in three dimensions. Moreover, there is an analogue of the Fourier series proof of Poisson's formula. It differs from the Fourier series proof in using expansions in the surface harmonics of Legendre and Laplace (see section 7) instead of the $e^{in\theta}$. This suggests that there might be some relationship between surface harmonics on a sphere and complex exponentials on the circle. There is, and understanding this relationship is the key to understanding why the theory of group representations may be regarded as a natural generalization of Fourier analysis. The rotation group in two dimensions acts transitively on the unit circle with center 0, 0 and takes each function $\theta \to e^{in\theta}$ into a constant multiple of itself. Moreover, every one-dimensional invariant subspace of measurable functions on the circle is the one-dimensional subspace of all multiples of $e^{in\theta}$ for some n. The rotation group in three dimensions acts transitively on the unit sphere with center 0,0,0 but has *no* invariant one-dimensional subspace of measurable functions except for the space of constants. On the other hand it has finite-dimensional invariant subspaces, which are irreducible in the sense that no proper sub-

space is invariant. These are mutally orthogonal with respect to integration over the sphere and each is spanned by surface harmonics. Thus the expansion theorem in surface harmonics may be looked upon more fundamentally as an expansion theorem in terms of members of irreducible invariant subspaces. Each such subspace defines a representation of $SO(3)$ by linear transformations, and it is these irreducible representations which replace characters in dealing with a transitive action of a non-commutative group. A commutative group has only one-dimensional irreducible representations. In a sense the theory of surface harmonics of Legendre and Laplace is an anticipation of non-commutative harmonic analysis made over a century before the world was ready to appreciate it as such.

Poisson's formula expressing a harmonic function inside a ball in terms of its values on the spherical surface bounding the ball can be used to show that for every continuous function ϕ on this surface, there is a unique solution of Laplace's equation continuous at the boundary and having the values of ϕ as boundary values. The corresponding result for a rectangular parallelepiped can easily be proved using Fourier series in three variables, and it is natural to conjecture a general theorem applying to more or less arbitrary closed surfaces. The first person to state and seriously to attempt a proof of this important theorem was G. Green (1791-1841). Green did not get to a university until he was forty, but he read Poisson's 1812 paper on electrostatics and Fourier's book on heat conduction while helping his father run the family mill and bakery in Nottingham. Starting to think about things for himself, Green managed to go much further than Poisson. In 1828 he published a remarkable paper entitled "An essay on the application of mathematical analysis to the theories of electricity and magnetism." Unfortunately he was too much out of touch with the scientific world to think of sending it to a journal. He had it privately printed and circulated in Nottingham; it did not become widely known until Lord Kelvin (then William Thompson) stumbled upon it and arranged for its publication in *Crelle's Journal* in 1850. In the meantime much of its contents had been rediscovered by Gauss and others. It contained the well-known Green's identities from which one deduces the uniqueness of solutions of Laplace's equation with given boundary values and Green's famous "physical" argument for the existence of what is now called a Green's function. From its alleged existence Green could deduce a generalization of Poisson's formula and the existence of a solution of Laplace's equation with arbitrary continuous boundary values. Gauss later gave a different proof, which was unsatisfactory in another way, and what came to be known as the "Dirichlet problem" remained open for many years. The first satisfactory solution for convex regions was given by C. Neumann (1832-1925) in 1870. Work on the problem was an important stimulus to the development of analysis—especially the theory of integral equations. The theory of integral equations led

in turn to the modern theory of operators in Hilbert space, which plays a central role in present day non-commutative harmonic analysis (see section 14).

10. ELLIPTIC FUNCTIONS AND EARLY APPLICATIONS OF THE THEORY OF FUNCTIONS OF A COMPLEX VARIABLE TO NUMBER THEORY

The usefulness of considering the functions of analysis for complex values of the argument became dramatically apparent in the years 1827-1829 when Abel (1802-1829) and Jacobi (1804-1851) published their memoirs, *Recherches sur les fonctions elliptiques* (in two parts) and *Fundamenta nova theoriae functionum ellipticarum* respectively. Independently (Abel slightly earlier) they had discovered and developed the consequences of the following important fact: The theory of the functions defined by the indefinite integrals that arise when one attempts to find the length of an arc of an ellipse becomes very much simpler if one a) concentrates attention on the inverse functions and b) considers complex as well as real values of the variables. These inverse functions are analytic except for poles in the whole complex plane, and moreover have two complex periods ω_1 and ω_2 which are not real multiples of one another: $f(z + \omega_1) = f(z) = f(z + \omega_2)$ for all z. Any function with these properties is called an elliptic function. Because of its two independent periods an elliptic function is uniquely determined by its values in a parallelogram with one vertex at 0, and because of its analyticity except for poles it is determined up to a multiplicative constant by the positions (and orders) of its (finitely many) zeros and poles in this parallelogram. It is thus possible to get a complete overview of all possible elliptic functions with given periods ω_1 and ω_2, and a very beautiful theory emerges.

Jacobi and Abel were interested in the dependence of their functions on ω_1 and ω_2 as well as on z and found innumerable elegant identities and relationships. In the course of such work Jacobi compared the power series expansions resulting from two different expressions for the same function and deduced the remarkable fact that $r_4(n)$, the number of representations of a positive integer n as the sum of four squares, is equal to eight multiplied by the sum of all divisors of n which are not divisible by 4. Since 1 and n count as divisors, it follows at once that *every* positive integer n can be written as a sum of four squares. This result, conjectured by Fermat, had been proved by Lagrange in 1773 after many years of unsuccessful attempts by Euler (Euler found his own proof a year later). The formula for the exact number of representations was completely new. Other identities in Jacobi's work led to similar formulae for r_2, r_6, and r_8. Of course that for r_2 was already known. Twenty years later Eisenstein (1823-1852) found purely arithmetical proofs of the formulae for r_6 and r_8 as well as similar formulae for r_5 and r_7,

but the analytical proofs for r_4, r_6, and r_8 could be understood only as curious accidents until well into the twentieth century. Between 1925 and 1940 Hecke and Siegel fitted them into two (somewhat different) beautiful general theories which will be described later in this article. Both theories are developments of the theory of "modular forms" begun by Dedekind, Klein, and Hurwitz in the last quarter of the nineteenth century and, since it is not difficult, I shall give a brief account here of how modular forms relate to elliptic functions on the one hand and Jacobi's number-theoretical results on the other.

For each pair ω_1 and ω_2 of independent periods, there is a canonical elliptic function $\wp$ characterized by the fact that it has second order poles at all points $n\omega_1 + m\omega_2$ of the discrete group of all periods and no other poles, and that the principal part at each pole z_j is $\frac{1}{(z - z_j)^2}$. A key (and not difficult) result in the theory of elliptic functions is that every elliptic function with periods ω_1 and ω_2 is a rational function of $\wp$ and its derivative $\wp'$, and that $\wp$ satisfies a differential equation of the form $\wp'(z)^2 \equiv 4(\wp(z))^3 - g_2\wp(z) - g_3$ where g_2 and g_3 are "constants" which of course depend upon ω_1 and ω_2. If one investigates g_2 and g_3 as functions of ω_1 and ω_2, two facts emerge very easily. Substitution of λz for z in the differential equation leads at once to the conclusion that g_2 and g_3 are homogeneous of orders -4 and -6 respectively; that is, $g_2(\lambda\omega_1,\lambda\omega_2) \equiv \frac{1}{\lambda^4} g_2(\omega_1, \omega_2)$ and $g_3(\lambda\omega_1, \lambda\omega_2) = \frac{1}{\lambda^6} g_3(\omega_1, \omega_2)$. It follows trivially that g_2 and g_3 may be written in the form $g_2(\omega_1, \omega_2) = \frac{1}{\omega_2^4}\phi_2\left(\frac{\omega_1}{\omega_2}\right)$ and $g_3(\omega_1, \omega_2) = \frac{1}{\omega_2^6}\phi_3\left(\frac{\omega_1}{\omega_2}\right)$ where ϕ_2 and ϕ_3 are functions of one variable. On the other hand it is clear that g_2 and g_3 depend only on the discrete group of all periods and not on ω_1 and ω_2 themselves. Suppose then that $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ is any two-by-two matrix of integers with $ad - bc = 1$. If $\omega'_1 = a\omega_1 + b\omega_2$ and $\omega'_2 = c\omega_1 + d\omega_2$, then ω'_1 and ω'_2 generate the same group that ω_1 and ω_2 do. Hence $g_2(\omega_1,\omega_2) = g_2(\omega'_1,\omega'_2)$ and $g_3(\omega_1,\omega_2) = g_3(\omega'_1,\omega'_2)$. Expressing g_2 and g_3 in terms of ϕ_2 and ϕ_3, one discovers immediately that ϕ_2 and ϕ_3 have the property that $\phi_2\left(\frac{az + b}{cz + d}\right) = (cz + d)^4\phi_2(z)$ and $\phi_3\left(\frac{az + b}{cz + d}\right) = (cz + d)^6\phi_3(z)$ for all integer matrices $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ of determinant one. The functions ϕ_2 and ϕ_3 can also be shown to be analytic throughout the upper half of the complex plane and to be bounded as z approaches infinity along the imaginary axis.

Quite generally, a function f analytic in the upper half-plane and bounded as just indicated is said to be a *modular form* of weight k (or dimension $-2k$) if $f\left(\frac{az + b}{cz + d}\right) = (cz + d)^{2k}f(z)$ for all z in the upper half-plane. The functions ϕ_2 and ϕ_3 defined by g_2 and g_3 as above were not only the first

modular forms to be considered as such, but turn out to generate all the others. In the general theory of modular forms it is shown that ϕ_2^3 and ϕ_3^2 span the two-dimensional vector space of all modular forms of weight 6, and a member of the one-dimensional subspace vanishing at ∞ is singled out and called Δ. Given a positive integer k, the equation $2\alpha + 3\beta + 6\gamma = k$ has a finite number of solutions in non-negative integers α, β, γ. Moreover, for each solution, $\phi_2^\alpha \phi_3^\beta \Delta^\gamma$ is evidently a modular form of weight k. These particular modular forms of weight k turn out to span the vector space of all modular forms of weight k.

If f is any modular form of weight k and we choose $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$, the identity $f\left(\frac{az + b}{cz + d}\right) = (cz + d)^{2k} f(z)$ reduces to $f(z + 1) = f(z)$ so that f has a Fourier series expansion $\sum_{n=-\infty}^{\infty} c_n e^{2\pi i n z}$ which converges in the upper half-plane. The condition on f at $i\infty$ implies that $c_n = 0$ for $n < 0$ so that $f(z)$ always has the form $\sum_{n=0}^{\infty} c_n e^{2\pi i n z}$. The c_n are called the Fourier coefficients of the form and c_0 is called the constant term. The constant term c_0 may be thought of as the value of f at $i\infty$, and forms for which it is zero are called *cusp forms*. For $k = 2, 3, 4, \cdots$ there is a very simple and straightforward way of defining a modular form of weight k which is *not* a cusp form. One simply sums the so-called Eisenstein series $G_k(z) = \sum_{n,m \neq 0,0}^{\infty} \frac{1}{(nz + m)^{2k}}$ where n and m are integers. This sum obviously satisfies the identity characteristic of modular forms of weight k, and it is not difficult to check that it has the necessary analyticity and boundedness properties. Simple arguments based on the easily established formula $\frac{1}{(\sin \pi z)^2} = \sum_{n=-\infty}^{\infty} \frac{1}{(z + n)^2}$ allow one to compute the Fourier coefficients of G_k, to show that it is not a cusp form, and more significantly to show that these coefficients have simple and suggestive number theoretical properties. Indeed $G_k(z) = c_0 + \frac{2(-1)^k 2\pi^{2k}}{(2k-1)!} \sum_{n=1}^{\infty} a_n e^{2\pi i n z}$ where $c_0 = 2[1 + \frac{1}{2^{2k}} + \frac{1}{3^{2k}} + \frac{1}{4^{2k}} \cdots] = 2\zeta(2k)$ and $a_n = \Sigma d^{2k-1}$ where d ranges over all the divisors of n. Evidently every modular form of weight k is uniquely a sum of a cusp form and a constant multiple of the Eisenstein series of weight k. The cusp forms also have Fourier coefficients with number-theoretical properties, but these are more subtle and were not discovered until well into the twentieth century.

The genesis of Jacobi's results on sums of squares can now be understood, at least in principle, by confronting the facts about the Fourier coefficients of Eisenstein series with a remarkable connection between sums of squares and modular forms which emerges when one applies the Poisson summation formula to the function $x \to e^{-ax^2}$. Let ϕ be any continuous function which goes to zero sufficiently rapidly at ∞ and let $\hat{\phi}(y) = \int_{-\infty}^{\infty} e^{ixy} \phi(x)\, dx$

be its Fourier transform. The Poisson summation formula is a simple consequence of elementary manipulations with Fourier series and asserts that $\sum_{n=-\infty}^{\infty} \phi(n) = \sum_{n=-\infty}^{\infty} \hat{\phi}(2\pi n)$. Now if $\phi(x) = e^{-ax^2}$, it is easy to compute that $\hat{\phi}(y) = \sqrt{\frac{\pi}{a}} e^{(-y^2)/4a}$ so that the Poisson summation formula yields $\sum_{n=-\infty}^{\infty} e^{-an^2} = \sqrt{\frac{\pi}{a}} \sum_{n=-\infty}^{\infty} e^{(-\pi^2 n^2)/a}$. This is valid for complex values of a having positive real part, and making the substitution $a = -\pi iz$ yields the identity

$$\sum_{n=-\infty}^{\infty} e^{\pi i n^2 z} = \sqrt{\frac{i}{z}} \sum_{n=-\infty}^{\infty} e^{(-\pi i n^2)/z}$$

for all z in the upper half-plane. Setting $\theta(z) = \sum_{n=-\infty}^{\infty} e^{\pi i n^2 z}$, this identity may be written as $\theta(z) = \sqrt{\frac{i}{\pi}} \theta(-\frac{1}{z})$, a relation found by Jacobi and sometimes called the Jacobi inversion formula. Using the Jacobi inversion formula and the fact that $\theta(z+2) = \theta(z)$, it is possible to show that for $k = 1, 2, \cdots$, θ^{4k} is "almost" a modular form—specifically that it is a modular form of weight k not for the whole "modular group" of all 2×2 integer matrices $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ with $ad - bc = 1$ but for the subgroup Γ_2 of all $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ with a and d odd and b and c even. On the other hand it is an immediate consequence of the definitions that $\theta^{4k}(z) = \left(\sum_{n=-\infty}^{\infty} e^{\pi i n^2 z}\right)^{4k} = \sum_{n=1}^{\infty} r_{4k}(n) e^{\pi i n z}$, so that the Fourier coefficients of this "almost modular form" are the representation numbers for sums of squares. To get actual theorems like Jacobi's by this route, the theory of Eisenstein series and their Fourier coefficients has to be extended to modular forms of "higher level" as was done by Hecke in 1927.

It is interesting to compare the role of the Poisson summation formula in the above considerations with its use in establishing the quadratic reciprocity law via the determination of the sign of a Gauss sum (see section 6). It is also interesting to note that Jacobi and Dirichlet, who introduced analysis into number theory in such different ways, were almost exact contemporaries and also very good friends.

Thirty years after the appearance of Jacobi's 1829 memoir on elliptic functions, Riemann (1826-1866) published a short note giving a still different application of the theory of functions of a complex variable to number theory and providing an important connecting link between the ideas of Jacobi and those of Dirichlet. Recall that in the mid-eighteenth century Euler proved that the sum of the reciprocals of the primes diverges by considering

the identity $\prod_p \frac{1}{1-\frac{1}{p^s}} \equiv \sum_{n=1}^{\infty} \frac{1}{n^s}$ for real values of the variable greater than 1, and that Dirichlet in 1837 extended Euler's ideas to the primes in an arithmetic progression. Riemann had the idea of obtaining more refined information about the distribution of the primes by considering Euler's function $\sum_{n=1}^{\infty} \frac{1}{n^s} = \zeta(s)$ for *complex* values of s. (The notation $\zeta(s)$ is due to Riemann.) It is evident that the series converges absolutely and uniformly in the right half-plane $\sigma > 1$ where $s = \sigma + i\tau$ and defines an analytic function there. Riemann went further, however, and showed that ζ could be continued to be analytic over the whole complex plane except for a first order pole at $s = 1$. Moreover, he found a simple functional equation connecting the values of ζ in the half-plane $\sigma > \frac{1}{2}$ with those in the half-plane $\sigma < \frac{1}{2}$. It reads $\frac{\zeta(s)\Gamma(s/2)}{\pi^{s/2}} = \frac{\zeta(1-s)\Gamma(\frac{1-s}{2})}{\pi^{(1-s)/2}}$ where Γ is Euler's well known gamma function. The gamma function is easily seen to be analytic in the entire complex plane except for simple poles at $-1, -2, -3, \cdots$ and to have no zeros. The significance of this identity is in the information it provides about the zeros and poles of ζ. The product formula $\zeta(s) = \prod_p \frac{1}{1-\frac{1}{p^s}}$ implies that there are no zeros or poles in the half plane $\sigma > 1$, and Riemann's functional equation allows one to conclude a) that the zeros and poles in the half plane $\sigma < 0$ are the poles and zeros of Γ and b) that the poles and zeros in the "critical strip" $0 < \sigma < 1$ are symmetrical about the center line. Thus ζ in addition to having a unique pole at $s = 1$ has no zeros outside the closure of the critical strip except for simple ones at $0, -1, -2, \cdots$. Since a function analytic except for poles (and well behaved at ∞) is almost determined by its zeros and poles, it is of interest to know more about the location of the unknown zeros inside and on the boundary of the strip. Riemann's famous unproved conjecture asserts that they all lie on the center line $\sigma = 1/2$. The much weaker result, that there are no zeros on the line $\sigma = 1$, was proved independently by Hadamard (1865-1963) and de La Vallée-Poussin (1866-1964) in 1896 and used by them to prove another of Riemann's conjectures.[2] This is the celebrated prime number theorem, which asserts that $\lim_{x\to\infty} \frac{\pi(x)\log x}{x} = 1$ where $\pi(x)$ is the number of primes less than or equal to x.

One of the two proofs that Riemann gave of his functional equation exhibits it as a corollary of Jacobi's inversion formula and hence of the Poisson summation formula. Moreover, the connection between the functional equation and Jacobi's formula is provided by the so-called Mellin trans-

form. This in turn is just the analytic continuation of the Fourier transform applied to functions on the multiplicative group of all positive real numbers.

11. THE EMERGENCE OF THE GROUP CONCEPT

While a modern mathematician looking back can see the pervasive role of group theory in the mathematics of the nineteenth century, this role was quite invisible to the mathematicians themselves until very late in the century. Until the end of the 1860s, group theory was the theory of finite permutation groups and its only application was to the theory of equations. Although Feit [5] prefers to regard Cauchy as having founded group theory in 1815, one can make a case for Lagrange's having done so forty-five years earlier. While Lagrange did not have the group concept — not even that of a group of permutations — he was the first to realize the significance of the study of permutations of the roots for the theory of equations. Moreover, his long memoir on the theory of equations published in 1770 stimulated the later work of Cauchy and Galois and contained in essence the proof of what is known today as Lagrange's theorem. This is the theorem that the order of a subgroup of a finite group necessarily divides the order of the group. The main purpose of Lagrange's memoir was to study systematically the various methods that had been found for solving polynomial equations of the second, third, and fourth degrees, to understand why they worked and what stood in the way of extending these methods to equations of the fifth and higher degrees. He found that the known methods could be understood in a unified way by considering what happened to rational functions of the roots when they were permuted amongst themselves. Lagrange's analysis gave strong indications that there was a difficulty in principle in solving fifth-degree equations in the same sense that this was possible for equations of lower degree — namely by formulas involving only rational operations and the taking of roots. However, an actual proof of the impossibility of such a solution was first given by Ruffini (1765-1822) in 1813. Abel gave another proof in 1824 without knowledge of Ruffini's work. Both men had read Lagrange's paper and were influenced by his ideas. Cauchy also read Lagrange's paper but was influenced in another way. In 1815 he introduced the concept of a group of permutations and in a series of papers developed some of the elementary theory of such groups. He is responsible for the notions of subgroup, transitive group, and conjugate elements, and he made an attempt at classifying his groups.

The first man to make really fundamental progress using the group concept was E. Galois (1811-1832), who combined the ideas of Lagrange, Cauchy, and Abel to produce a beautiful general theory explaining in terms

of group theory just why some equations are solvable in terms of radicals and others are not. He distinguished a certain subgroup of the group of all permutations of the roots of an equation by the symmetry properties of its elements. This group is now called the Galois group of the equation, and Galois showed that the solvability of the equation depends completely on the structure of the Galois group G. The equation is solvable if and only if there exists a family $N_0 \subset N_1 \subset N_2 \cdots N_k = G$ of subgroups such that N_j is normal in N_{j+1} for $j < k$ and the quotient group N_{j+1}/N_j is abelian. The concept of a normal subgroup and of a quotient group are due to Galois — and so is the word group. Galois wrote up his work for publication rather hurriedly just before engaging in the duel which was to take his life. His concise exposition of very original ideas was difficult for many to follow, and for this and other reasons publication was delayed for many years. The paper did not appear until 1846.

Kronecker (1823-1891) seems to have been the first to understand Galois's ideas thoroughly enough to carry them further. He published on the subject as early as 1853. His first really striking achievement, however, was his use of Galois theory to give a more perspicacious proof of an astonishing discovery of Hermite. Hermite (1822-1905) became interested in the theory of equations at an early age and, like Ruffini and Abel before him, succeeded in proving the impossibility of solving fifth-degree equations by radicals after reading Lagrange's memoir of 1770. Later he became a close student of the work of Abel and Jacobi on elliptic functions and in 1858 blended his two interests by showing that the roots of the general equation of the fifth degree could be expressed in terms of the coefficients if one used certain transcendental functions arising in the theory of elliptic functions and related to modular forms. Kronecker not only showed how to prove and understand this result, using Galois theory, but went on to study in depth the intricate relationships that exist between groups, polynomial equations, and elliptic functions.

The concept of an abstract group was formulated in 1854 by Cayley (1821-1895), but until around 1870 group theory remained a very specialized topic, known and understood by relatively few mathematicians and having the theory of equations as its only significant application. The first exposition of the theory to occur in a textbook appeared in 1866 as a section in Serret's *Cours d' Algebre Superieure*. The first book to be completely devoted to group theory was Jordan's (1838-1922) very influential *Traités des substitutions et des equations algébriques*, published in 1870. Almost simultaneously with the appearance of Jordan's comprehensive treatise, the scope of group theory began to increase dramatically as a consequence of the activities of two young mathematicians just beginning their careers. Sophus Lie (1842-1899) began in 1869 to study continuous groups with a view of doing for differential equations what Galois had done for algebraic

equations. In 1872 Felix Klein (1849-1925) devoted his inaugural lecture as a professor at Erlangen to announcing his celebrated program for unifying geometry through group theory and soon was developing the ideas of Hermite and Kronecker into an elaborate study of the interplay between discrete groups (both finite and infinite) and the theory of functions of a complex variable. We have already mentioned (see section 10) Klein's connection with the theory of modular forms. The first systematic treatment of this theory occurs in the thesis (published in 1881) of Klein's student Hurwitz (1859-1919) and is based quite directly on ideas of Klein and Dedekind published in the late 1870s. The so-called modular group of all matrices $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ with integer coefficients and determinant one could now of course be recognized and dealt with as a group. Above all, however, Klein was enthusiastic about the importance, unifying power, and wide applicability of group theoretic ideas and viewpoints. Being an energetic organizer and proselytizer by temperament, he did as much as or more than Jordan's book in spreading and popularizing group theory and the group concept.

After 1872 group theory developed rapidly in many directions, and mathematicians slowly became more and more group-conscious. An early development in the latter direction was the realization that Gauss not only had been working with finite commutative groups in his composition of quadratic forms, but in effect had proved that every such group is a direct product of cyclic groups. Deeper structure theorems for finite groups began to be found, beginning with the important Sylow theorems in 1872, and by 1897 Burnside (1852-1927) was able to publish a book on finite groups alone going far beyond the 1870 book by Jordan. Lie worked out his ideas on continuous groups between 1869 and 1884, and (with the collaboration of Engel) presented them in a three-volume treatise, the final volume of which appeared in 1893. Lie's central concept — that of the Lie algebra of a continuous group — made it possible to reduce many questions about the group (more precisely about its local behavior) to purely algebraic questions about the Lie algebra. In his thesis of 1894 E. Cartan (1869-1951) corrected the work of Killing (1847-1923) published between 1888 and 1890, and thereby gave a complete classification of all "simple" Lie algebras over the complex numbers. Not quite a decade later he found the corresponding classification for real Lie algebras. Klein's earlier work on discrete groups and complex variable theory led to his celebrated book of 1884, *Vorlesungen über das Ikosadeder und die Auflösung der Gleichung vom fünften Grade,* and his two volume treatise with Fricke, *Vorlesungen über die theorie der elliptischer Modulfunktionen,* published in 1890 and 1892. In 1881 a formidable rival to Klein appeared in the person of a young Frenchman named Henri Poincaré (1854-1912) who at that time began to develop his ideas on what are now called automorphic functions. Inspired by earlier work of Fuchs (1833-1902) and Schwarz (1843-1921) on the inverse functions of solutions

of second order differential equations, Poincaré was led to consider functions invariant under much more general discrete groups of two-by-two matrices than the subgroups of finite index of the modular group. Klein had been led in the same direction by his studies of Riemann's work on integrals of algebraic functions, and the final theory is a result of their combined efforts. Klein's version of the theory appeared in 1897 and 1901 in another two-volume treatise co-authored with Fricke. While the theory of automorphic functions was developing, infinite discrete groups appeared in a rather different connection. Their relationship to earlier studies in crystallography was realized, and with this application in mind Federov (1853-1919), Schoenflies (1853-1928), and Barlow (1845-1934) independently classified the so-called "space groups." A quarter of a century after Klein's famous inaugural lecture, group theory was a well established subject with lengthy treatises devoted to its various aspects.

12. INTRODUCTION TO SECTIONS 13-16

The group theoretical nature of expansions in Fourier series could not very well have been recognized until Jordan's book and Klein's missionary activities had had time to do their work, that is, until the very end of the nineteenth century. It was only in 1882 that Weber (1842-1913) formally defined the character notion for abstract finite commutative groups. Actually another three decades were required. The delay is not difficult to understand and may be attributed to the fact that (in spite of the efforts of Klein) analysts and mathematical physicists resisted thinking in group theoretical terms until well into the twentieth century. Moreover, noticing that the functions $x \to e^{inx}$ are group characters might have seemed of minor interest unless it also occurred to one to seek non-commutative analogues of these characters and so extend the scope of harmonic analysis. What actually happened was that a non-commutative extension of group characters was discovered in 1896 in a purely algebraic context. It was studied purely algebraically for over a quarter of a century and then extended to compact Lie groups. Specialization to the one-dimensional torus group made the connection obvious.

Fredholm's work on integral equations in 1900 and the introduction of the Lebesgue integral in 1902 inspired and made possible the modern L^2 theory of Fourier series and integrals as well as the spectral theory of self-adjoint and unitary operators in Hilbert space. This development took place more or less simultaneously with that described in the preceding paragraph and was destined to be blended with it in the general theory of unitary group representations, which emerged later. Although group theory played no role as such, one could see looking back that the work of Hilbert and his students on spectral theory (augmented by later work of Stone and von Neumann) was equivalent to a very thorough and complete reduction theory for

unitary representations of the additive group of the real line. Moreover, when properly formulated, this reduction theory turned out to have an essentially verbatim generalization to arbitrary locally-compact commutative groups.

Before these two lines of development could be blended as indicated, they both found extensive applications to physics in connection with the marvelously effective and subtle refinement of classical mechanics known as quantum mechanics. Quantum mechanics emerged between 1924 and 1927 after a quarter-century of confused, inconsistent, and semi-successful attempts to deal with the anomalies that arose when one attempted to explain phenomena by applying classical mechanics to the atoms of which matter was supposed to be composed. It seems almost miraculous that two sets of necessary and appropriate mathematical tools (later to be seen as parts of one whole) were being independently forged just as the need for them was arising.

Quantum mechanics not only provided a rich source of applications for the mathematical developments of the three preceding decades but also inspired and strongly influenced the unified theory to which they led. Before giving details, I shall devote the next three sections to independent discussions of the three components, beginning with a section on physics and proceeding through spectral theory, etc., to non-commutative characters.

13. THERMODYNAMICS, ATOMS, STATISTICAL MECHANICS, AND THE OLD QUANTUM THEORY

One of the more striking aspects of nineteenth-century physical science is the extent to which profound relationships between apparently independent phenomena were discovered. In earlier sections I have already mentioned the relationship between chemistry and electricity, between electricity and magnetism, and between electromagnetism and light. Another and by no means the least important was suggested at the turn of the century by work of Rumford (1753-1814) and Davy (1778-1829), but only became worked up into an exact quantitative theory around 1850. This is the relationship between mechanics and heat, whose principal features constitute the first and second laws of thermodynamics. Black's distinction between temperature and quantity of heat and his concept of specific heat (see section 7) depended on the (approximately verifiable) notion that heat behaves in many respects like a fluid and that when a hot body cools down because of contact with a colder one, the quantity of heat lost by one equals the quantity gained by the other. There is in effect a law of conservation of heat. In mechanical processes where there is "friction," however, heat seems to be created out of nothing, and when one deals with expanding gases, heat seems to disappear without reappearing elsewhere. Moreover, the conservation

law for mechanical energy implied by Newton's laws also seems to fail when bodies change temperature because of friction or the expansion and contraction of gases.

The first law of thermodynamics is in essence the statement that these two failures are mutually self-correcting. The quantity of heat that appears or disappears is directly proportional to the amount of mechanical energy that disappears or appears. There is a so-called "mechanical equivalent of heat" and, while neither mechanical energy nor quantity of heat is conserved by itself, there is a conservation law for the sum of the mechanical energy and the mechanical equivalent of the quantity of heat. While the existence of a well defined mechanical equivalent of heat was suggested by the experiments of Rumford and Davy, it did not become an accepted part of physics until Kelvin (1824-1907) drew attention to very careful measurements made by Joule (1818-1889) around 1840, and Helmholtz (1821-1894) published an influential paper in 1847. Helmholtz generalized the observation that heat and mechanical energy can be converted into one another, pointing out that they can be converted into other things involving electricity, chemistry, etc., as well, and proposing a universal energy conservation law. (Similar ideas were expounded five years earlier by J. R. Meyer (1814-1878), but Helmholtz was more detailed, specific, and persuasive.)

The second law of thermodynamics is more subtle and is concerned with certain limits on the possibility of converting energy in the form of heat back into mechanical energy. It has its origin in a paper on the efficiency of heat engines published in 1824 by S. Carnot (1796-1832). While Carnot had the necessary key ideas, he could not express them clearly because he did not think in terms of conservation of energy and along with Lavoisier thought of heat as an element. Carnot's paper was forgotten for a quarter of a century, but Kelvin called attention to it in 1848, and by 1850 Kelvin and Clausius (1822-1888) had combined its ideas with those of Joule and Helmholtz to formulate the second law as we understand it today.

In understanding this law it is useful to realize that it has two aspects. On the one hand it is a general principle which is a bit awkward to formulate in terms that are both sufficiently general and sufficiently precise. On the other hand it is a collection of exact laws about the properties of matter that can be deduced as consequences of this general principle. Consider for example unit mass of some gas (not assumed to be "perfect" but assumed incapable of chemical change) enclosed in a container of variable volume V. It follows from the laws of continuum mechanics (see section 7) that at each temperature T there is a function $V \to p(V, T)$ characteristic of the gas in question, and giving the pressure as a function of the variable volume. In addition it follows from the first law that there is a function U of V and T defined up to an additive constant by the formulae $C_v = \frac{1}{\lambda}\frac{\partial U}{\partial T}$ and C_p

$= \frac{1}{\lambda}\left[\frac{\partial U}{\partial T} - \left(p + \frac{\partial U}{\partial v}\right)\left(\frac{\partial p}{\partial T} / \frac{\partial p}{\partial v}\right)\right]$ which is called the *internal energy function.* Here C_v and C_p are the specific heats at constant volume and constant pressure respectively, and λ is the conversion factor from heat units to mechanical energy units. Until one takes the second law into account there are no restrictions on U and p. There is no *a priori* reason why there should not exist a gas having two more or less arbitrary functions as U and p. However, for every known gas there is a relationship between U and p which holds with great exactitude. This relationship consists in satisfying the partial differential equations equivalent to the existence of a function θ such that $\frac{dU + pdV}{\theta(T)}$ is an exact differential. The function θ is uniquely determined up to a multiplicative constant and is the same for all gases. One can redefine temperature so that $\theta(T) \equiv T$ and so obtain the "absolute" temperature scale introduced by Kelvin. The connection of this very useful fact about gases to the principle known as the second law of thermodynamics is as follows: If one could discover a single gas for which the relationship did not hold, one could use this gas in a manner described by Carnot to take heat energy from a lower to a higher temperature without at the same time turning any mechanical energy into heat. One could thereby turn heat energy back into mechanical energy without restriction and so construct a "perpetual motion machine of the second kind." The various formulations of the second law are equivalent ways of asserting the impossibility of such a machine.

In the case of our chemically stable gas, the exactness of $\frac{dU + pdV}{T}$ (where T is now the absolute temperature) implies the existence of a function S of V and T such that $dU + pdV = TdS$. S is determined up to an additive constant and is called the *entropy* function for the gas. Let $F = U - TS$. Then F is uniquely determined by the gas up to an arbitrary linear function of T. Moreover, straightforward calculations show that the functions U and p can be computed from F by the formulae $p = -\frac{\partial F}{\partial V}$, $U = F - T(\frac{\partial F}{\partial T})$. Thus it suffices to know the single function F to know both U and p and hence all mechanical and thermal properties of the gas. Its value at any V and T is called the *free energy* of the gas in the state defined by V and T.

Exactly the same ideas can be applied to liquids and solids and more importantly to relationships between different substances or substances in different states of aggregation; as in chemical reactions, evaporation, and melting. In every case the impossibility of a perpetual motion of the second kind can be shown to imply exact quantitative relationships between different experimentally measurable quantities and thus to cut down enormously

on the amount of experimental work that has to be done. The pioneer in applying thermodynamics to chemistry was J. Willard Gibbs (1839-1903).

Throughout most of the nineteenth century it was a moot question whether matter is best conceived as a continuum or as built out of discrete point-like entities called atoms. In 1803 John Dalton (1766-1844) had shown how the laws of definite proportions and multiple proportions in chemistry could be explained in terms of atoms and molecules and introduced the concepts of atomic weight and molecular weight. While Dalton's view soon became the accepted one in chemistry, and the experiments of Faraday on electrolysis in the early 1830s suggested that there were also atoms of electricity, many scientists continued to doubt the real existence of atoms of any kind. They preferred to think of them as fictional entities useful in organizing the facts of chemistry, but otherwise not to be taken seriously. It was difficult to decide the matter experimentally because the laws of continuum mechanics were the same whether one postulated a continuum or a sufficiently fine particle structure. Indeed, a popular method for deriving the partial differential equations of motion of a continuum was to begin with an atomic model and pass to the limit as the atoms became lighter and more numerous.

The position changed radically, however, in the final decades of the nineteenth century, when serious quantitative efforts were made to explain heat as mechanical energy due to the motions of the atoms and an algorithm was found for computing the free energy function of an arbitrary gas (or other homogeneous matter) from the kinetic and potential energy functions for the mechanical system defined by its constituent atoms. This "statistical mechanics" evolved from work on the kinetic theory of gases published by Maxwell (1831-1879) in 1859 and Boltzmann (1844-1906) in 1868 and was extensively developed by Boltzmann and Gibbs. The fundamental result can be derived by several probabilistic arguments, none of which is entirely satisfactory, and can be stated very simply. Let Ω denote the "phase space" of the underlying atomic system, that is, the space of all possible position and momentum coordinates of the constituent atoms, and let H denote the real valued function on Ω giving the total energy of the system in terms of the positions and momenta of its atoms. Then the free energy of the substance as a function of V and T is $-kT \log \int_{\Omega} e^{-H/kT} dq_1 \cdots dq_{3n}\, dp_1 \cdots dp_{3n}$ where the q_j and the p_j are the position and momentum coordinates of the n atoms and k is a universal constant known as Boltzmann's constant. The dependence on V results from the dependence of H on V. One of the interesting elementary consequences of this algorithm is obtained by thinking of a solid as a set of n atoms vibrating about their equilibrium positions in some regular lattice. To the extent that the vibrations are small enough so that the equations of motion may be taken to be linear, one can conclude that the specific heat at constant volume is independent of the temperature

and equal to $3nk$. This implies that the specific heat of unit mass of a substance is inversely proportional to the atomic weight — a fact discovered empirically by Dulong and Petit in 1819. It also implies that in the limit of infinitely many zero-mass atoms, the specific heat per unit mass will be infinite. The continuum limit of statistical mechanics gives absurd results it if can be said to exist at all.

While the result just described could be considered almost as an experimental verification of the existence of atoms, the disbelievers had a way out. It was possible to object to statistical mechanics on at least two grounds. First of all, the arguments leading to the algorithm for computing the free energy from H were far from compelling. Second and even more seriously, the results of this calculation often gave results in pronounced and inexplicable disagreement with experiment. For example, the specific heats of solids are not always independent of the temperature. Indeed, if one goes to low enough temperatures they *never* are. Instead, one finds a monotone increase to a constant asymptotic value and it is only at this high temperature limit that the law of Dulong and Petit applies. Similar difficulties were found with the specific heats of gases having polyatomic molecules. It was as though some degrees of freedom were frozen at low temperatures. Gibbs and Boltzmann fought a losing battle against their critics and became quite discouraged. Unfortunately they both died in the early twentieth century just before being vindicated by Einstein's application of quantum ideas to the theory of specific heats. It is rather ironic that J. J. Thomson's discovery of the atom of negative electricity at the very end of the century was particularly discouraging to Gibbs. It suggested that an atom of matter was a complex object with many degrees of freedom. Since none of these appeared to have any influence on specific heats, the ideas of Gibbs and Boltzmann seemed more inadequate than ever.

The immediate stimulus leading to the theory that was to rehabilitate statistical mechanics and revolutionize physics was an anomaly closely related to the failure of the Dulong and Petit law to hold at low temperatures. With the development of Maxwell's ideas about the identity of light waves with oscillating electric and magnetic fields (see section 9), it became clear that one could think of an electromagnetic field as a generalized oscillating continuum and apply the concepts of dynamics and even thermodynamics to its study. In particular one could try to understand the relationship between heat and light suggested by the fact that hot objects are "red hot" at one temperature and "white hot" at a higher temperature. The precise problem actually considered was that of a perfectly reflecting enclosure with walls at a fixed temperature and electromagnetic radiation being reflected back and forth across it. One could let some of the radiation escape through a small hole, pass it through a spectroscope, and study the distribution of energy across the spectrum. There is a characteristic distribution at each tempera-

ture that is independent of the material of which the enclosure is made. The problem was to explain this distribution theoretically.

Using the linearity of the equations of motion, it is not difficult to show that the equivalent dynamical system is equivalent in turn to a system consisting of a countable infinity of *independent* harmonic oscillators of natural frequencies $\nu_1, \nu_2, \cdots$ where the ν_j are uniquely determined by the shape and size of the container. In a state of the system in which the jth oscillator has energy E_j, the radiation with wave length between λ and $\lambda + \Delta\lambda$ will have an energy $\Sigma' E_j$ where the sum Σ' is extended over all j with ν_j between c/λ and $\frac{c}{\lambda + \Delta\lambda}$ and c is the velocity of light. The problem reduces to computing the E_j as a function of the temperature, and this can be dealt with (as Rayleigh [1842-1919] saw in 1900) by applying the fundamental algorithm of statistical mechanics. In fact, except for having infinitely many oscillators instead of $3n$, the dynamical system has the same structure as the one that arises in computing the specific heat of a solid made up of n atoms. One finds easily that the energy of the jth oscillator is independent of ν_j and equal to kT where k is Boltzmann's constant. Ignoring the fact that $\sum_j E_j$ is predicted to be infinite (classical statistical mechanics and classical continuum mechanics are incompatible as indicated above), one immediately deduces that the amount of energy in the part of the spectrum with wave length between λ and $\lambda + \Delta\lambda$ is just kT times the number of ν_j such that c/ν_j is between λ and $\lambda + \Delta\lambda$. When the container is a rectangular parallelepiped, a simple application of Fourier analysis permits an explicit determination of the possible ν_j and the conclusion that for a sufficiently large volume V there is an approximately continuous distribution with density $\frac{8\pi\nu^2}{c^3} V$. This immediately implies Rayleigh's law stating that the energy density in the spectrum at wave length λ is

$$\frac{8\pi(\frac{c}{\lambda})^2}{c^3} kTV \frac{d}{d\lambda}\left(\frac{-c}{\lambda}\right) = \frac{8\pi kTV}{\lambda^4}.$$

While Rayleigh's law was in gross disagreement[3] with experiment for small λ (experiments did *not* suggest an infinite total energy), agreement was good for large λ. More precisely agreement was good when $T\lambda$ was large. Thus the range of good agreement increased with temperature. Once again statistical mechanics seemed to be valid only at sufficiently high temperatures. A few years earlier Wien (1864-1928), with some theoretical justification, had found that one could fit the data well for small λT with an empirical formula of the form $\frac{A e^{-b/\lambda T}}{\lambda^5}$ where A and b are constants.

Quantum theory began in 1900 when Max Planck (1858-1947) found a startlingly simple but rather mysterious argument leading to a reconciliation of the contradictory formulae of Rayleigh and Wien. If one notices that Wien's formula differs from Rayleigh's only in replacing λT by $be^{-b/\lambda T}$, it is

easy to guess Planck's formula. Indeed $\frac{1}{e^{1/x}-1}$ is very close to $e^{-1/x}$ when x is small, and very close to x when x is large. Planck's formula may be obtained from Rayleigh's by writing $\frac{8\pi kTV}{\lambda^4} = \frac{8\pi k\lambda TV}{\lambda^5}$ and replacing λT by $b/e^{b/\lambda T}-1$ where $b = \frac{hc}{k}$ and h is a new constant of nature introduced by Planck. This is not how Planck obtained his formula, however, and what is really significant and interesting is the physical hypothesis that led him to it.

Consider the fundamental algorithm of statistical mechanics. It may be written in the form $F = -kT \log P$ where $P = \int \cdots \int_{\Omega} e^{-H/kT}\, dq_1 \cdots dq_{3_n}\, dp_1 \cdots dp_{3_n}$ and $P = e^{-F/kT}$ is the so called *partition function*. The formula for P in terms of H may be rewritten as $\int_{-\infty}^{\infty} e^{-x/kT}\, d\beta(x)$ where β is the measure on the real line such that β of any interval I is the $dq_1, \cdots, dp_{3_n}$ measure of $H^{-1}(I)$. When the dynamical system is a harmonic oscillator of frequency ν, one computes easily (since Ω is two-dimensional) that $\beta(I)$ is just $\frac{1}{\nu}$ times the length of I when I lies in $x \geq 0$ and $\beta(-\infty,0) = 0$. Thus $P(T) = \int_0^{\infty} e^{-x/kT}\frac{dx}{\nu} = \frac{kT}{\nu}$. It follows at once from the thermodynamical relationship between energy and free energy that $E(T) = kT^2\frac{P'(T)}{P(T)}$, so that in the case of a harmonic oscillator, $E(T) = kT$ as stated earlier. Suppose now that the measure $\beta = \frac{dx}{\nu}$ is replaced by a discrete approximation to it. Choose an arbitrary number h and let us replace β by a measure β_h concentrated at and having the value h at each of the equally spaced points $0, a, 2a, 3a, \cdots$. An interval with n such points will have β_h measure nh and β measure between $\frac{(n-1)a}{\nu}$ and $\frac{na}{\nu}$. Thus β and β_h will agree asymptotically if and only if $a = h\nu$. Using β_h instead of β, one finds $P_h(T) = \sum_{n=0}^{\infty} e^{-kn\nu/kT} = \frac{h}{1-e^{-h\nu/kT}}$ and $E_h(T) = \frac{h\nu}{e^{k\nu/kT}-1}$. Of course as h tends to zero so that β_h approaches β, $E_h(T)$ has kT as a limit as one would expect. However, if one stops short of zero at just the right value (Planck's constant) one gets a formula for $E_h(T)$ which, when substituted for kT in Rayleigh's derivation, gives the correct distribution law for all T and λ.

This is not of course how Planck proceeded.[4] He couched his argument in more physical terms. In essence, however, the argument was the one I have given. Planck did not intend to stop short at a non-zero value of h. He found it convenient to follow the time-honored custom of beginning his analysis with a discrete approximation and meant to pass to the limit of zero h. It is fortunate for physics that he was alert enough to notice that the right answer fell out before the limit was reached.

As Einstein (1879-1955) pointed out in 1907, the failure of the law of Du-

long and Petit at low temperatures can be understood in exactly the same way. Replacing $\frac{kT}{\nu}$ by $1 - e^{-h\nu/kT}$ as the partition function of a harmonic oscillator of frequency ν, one finds that the specific heat of an n atom solid is not $3nk$ but $\sum_{j=1}^{3n} \frac{d}{dT} \frac{h\nu_j}{e^{h\nu_j/kT} - 1}$. Each term tends to k as T approaches ∞, but its distance from the limiting value k depends upon T/ν_j. Thus the high frequency components will be suppressed at low temperatures. Some degrees of freedom will indeed be "frozen."

In spite of these successes, the physical significance of the argument remained completely obscure. To make any sense at all out of it one had to assume that for some mysterious reason an oscillator of frequency ν could not occupy the whole continuum of energy states previously thought possible. It was restricted to energies of the form $nh\nu$ where n is a non-negative integer. It was to be a quarter of a century before this mysterious quantization was to be "understood" in the sense of being a consequence of the laws of a new mechanics — the subtle refinement of classical mechanics known as quantum mechanics. In the meantime a number of other cases were found in which it was possible to "explain" physical phenomena and derive formulae agreeing with experiment by arbitrarily "quantizing" energy or momentum. Perhaps the most striking examples are Einstein's explanation of the photo-electric effect in 1906 and Bohr's (1885-1962) explanation of the spectrum of the hydrogen atom in 1913. The body of results so obtained between 1900 and 1925 is sometimes referred to as "the old quantum theory."

14. THE LEBESGUE INTEGRAL, INTEGRAL EQUATIONS, AND THE DEVELOPMENT OF REAL AND ABSTRACT ANALYSIS

In 1900 I. Fredholm (1866-1927) announced an interesting new approach to the theory of certain linear integral equations, and in the winter of 1900-1901 this work was reported upon in the seminar of David Hilbert (1862-1943). In 1902 the thesis of H. Lebesgue (1875-1941) appeared. It contained a theory of measure and integration which came to be more or less universally accepted as the appropriate one for most purposes. Both events had their roots in the work of the early nineteenth century on methods for dealing with the partial differential equations of mathematical physics (see sections 8 and 9), and both were of fundamental importance for the future development of analysis.

Lebesgue's thesis was the culmination of a long development which I shall not attempt to trace in any detail. The failure of Fourier series to converge everywhere for continuous functions having insufficient smoothness had led to much debate and soul-searching about the true nature of functions and the best way of defining the definite integral. Well-known early

studies of the question were made by Cauchy in 1823 and Riemann in 1854. The problems involved led Cantor (1845-1918) to his general theory of sets in 1884-85 and to various attempts to measure their "size." The correct way of going about this was perceived by E. Borel (1871-1956) and briefly indicated by him in 1898. Lebesgue developed Borel's ideas about measure and applied them to integration. For a full account, including the contributions of Jordan, Baire, and others, the reader may consult the recent book by Thomas Hawkins [8].

This new flexible and much more general theory of integration had a considerable impact not only on the theory of Fourier series (and integrals), but also on the theory of integral equations and probability theory. One highly unsatisfactory feature of the old theory of Fourier series was the lack of any necessary *and* sufficient conditions for a particular sequence of complex numbers to be the Fourier coefficients of a function in a certain class or vice versa. Already by 1907 this circumstance had been remedied by the beautiful and important Riesz-Fischer theorem. This asserts that the mapping $f \to \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-inx}dx$ sets up a one-to-one correspondence between *all* Lebesgue measurable complex valued functions f on the interval $[0, 2\pi]$ such that $\int_0^{2\pi} |f(x)|^2 dx < \infty$ and *all* sequences $\{c_n\}$ of complex numbers such that $\sum_{n=-\infty}^{\infty} |c_n|^2 < \infty$. In this correspondence $\sum_{n=-\infty}^{\infty} |c_n|^2 = \frac{1}{2\pi}\int_0^{\infty} |f(x)|^2 dx$, and one identifies two functions f when they differ only on a set of measure zero. The interesting part, of course, is that every square summable sequence actually arises as a sequence of Fourier coefficients for some function, and this of course could not be true without including the quite general Lebesgue integrable functions. Actually Riesz and Fischer (who wrote independent papers) proved a considerably more general theorem applying to orthogonal functions in general. Moreover, they made it clear that their result was an easy consequence of the central fact that the set of all square integrable functions is "complete" with respect to "mean convergence." Every sequence which is a Cauchy sequence in the sense of mean convergence has a square integrable limit. The analogue for Fourier integrals was formulated and proved by Plancherel and published in 1910.

A necessary and sufficient condition of equal importance but of a rather different character is due to G. Herglotz (1881-1953). If μ is any finite nonnegative measure on the interval $0 \le x \le 2\pi$ (and one identifies 0 and 2π), then one calls the complex numbers $c_n = \int_0^{2\pi} e^{-inx}d\mu(x)$ the Fourier coefficients of μ. It is trivial to verify that the sequence $n \to c_n$ is *positive definite* in the sense that $\sum_{n,m} c_{n-m}Z_n\overline{Z}_m \ge 0$ for every finite sequence $Z_{-\ell}, Z_{-\ell+1}, \cdots, Z_0, Z_1, \cdots, Z_\ell$, of complex numbers. Conversely, as shown by Herglotz in 1911, every positive definite sequence is the sequence of Fourier coefficients for a

unique positive measure μ. The analogue for Fourier integrals was first stated and proved twenty-one years later by Bochner (1899—).

Fredholm's work on integral equations was stimulated by slightly earlier work of Volterra (1860-1940), and this work in turn can be understood as the result of looking at Neumann's solution of the Dirichlet problem in 1870 (see section 9) from a new point of view. This new point of view is of central importance in modern abstract analysis (functional analysis) and has become so familiar that it is now hard to believe that it was startlingly new in 1887. In that year Volterra began to publish a series of papers systematically developing a theory of functions in which the arguments need not be k-tuples of numbers but may be other functions. In particular, he introduced and stressed the point of view that when one performs an operation on a function which leads to a number (e.g., when one evaluates a direct integral), this is analogous to substituting a number in a formula to get another number. He called such functions "functions of lines" — a term soon replaced by "functional." A closely related step, which became part of the same program, was to think of the operations of analysis such as differentiation as again analogues of functions. In other words, one can have generalized functions in which the ranges as well as the domains are sets of functions. These Volterra called *operations* or *operators*.

Neumann had solved the Dirichlet problem by way of a preliminary reduction to a linear integral equation — for functions on the surface of the region in question. Then in 1896 Volterra studied a class of integral equations that could be solved by the same method. Moreover, he interpreted and clarified the method using his operator point of view. Let $T_K(f)(y) = \int_a^y K(x, y)f(x)dx$ where K is a continuous function of two variables and a is a fixed constant. Then to solve the integral equation

$$f(y) = \varphi(y) - \int_a^y \varphi(x)K(x, y)dx$$

is to solve the operator equation $f = \varphi - T_K(\varphi) = (I - T_K)\varphi$, and one has $(I - T_K)(I + T_K + T_K^2 + \cdots) = I$ where I is the identity. T_K^n is an integral operator defined by a kernel K_n, and $T_K + T_K^2 + \cdots$ makes sense as the integral operator defined by $K_1 + K_2 + \cdots$. Thus $\varphi = (I + T_K + T_K^2 + \cdots)f = f + T_{K_1}(f) + T_{K_2}(f) + \cdots$. Fredholm showed how to use analytic continuation to deal with more general cases in which $I + T_K + T_K^2 + \cdots$ does not converge. The idea is that $I - \lambda T_K$ will fail to have an inverse only at the zeros of a certain entire function δ, and $(I - \lambda T_K)^{-1}\delta(\lambda)$ can be developed as an everywhere convergent power series in λ.

Fredholm's results were quite beautiful and interesting, but there is little doubt that their greatest importance lies in the fact that they inspired Hilbert to spend a decade making an intensive study of linear integral operators. Hilbert's work on the subject was first published in a series of six articles between 1904 and 1910, and then again as a book *Grundzüge einer*

allgemeinen Theorie der Linearen Integralgleichungen, which appeared in 1912. Hilbert concerned himself above all with the existence of eigenfunctions and eigenvalues for his integral operators, that is, with functions f and constants λ such that $T(f) = \lambda f$ for the operators T. Moreover, he took an abstract algebraic approach, thinking of an integral operator as an infinite analogue of a matrix, and introduced the "Hilbert space" of all infinite sequences $c_1, c_2, \cdots$ of complex numbers such that $|c_1|^2 + |c_2|^2 + \cdots < \infty$ as the corresponding analogue of the space of all n-tuples of complex numbers. He made the connection with function spaces via expansions in systems of orthogonal functions and thereby inspired the work of Riesz and Fischer discussed above.

In the purely algebraic case, an $n \times n$ matrix a_{ij} with $a_{ij} = \overline{a}_{ji}$ was known to be "diagonalizable" in the sense that there exists a basis of mutally orthogonal eigenvectors for the operator defined by the matrix. Hilbert set out to generalize this result to infinite matrices in Hilbert spaces working with the quadratic form $\sum a_{ij}x_i\overline{x}_j$ instead of the matrix itself. Assuming $a_{ij} = \overline{a}_{ji}$ and a strong continuity property called complete continuity, he was able to prove the obvious analogue of the algebraic result. Since many integral operators can be shown to be completely continuous, Hilbert's theorem had wide applicability to the differential operators that occur in mathematical physics. On the other hand, complete continuity is too strong a requirement for many purposes, and Hilbert's most striking achievement in this area is his formulation and proof (in the special case of a bounded self-adjoint operator) of what is today called the spectral theorem. From Hilbert's point of view, he found the analogue for a bounded quadratic form in infinitely many variables of the classical theorem asserting that a linear change of coordinates reduces every quadratic form to a linear combination of squares. The difficulty produced by dropping the hypothesis of complete continuity is that sums have to be replaced by integrals. Anticipating later concepts one can say that, in effect, Hilbert showed that every bounded self-adjoint operator in Hilbert space can be decomposed as a "continuous direct sum" or "direct integral" of constant operators. The replacement of sums by integrals makes it difficult to attach a meaning to such concepts as the "multiplicity of occurrence" of an eigenvector and so to discuss "equivalence" of operators or forms. This difficulty can be met, and a preliminary version of the resulting "spectral multiplicity theory" appeared in the 1907 thesis of Hilbert's student Hellinger (1883-1950). Hellinger published an improved version in 1909, and Hahn (1879-1934) showed in 1911 that further improvements and simplifications could be made by making systematic use of the theory of the Lebesgue integral. One usually speaks today of the Hahn-Hellinger theory. The modern theory of unitary group representations, which contains classical harmonic analysis as a very special case, may be regarded as a sort of blend of the spectral theorem combined with the

Hahn-Hellinger theory on the one hand and the group representation theory of Frobenius and Schur (see section 15) on the other. Thus this work of Hilbert, Hellinger, and Hahn is of the greatest importance for our main theme. I shall formulate their results more precisely in section 16.

The work of Hilbert and Lebesgue described above led naturally to a very fruitful approach closely allied to the operator point of view advocated and developed by Volterra twenty years earlier. This approach consists in looking at sets of functions as infinite-dimensional analogues of Euclidean space with individual functions as "points." One can then think of convergence of functions in geometrical terms and apply one's geometric intuition to get new insights. The concept of an abstract metric space was introduced by Fréchet (1878-1973) in 1906. The following year Fréchet and E. Schmidt introduced geometric language into the study of Hilbert's sequence space, and Fréchet and F. Riesz (1880-1956) observed that the same language could be applied to the space of square summable functions. Indeed, the Riesz-Fischer theorem implied that the two spaces were isomorphic. Between 1907 and 1918 this geometric point of view was applied to a number of concrete function spaces different from Hilbert space by several mathematicians, among whom the undisputed leader was F. Riesz. The modern axiomatic approach was started in 1922 by a paper of S. Banach (1892-1945).

Concurrently with the discovery of the Riesz-Fischer theorem and the rise of the function space concept, another less abstract branch of analysis was developing out of applications of "summability" methods to the study of divergent Fourier series. The main initial stimulus was a paper by Fejér (1880-1957) published in 1904. Let S_n be the sum of the first n terms of the Fourier series of a function f. Fejér proved that $\frac{S_1 + \cdots S_n}{n}$ converges uniformly to f whenever f is continuous, and in 1905 Lebesgue proved that if f is measurable and such that $\int |f(x)|\, dx < \infty$, then $\frac{S_1(x) + \cdots + S_n(x)}{n}$ converges to $f(x)$ for all x except for a set of measure zero. Two years after that, Fatou (1878-1929) proved an analogue of Lebesgue's theorem using "Abel summability" instead of the "Cesarò summability" of Fejér and Lebesgue. In Abel summability one replaces

$$\frac{S_1 + S_2 + \cdots + S_n}{n} = \frac{na_1 + (n-1)a_2 + \cdots + a_n}{n}$$

where $S_n = a_1 + \cdots + a_n$ by $a_1 + ra_2 + \cdots$ where $0 < r < 1$ and calls $\lim_{r \to 1} (a_1 + ra_2 + \cdots)$ the Abel sum if it exists. For Fourier series of the form $\sum_{n=0}^{\infty} c_n e^{in\theta}$, the Abel sum is $\lim_{r \to 1} \sum_{n=0}^{\infty} c_n r^n e^{in\theta} = \lim_{r \to 1} \sum_{n=0}^{\infty} c_n z^n$ where $z = re^{i\theta}$. Thus Fatou's theorem implies that a measurable function on the circle whose absolute value has a finite Lebesgue integral

and whose negative Fourier coefficients are zero is the set of boundary values of a function of a complex variable analytic in $|z| < 1$. A closely related corollary connects harmonic functions in the unit circle with arbitrary measurable, absolutely integrable boundary functions. An important new element was introduced in 1909 and 1910 when Hardy (1877-1947) and Littlewood (1885-1977) began to study what came to be called Tauberian theorems. These are theorems allowing one to deduce the convergence of a series from its summability in various senses when appropriate auxiliary conditions are satisfied. Such a theorem was proved by Tauber in 1897 under rather strong conditions. In the decade or so following 1910, Hardy and Littlewood collaborated in a series of papers that improved Tauber's theorem in a number of non-trivial ways.

These Tauberian theorems proved by Hardy and Littlewood (with contributions from other mathematicians as well) turned out to be of central importance in applying Fourier analysis to number theory in a new way. Let $n \to \varphi(n)$ be a complex-valued function on the integers. The most general character on the additive group of all the integers is $n \to z^n = r^n e^{in\theta}$ where $z \neq 0$ and the "Fourier transform" of φ is $\sum_{n=-\infty}^{\infty} \varphi(n) z^n$. When $|z| = 1$ so that the corresponding character is unitary, this reduces to $\sum_{n=-\infty}^{\infty} \varphi(n) e^{in\theta}$, the function whose Fourier coefficients are the $\varphi(n)$. Of course this latter function will not exist unless $\varphi(n) \to 0$ as $|n| \to \infty$. On the other hand, if $\varphi(n) = 0$ for $n < 0$ and $\varphi(n)$ is not too badly unbounded as $n \to \infty$, then $\sum_{n=-\infty}^{\infty} \varphi(n) z^n = \sum_{n=0}^{\infty} \varphi(n) r^n e^{in\theta}$ will be defined and analytic for all z with $|z| < 1$, and this analytic function may be regarded as an analytic continuation of the "non-existent" function $\sum_{n=0}^{\infty} \varphi(n) e^{in\theta}$. Moreover, when $\sum_{n=0}^{\infty} \varphi(n) e^{int}$ does exist, studying Abel summability and Tauberian theorems amounts to studying the relationship of $z \to \sum_{n=0}^{\infty} \varphi(n) z^n$ to its boundary values. More generally, one can use similar methods to relate the asymptotic behavior of $\sum_{n=0}^{\infty} \varphi(n) z^n$ as $|z| \to 1$ to the behavior of $\varphi(n)$ for large n. For various functions φ of number-theoretical interest, such as the number of partitions of n, one can show by direct arguments that the "dominant" singularities of $z \to \sum_{n=0}^{\infty} \varphi(n) z^n$ as $|z| \to 1$ are at points of the form $e^{(2\pi ip)/q}$ where p and q are integers. Moreover, one can obtain quite precise information about the way in which $\sum_{n=0}^{\infty} \varphi(n) r^n e^{(2\pi ipn)/q}$ approaches ∞ as r tends to 1. Using this information, Hardy and his collaborators were able to apply Cauchy's theorem, Tauberian theorems, and delicate estimates to obtain useful asymptotic formulae for $\varphi(n)$. The method is now known as the circle method. It was

used first by Hardy and Ramanujan (1887-1920) to study the number of partitions of n—the results being published between 1917 and 1919. Slightly later Hardy and Littlewood used it to study Waring's problem and in particular to obtain asymptotic formulae for the number of representations of n as a sum of a fixed number r of integer kth powers. They also showed that the prime number theorem of Hadamard and de la Vallée Poussin (see section 10) is deducible from a Tauberian theorem.

In 1914 Hardy introduced and studied a new class of functions—called H^p functions in his honor. If $p \geq 1$, the class H^p consists of all analytic functions in $|z| < 1$ have the property that $\int |f(re^{i\theta})|^p d\theta$ is bounded as a function of r. The study of this class was continued by F. and M. Riesz and led to a certain blending of function space ideas with those of the more concrete analysts such as Hardy and Littlewood.

15. GROUP REPRESENTATIONS AND THEIR CHARACTERS

In 1881 Weber defined a character of a finite commutative group G to be a complex valued function χ on G such that $\chi(xy) = \chi(x)\chi(y)$ for all x and y in G. This definition was an abstract generalization of one given three years earlier by Dedekind in connection with his work on algebraic number theory, which was inspired in turn by early work of Gauss and Dirichlet (see sections 6 and 12). While Weber's definition makes sense for arbitrary finite groups, it is more or less vacuous except insofar as the group has commutative aspects. Specifically, every character is identically one on the commutator subgroup and consequently the only characters not identically one are derived trivially from characters of commutative quotient groups. Group theory acquired a powerful new tool that was soon to become almost indispensable when G. Frobenius (1849-1917) published a paper in 1896 showing that there is a natural generalization of the character notion that involves the whole group G in a significant and interesting way—even when G is non-commutative. Considering the impact that this generalization was to have on group theory, it is interesting to note that Frobenius and Klein were born in the same year. It is even more interesting that the new definition was more or less directly inspired by Dedekind and the needs of this work on algebraic number theory.

After Dirichlet's work of 1837-1840, the theory of binary quadratic forms was generalized in two different directions. On the one hand, the preliminary work of Legendre and Gauss on ternary quadratic forms developed into a general theory of quadratic forms in n variables in the hands of Eisenstein (1823-1852), Hermite (1822-1905), and H. J. S. Smith (1826-1883). On the other hand, attempts to prove Fermat's "last theorem" and to generalize the quadratic reciprocity law led first Kummer (1810-1893) and then Kronecker (1823-1891) and Dedekind (1831-1916) to develop the theory of

algebraic number fields. The theory of binary quadratic forms is more or less equivalent to the special case of the latter theory in which the field is generated over the rationals by a root of a quadratic equation with integer coefficients. In that case the group of automorphisms of the field (the Galois group) is of order two and one can deal with it without thinking in group theoretic terms. More generally, a theory as complete as that of Gauss and Dirichlet is not available even today (except when the Galois group is commutative). It was in working out aspects of this still incomplete theory (see section 19) that Dedekind was led to the problem that inspired Frobenius to introduce his "higher dimensional characters." Let G be a finite group of order h and let $g_1 \cdots g_h$ be the elements. Let $x_{g_1} \cdots x_{g_h}$ be h independent variables parameterized by the elements of G and let $\theta(x_1 \cdots x_h)$ denote the determinant of the matrix $||x_{g_i g_j^{-1}}||$. Then θ is a polynomial in h variables, which Dedekind called the group determinant. In a letter written to Frobenius in 1896 (and published in Dedekind's collected works), Dedekind states that many years earlier (around 1880) he had been led to study the group determinant through a consideration of the discriminant of an algebraic number field. He had soon discovered the interesting fact that θ factorizes into linear factors parameterized by the characters of G whenever G is commutative, and he had also factorized θ for various special non-commutative groups. But his attempts to generalize his theorem about commutative groups to general non-commutative groups had failed, and one purpose of his letter was to interest Frobenius in the problem (see Thomas Hawkins's article in this volume).

Frobenius's response was prompt and effective. A correspondence ensued, and before the end of the year, Frobenius had published a paper on the theory of his new characters and another applying them to the solution of Dedekind's problem. Each Frobenius character χ has a *degree* equal to its value at the identity element e, and it turns out that the *distinct* irreducible factors of θ are parameterized by the Frobenius characters. Each factor has a degree equal to the degree of the corresponding Frobenius character and occurs with a multiplicity equal to this degree.

Frobenius's original definition of character was a complicated one, which emerged from his analysis of Dedekind's problem. A year later, however, he showed that his definition is equivalent to another that is much simpler and more natural. Let us define an n-dimensional *representation* of the group G to be a homomorphism L of G into the group of all $n \times n$ complex matrices of non-zero determinant. Let us define L to be reducible if a change of basis can be made which throws all matrices L_x simultaneously into the form $\left(\begin{pmatrix} A_x & 0 \\ B_x & C_x \end{pmatrix}\right)$ and let us define L to be irreducible if it is not reducible. For each representation L of G one obtains a complex valued function χ^L on G by setting $\chi^L(x) = \text{Trace } (L_x)$. χ^L is called the *character* of L. The

classical characters of Dedekind and Weber are just the characters of the *one-dimensional* representations of G, and the new characters introduced by Frobenius are just the characters of the other irreducible representations of G. When the representation L is reducible as explained above, it is clear that $x \to A_x$ and $x \to C_x$ are also representations, and that $\chi^L(x) = \chi^A(x) + \chi^C(x)$. It follows that the character χ^L of any representation L is a sum of a finite number of characters of irreducible representations and hence of characters in the new sense introduced by Frobenius. It will be convenient to adopt the following more or less standard terminology. A character in the sense of Dedekind and Weber is a *one-dimensional character.* The character of an irreducible representation is an *irreducible character.* A finite sum of irreducible characters, or equivalently the character of a (possibly reducible) representation, is a *character.* It follows at once from the definition that the characters of two representations are equal whenever one representation can be obtained from the other by a change of basis, and that every character is a constant on the conjugate classes of the group. Since Frobenius was able to prove that any two distinct irreducible characters χ_1 and χ_2 are *orthogonal* in the sense that $\sum_{x \in G} \chi_1(x)\overline{\chi_2(x)} = 0$, it follows that the irreducible characters are linearly independent and hence that there can be only finitely many of them. Indeed, it follows that there can be no more than h where h is the number of conjugate classes. Frobenius proved further that there are exactly h, and succeeded in his first paper in determining all of them for several different non-commutative groups.

Burnside (1852-1927) became interested in Frobenius's new theory almost immediately and began to publish papers on the subject in 1898. He found different proofs of Frobenius's main results and was a pioneer in emphasizing the advantages of taking representations rather than their characters as the basic objects of the theory. Frobenius's student Schur (1875-1941) saw things in this way also. In 1905 he published a systematic account of the whole theory from the representation theory point of view. In the theory of representations a key role is played by a theorem discovered by Maschke in a slightly different context and published by him in 1899. It asserts that every reducible representation is actually *completely reducible* in the sense that the basis may be chosen so that the matrices take the form $\left(\begin{pmatrix} A_x & 0 \\ 0 & C_x \end{pmatrix}\right)$.

The theory of representations takes a more perspicuous form if one avoids a choice of basis and thinks in terms of abstract linear transformations. A representation L of G is then a homomorphism $x \to L_x$ of G into the group of all non-singular linear transformations of some finite-dimensional complex vector space $V(L)$, and two representations L and M are equivalent if there exists a non-singular linear transformation T from $V(L)$ onto $V(M)$ such that $TL_xT^{-1} = M_x$ for all x. L is reducible if there exists a

proper subspace V_1 of $V(L)$ such that $L_x(V_1) = V_1$ for all x. The restriction of the L_x to V_1 defines a *subrepresentation* L^{V_1} whose space is V_1. Maschke's argument shows that for every such subspace V_1 there is another V_2 with $V_1 + V_2 = V(L)$ and $V_1 \cap V_2 = 0$. Thus in an obvious sense L is the "direct sum" of L^{V_1} and L^{V_2}. Iterating this procedure, one shows that every representation L is equivalent to a direct sum $M^1 \oplus M^2 \oplus \cdots \oplus M^k$ where the M^j are irreducible. Moreover, it is not hard to show that this direct sum decomposition is essentially unique in the sense that if also $L \simeq N^1 \oplus N^2 \oplus \cdots \oplus N^\ell$ where the N^j are irreducible, then $\ell = k$ and there exists a permutation π such that M^j and $N^{\pi(j)}$ are equivalent. Thus one knows the most general representation of G to within equivalence when one knows the most general irreducible representation of G to within equivalence. Since it can be shown that two representations are equivalent if and only if their characters are equal, it follows that there are only finitely many equivalence classes of irreducible representations. A representation of particular interest is the so-called regular representation. Its space is the space of all complex-valued functions on G, and one defines the representation by translation: $L_x(f)(y) = f(yx)$. A fundamental theorem asserts that the regular representation is equivalent to a direct sum in which each equivalence class of irreducibles occurs, and occurs with a multiplicity equal to the degree of its character; that is, the dimension of the representation space. It is suggestive to compare this fact with Frobenius's theorem on the factorization of group determinants.

There is an alternative route to the decomposition theory of group representations which takes its origin in the theory of algebras (or hypercomplex number systems as they used to be called). Influenced by the work of Killing cited in section 11, Molien (1861-1941) published two papers in 1893 giving the first deep and general theorems about the structure of associative algebras over the complex field. He introduced the notions of simplicity and semi-simplicity, and more or less proved the celebrated Wedderburn structure theorems in the complex case. (Similar results were obtained independently but slightly later by Cartan.) Then in 1897 (and apparently without knowledge of Frobenius's work) Molien applied his ideas to a particular algebra that one can associate with a finite group—the so-called group algebra—and obtained a number of Frobenius's more important results. For more particulars about the relationship between the work of Frobenius, Molien, and Burnside, as well as a detailed analysis of the correspondence between Dedekind and Frobenius and how Frobenius was led to invent characters, the reader is referred to three excellent articles by Thomas Hawkins [*9*, *10*, and *11*]. I am indebted to these articles for many of the historical facts stated in this section.

In addition to solving Dedekind's problems and opening the door to a far-reaching extension of the method of harmonic analysis, Frobenius's dis-

covery of non-one-dimensional irreducible characters provided group theory itself both with a powerful new tool and with a fascinating and difficult new problem. In 1900 Burnside published two papers using characters to prove new theorems about the structure of finite groups, and shortly thereafter Frobenius did likewise. Then in 1904 Burnside used characters to prove the very striking theorem that any group whose order is of the form $p^\alpha q^\beta$ where p and q are primes is necessarily solvable. For well over half a century no proof not using characters was known, and even today the character proof is by far the simplest. The new problem is that of actually finding the irreducible representations and their characters for particular finite groups. This problem is easily solved in some cases but quite difficult and challenging in others. In the special case of the so-called "Chevalley groups" (analogues of semi-simple Lie groups in which the real and complex fields are replaced by finite fields), study of this problem is a field of research of considerable current interest and one in which important progress has recently been made.

In studying the problem of finding the irreducible representations and characters of a finite group G, it is useful to consider the relationship of the representations of G to those of its various subgroups. Already by 1898 Frobenius had published a paper on the subject. In this paper he introduced the very important concept of an *induced character*. Let χ be an arbitrary character of the subgroup H of the group G and let χ^0 be the function on G which agrees with χ on H and is otherwise zero. Then define χ^* on G by the formula $\chi^*(x) = \frac{1}{o(H)} \sum_{y \in G} \chi^0(yxy^{-1})$. Frobenius proved that χ^* is always a character and called it the character of G induced by the character χ of H. While χ^* need not be irreducible when χ is, it is irreducible in many cases, and inducing is one of the most important ways of constructing non-one-dimensional irreducible characters. For nilpotent groups *every* irreducible character which is not one-dimensional is induced by a one-dimensional character of a suitable subgroup, and for many non-commutative groups a significant fraction of their irreducible characters may be so obtained. Further insight into the nature of the inducing process may be obtained by considering the celebrated Frobenius reciprocity theorem. Let χ_1 and χ_2 be irreducible characters of G and a subgroup H respectively. Then χ_2^* expressed in terms of the irreducible characters of G contains χ_1 exactly as many times as the restriction of χ_1 to H contains χ_2. One verifies easily that the character of the regular representation of G is equal to the character induced by the unique irreducible character of the subgroup $\{e\}$ consisting of the identity alone. The Frobenius reciprocity theorem applied to this case yields at once the facts about the structure of the regular representation stated earlier. Another important elementary fact relating characters of subgroups to characters of groups is concerned with product groups. If $G = G_1 \times G_2$ and

χ_1 and χ_2 are characters of G_1 and G_2 respectively, then $x, y \to \chi_1(x)\chi_2(y)$ is a character of G which is irreducible if and only if χ_1 and χ_2 are both irreducible. Moreover, every irreducible character of G can be so obtained by composition from characters of G_1 and G_2 respectively.

That characters and group representations might have something to do with Fourier analysis seems to have first been recognized by Hermann Weyl (1885-1955) in 1927. But an essential first step was taken by Schur in 1924. Because of connections with the branch of algebraic geometry known as "invariant theory," Schur became interested in studying representations of the rotation group in n dimensions and discovered that he could carry over the main features of the character and representation theory of finite groups if he replaced summation over the elements of a finite group by a suitable integration over the compact manifold constituted by the elements of the rotation group. Hurwitz had made use of such an integration in 1897 in a method he discovered for constructing invariants. Schur adapted Hurwitz's integral to his needs. From a modern point of view, Schur and Hurwitz made use of the fact (proved by A. Haar in 1933) that every separable locally-compact group admits a measure (unique up to a multiplicative constant) that is defined on all Borel sets, is finite on compact sets, is invariant under right translation, and is not identically zero. When the group is a Lie group, the existence of this measure can be established easily using concepts from differential geometry. Using integration with respect to "Haar measure" to replace sums over the group elements, Schur was able to carry over Maschke's argument and prove the decomposability into irreducibles of an arbitrary representation of the rotation group. He was able to show also that an irreducible representation is determined to within equivalence by its character and to find all irreducible representations together with their characters for the groups with which he concerned himself.

Weyl had been informed of Schur's results in advance of publication. In the same year he published excerpts of a letter to Schur explaining how his results could be generalized to arbitrary semi-simple Lie groups by making use of work of Cartan that had appeared eleven years earlier in 1913. Cartan, working with Lie algebras, had solved the analogous infinitesimal problem. Weyl worked out his ideas in detail and published them in three papers which appeared in 1925 and 1926. Then in 1927 Weyl took the crucial step toward relating characters and representations to Fourier analysis.[5] In that year, in collaboration with his student Peter, he published a proof of the celebrated Peter-Weyl theorem. If L is an irreducible representation of any group, then the vector space spanned by the matrix elements with respect to a basis is independent of the basis chosen and is invariant under right and left translations. Moreover, the vector spaces so obtained from inequivalent representations are mutally orthogonal with respect to "Haar measure" when the group is a compact Lie group. In one formulation the

Peter-Weyl theorem asserts that for every compact Lie group, the linear span of these finite-dimensional orthogonal subspaces is uniformly dense in the space of continuous functions on the group. In another it asserts that one obtains a complete system of orthogonal functions for the group by choosing an orthogonal basis in each subspace. Specialized to the case of the one-dimensional torus group, the Peter-Weyl theorem is just the completeness of the functions $x \to e^{inx}$ with its implications for expansibility in Fourier series. The proof of the Peter-Weyl theorem is closely related to the proof (see section 14) that a completely continuous self-adjoint operator has a basis of eigenvectors. In fact, it is possible to deduce the Peter-Weyl theorem from the latter result. More generally, let the compact Lie group G act on the Riemannian space S so as to preserve the underlying metric and let it act transitively in the sense that for each s_1 and s_2 in S there exists an x in G with $(s_1)x = s_2$. For each x in G and each complex-valued function f on S, let $V_x(f)$ denote the function $s \to f((s)x)$. Then each V_x is a linear transformation and $x \to V_x$ defines a representation of G in any finite-dimensional vector space M of complex-valued functions on S which happens to be mapped onto itself by all V_x. Let us call this subspace irreducible if the corresponding representation is irreducible. In 1929 Cartan published a paper (admittedly inspired by the paper of Peter and Weyl) proving that M_1 and M_2 are orthogonal whenever the corresponding irreducible representations of G are inequivalent, and moreover that there exists a complete orthogonal set of continuous functions on S whose members belong to irreducible invariant subspaces M. The Peter-Weyl theorem is the special case in which $S = G$ and the action is by group multiplication. In the special case in which S is "symmetric," Cartan proved in addition that any two invariant irreducible subspaces which are not identical define inequivalent representations of G. In the subspecial case in which S is the surface of a sphere in three spaces and G is the rotation group, Cartan's result implies the completeness of the surface harmonics of Legendre and Laplace (see section 7). The connection between group representations and surface harmonics was recognized by Weyl in his book on group representation and quantum mechanics published in 1928.

16. GROUP REPRESENTATIONS IN HILBERT SPACE AND THE DISCOVERY OF QUANTUM MECHANICS

The extension of the theory of group representations and characters from finite groups to compact Lie groups does not produce many significant changes. One of the few is that the number of irreducible characters is countably infinite rather than finite. This implies of course that no representation can behave like the regular representation of a finite group in containing a representative of every equivalence class of irreducibles unless one per-

mits infinite-dimensional representations in some sense. Nowadays, nothing seems more natural than to define the regular representation of a compact Lie group G to be a representation L whose space is the Hilbert space of all complex-valued functions on G which are square integrable with respect to Haar measure and where $L_x(f)(y) = f(yx)$. If one does so (and suitably generalizes the direct sum notion to apply to infinite sums) one finds that the Peter-Weyl theorem implies a very straightforward generalization of the structure theorem for the regular representation of a finite group. This generalization states that the regular representation of a compact Lie group is a direct sum of finite-dimensional irreducible subrepresentations and that each equivalence class occurs with a multiplicity equal to its dimension. Similarly, Cartan's theorem about functions on S can be stated in terms of the decomposition of the representation L of G in $\mathcal{L}^2(S, \mu)$ defined by setting $L_x(f)(s) = f((s)x)$. The Lebesgue integral and Hilbert spaces of square integrable functions were still strange and unfamiliar objects to most mathematicians in the 1920s, however, and a systematic theory of group representations in an infinite-dimensional Hilbert space was slow to develop. When it did, this development was directly inspired by the discovery of quantum mechanics in the period between 1924 and 1927, especially by von Neumann's success in putting this theory into a rigorous, mathematically coherent form based on the theory of operators in Hilbert space.

The "old quantum theory" initiated by Planck's paper of 1900 was replaced by the much more satisfactory quantum mechanics during a period of about three years beginning at the end of 1924. In late 1924 and early 1925, Heisenberg (1901-1976) and Schrödinger (1887-1961) respectively published two apparently very different methods for deducing the spectrum of the hydrogen atom without imposing arbitrary quantization rules. These methods were later shown to be equivalent. More importantly, they turned out to provide the key to the puzzle. After a few years of intensive activity, difficult to trace in detail, physicists were in possession of a subtle refinement of classical mechanics which had classical mechanics as a limiting case and from which the quantum rules of the old quantum theory followed in a logical and consistent manner. Besides Heisenberg and Schrödinger, the chief architects of this new quantum mechanics were Born (1882-1970), Jordan (1902—), and Dirac (1902—).

The key idea in the finished theory is that one must give up the naïve notion that the state of a physical system at a given time can be described by the positions and velocities of its particles. Measurements interfere with one another in a manner that becomes more pronounced the lighter the particles are, and in the case of electrons it is very pronounced indeed. It turns out, however, that it makes sense to assign simultaneous probability distributions to all positions, velocities, and various functions of these and that such a collection of probability distributions may be considered to be a state

of the system. Indeed, the laws of quantum mechanics permit one to calculate all probability distributions at time t_1 when they are known at time $t_2 < t_1$.

The physicists' formulations of the laws permitting one to make these calculations and draw various conclusions from them were somewhat vague and unsatisfactory from the standpoint of a pure mathematician, and von Neumann (1903-1956) became interested in clarifying them. He succeeded admirably, and in 1927 published a remarkable paper showing how a subtle generalization of Hilbert's spectral theorem was the key to the whole question. Altering slightly von Neumann's definition and terminology, let us define a *projection-valued measure* on the real line to be a mapping $E \to P_E$ assigning a projection operator P_E in a separable Hilbert space $\mathscr{H}(P)$ to each Borel set E in the line R in such a fashion that 1) $P_{E\cap F} = P_E P_F$ for all E and F, 2) $P_\emptyset = 0$ and $P_R = I$ where $\emptyset$ is the empty set and I is the identity operator, and 3) $P_{\cup E_j} = \sum P_{E_j}$ whenever the E_j are pairwise disjoint. Let us say that the projection-valued measure P has *bounded support* if $P_{[a,b]} = I$ for some finite interval $[a,b]$ and *countable support* if there exists a countable subset Λ of R such that $P_\Lambda = I$. Given a vector φ in the Hilbert space $\mathscr{H}(P)$ it is easy to check that the function $E \to (P_E(\varphi) \cdot \varphi)$ is a measure on the line which is finite when P has bounded support. When this is the case, one can form the integral $\int_{-\infty}^{\infty} x d(P_x(\varphi) \cdot \varphi)$, and it is not difficult to show that there exists a unique bounded linear operator A such that $(A(\varphi) \cdot \varphi) = \int_{-\infty}^{\infty} x d(P_x(\varphi) \cdot \varphi)$ for all vectors φ. Moreover, the bounded linear operator A is self-adjoint in the sense that $(A(\varphi) \cdot \psi) = (\varphi \cdot A(\psi))$ for all φ and ψ in $\mathscr{H}(P)$. In terms of projection-valued measures, Hilbert's spectral theorem can now be stated very simply. It is just a converse of the result just stated. For every bounded self-adjoint operator A there exists a unique projection-valued measure P^A defined on the real line and having bounded support such that $(A(\varphi) \cdot \varphi) = \int_{-\infty}^{\infty} x d(P_x(\varphi) \cdot \varphi)$ for all φ in $\mathscr{H}(P)$. To understand the one-to-one correspondence between bounded self-adjoint operators and projection-valued measures on R with bounded support set up by the spectral theorem, it is useful to consider the special case in which A has a basis of eigenvectors φ_j, $A(\varphi_j) = \lambda_j \varphi_j$. In that case, P has the set $\bigcup_{j=1}^{\infty} \{\lambda_j\}$ of eigenvalues as countable support. $P_E = \sum_{\lambda_j \in E} P_{\{\lambda_j\}}$ and $P_{\{\lambda_j\}}$ is the projection on the vector space of all φ with $A(\varphi) = \lambda_j \varphi$.

While the spectral theorem was in a sense just what was needed, it did not go far enough. Von Neumann's first task was to extend it so that all projection-valued measures were involved and not just those of bounded support. The problem was to find a corresponding extension of the class of bounded self-adjoint operators. Extending an idea of E. Schmidt, von Neumann

found the answer in a certain class of operators that are not necessarily bounded and are defined on a dense subspace $\mathscr{D}$ of $\mathscr{H}(P)$. Given such an operator A, let A^* be defined as follows: $A^*(\psi)$ is defined if and only if $(A(\varphi) \cdot \psi)$ is continuous as a function of φ and then is the unique vector θ in $\mathscr{H}(P)$ such that $(A(\varphi) \cdot \psi) \equiv (\varphi \cdot \theta)$. One says that A is self-adjoint if A^* is defined on $\mathscr{D}$ and only on $\mathscr{D}$, and there agrees with A. With the concept of self-adjointness extended in this way to possibly unbounded operators, it became possible to extend the spectral theorem so that it set up a one-to-one correspondence between all self-adjoint operators on the one hand and all projection-valued measures on R on the other.

With the spectral theorem so extended, von Neumann was now able to unify the various probabilistic statements of the physicists in one simple general principle. This may be stated as follows: To each physical system there corresponds a complex Hilbert space (usually separable) whose one-dimensional subspaces define the states of the system. To every observable (position coordinates, momentum coordinates, etc.) there corresponds a self-adjoint operator. Given the self-adjoint operator A corresponding to a particular observable, let P^A denote the associated projection-valued measure and let φ denote any unit vector. Then $E \to (P^A_E(\varphi) \cdot \varphi)$ is a probability measure on the real line which depends only on the one-dimensional subspace to which φ belongs. This is the probability distribution assigned to the observable corresponding to A by the state corresponding to the one-dimensional subspace of complex multiples of φ.

Notice that in every state the observable corresponding to A takes a value outside of the spectrum of A with probability zero. When the spectrum of A is discrete or partially discrete, the values taken on by the observable are correspondingly restricted. This is the source of the mysterious "quantization rules" of the old quantum theory and explains why some observables are quantized and others are not. Similarly, the impossibility of finding states in which two different observables have highly concentrated probability distributions at the same time may be traced to the lack of commutativity of the corresponding self-adjoint operators. In fact, the celebrated Heisenberg uncertainty principle may be formulated as an inequality on the product of the dispersions of two probability distributions where the lower bound involves the commutator of the corresponding operators.

In the special case of a system of n "spinless," "distinguishable" particles it is possible to make these abstract statements more concrete. Let the masses of the n particles be $m_1, m_2, \cdots, m_n$ and let their coordinates $x_1\, y_1\, z_1 \cdots x_n\, y_n\, z_n$ be relabeled as $q_1, q_2, \cdots, q_{3n}$. Let $p_1, \cdots, p_{3n}$ be a relabeling of $m_1(dx_1/dt), m_1(dy_1/dt), \cdots, m_n(dz_n/dt)$. Then the Hilbert space of the system may be taken to be the space of all complex-valued functions on Euclidean $3n$ space which are square integrable with respect to Lebesgue measure, and when this is done, the operators Q_j and P_j corresponding to

the q_j and p_j are the following: $Q_j(\varphi)$ is defined whenever $q_j\varphi$ and φ are both square integrable and $Q_j(\varphi)(q_1 \cdots q_{3n}) = q_j\varphi(q_1 \cdots q_{3n})$. $P_j(\varphi)$ is defined whenever φ is absolutely continuous in the jth variable and φ and $\frac{\partial\varphi}{\partial q_j}$ are both square integrable and $P_j(\varphi)(q_1 \cdots q_{3n}) = \frac{h}{2\pi i}\frac{\partial\varphi}{\partial q_j}(q_1 \cdots q_{3n})$ where h is Planck's constant (see section 13).

Since the Q_j commute with one another there is no problem in discussing the joint probability distribution of the q_j and it is easy to deduce from von Neumann's general principle that in the state described by the function φ the probability that the coordinates q_j be in the region R of $3n$ space is $\int_R \cdots \int |\varphi(q_1 \cdots q_{3n})|^2 dq_1 \cdots dq_{3n}$. While not every classical observable has a quantum counterpart, there are many that do, and those that do are usually ones whose classical definition may be written in such a form in terms of the q_j and p_j that the same formula applied to the Q_j and P_j makes sense as a self-adjoint operator. For example, the classical angular momentum of a particle about the z axis is $m(x\frac{dy}{dt} - y\frac{dx}{dt}) = q_1p_2 - p_1q_2$ and $Q_1P_2 - P_2Q_1$ is i times a self-adjoint operator which defines the quantum analogue of angular momentum about the z axis. This operator (when its domain is suitably defined) has a discrete spectrum with the integer multiples of $\frac{h}{2\pi}$ as eigenvalues. In 1913, Bohr derived the spectrum of the hydrogen atom from the *assumption* that angular momentum could only take on the values $0, \pm\frac{h}{2\pi}, \pm 2\frac{h}{2\pi}$, etc.

Thus far we have discussed only quantum statics and said nothing about how the state of a system varies with the time. The physicists' answer to this question carries over immediately to von Neumann's formulation. If $t \to \varphi_t$ is a one parameter family of unit vectors describing the state of the system at any time t, then this vector-valued function of t must satisfy a differential equation of the form $\frac{d\varphi_t}{dt} = \frac{2\pi i}{h}H(\varphi_t)$ where H is the self-adjoint operator defining the total energy of the system. Up to possible ambiguities of domain this operator H is well defined whenever the classical expression for the energy is the sum of a potential energy term which is a function V of the q_j alone and the usual kinetic energy term $\sum_{j=1}^{n}\frac{m_j}{2}\left[\left(\frac{dx_j}{dt}\right)^2 + \left(\frac{dy_j}{dt}\right)^2 + \left(\frac{dz_j}{dt}\right)\right]$. One simply rewrites the kinetic energy in terms of the p_j, substitutes $\frac{h}{2\pi i}\frac{\partial}{\partial q_j}$ for each p_j and adds on the operator of multiplication by $V(q_1 \cdots q_{3n})$. Written out in concrete form, $\frac{d\varphi_t}{dt} = \frac{2\pi i}{h}H(\varphi_t)$ becomes the celebrated Schrödinger equation — a linear partial differential equation in $3n+1$ variables. An immediate consequence of Schrödinger's equation is that the ei-

genvectors of the self-adjoint operator H corresponding to the energy play a special role. Indeed, if $H(\varphi) = \lambda\varphi$, then $e^{i\lambda t}\varphi$ satisfies Schrödinger's equation, and since φ and $e^{i\lambda t}\varphi$ define the same state, φ defines a so-called stationary state — a state which does not change with time. Conversely, every stationary state is easily shown to have this form. In other words, the eigenvectors of H are on the one hand the stationary states and on the other the states in which the energy has a definite sharp value. It turns out that when some external perturbation causes an atom to shift from one stationary state to another of lower energy, the energy difference manifests itself as a quantum of electromagnetic radiation of frequency equal to $\frac{\Delta E}{h}$. Thus the problem of predicting the spectral lines emitted by an atom reduces to finding the eigenvalues of the appropriate operator H. (Reduces is perhaps too strong a word. When one investigates the mechanism more closely, one finds that some lines occur with zero intensity and so are not observed.) While Bohr's simple idea of quantizing angular momentum was sufficient to predict the eigenvalues of the H for the hydrogen atom, the old quantum theory was quite incapable of dealing with atoms having more than one electron. Even with quantum mechanics the problem is difficult, because it is far from trivial to find the eigenvalues of the appropriate H. Indeed, one has to resort to approximate methods of various kinds.

Von Neumann did not content himself with clarifying and rigorizing the conceptions of the physicists. In two further papers published in 1927 he made a basic contribution to physics in that he showed how to combine the ideas of quantum mechanics with those of statistical mechanics and produce a "quantum statistical mechanics." This seems difficult if not impossible at first, because classical statistical mechanics is based on the consideration of joint probability distributions for all the dynamical variables; that is, on the consideration of probability measures in the $6n$-dimensional "phase space" of all possible $6n$-tuples of position and momentum coordinates. The Heisenberg uncertainty principle seems to preclude a quantum mechanical analogue of phase space. However, von Neumann found an analogue of probability measures in phase space in what are now called von Neumann density "matrices" (more accurately density operators). We already have seen that a unit vector φ in the underlying Hilbert space assigns a probability measure on the line to each self-adjoint operator and hence to each observable. More generally, let T be a self-adjoint operator with a basis $\varphi_1, \varphi_2, \cdots$ of eigenvectors and let the eigenvalues $\gamma_1, \gamma_2, \cdots$ be non-negative and such that $\gamma_1 + \gamma_2 + \cdots = 1$. Then for each self-adjoint operator A we may consider the set function $E \to \text{Trace}(TP^A_E) = \sum_{j=1}^{\infty} \gamma_j (P^A_E(\varphi_j) \cdot \varphi_j)$ and verify at once that it is an infinite convex combination of the probability measures $E \to (P^A_E(\varphi_j) \cdot \varphi_j)$ and hence a probability measure itself. Here, of course, P^A is

the projection-valued measure corresponding to A by the spectral theorem. The generalized states so defined by non-negative self-adjoint trace operators with Trace 1 are related to the states defined by unit vectors just as probability measures in phase space are related to points in phase space. In each case, it is a question of comparing a convex set with its set of extreme points. To distinguish them one now speaks of *mixed states* and *pure states*. The mixed states (or *mixtures* as von Neumann called them) are von Neumann's substitutes for probability measures in phase space. His substitute for $\int_\Omega e^{-H/kT}d\zeta$ is Trace $(e^{-H/kT})$ where in the second expression H is the self adjoint operator corresponding to the energy observable. Just as $\int_\Omega e^{-H/kT}d\zeta$ may be written in the form $\int_{-\infty}^{\infty} e^{-x/kT}d\beta(x)$ where β is the image of ζ by H (see section 13), so Trace$(e^{-H/kT})$ may be written in the form $\sum_{j=1}^{\infty} e^{-E_j/kT}$ where $E_1, E_2, \cdots$ are the (not necessarily distinct) eigenvalues of H. In the systems to which statistical mechanics applies, H has a pure point spectrum.) This may be rewritten as $\int_{-\infty}^{\infty} e^{-x/kT}d\beta_q(x)$, where β_q is the measure that counts the number of eigenvalues in each set. In other words, as suggested by Planck's discovery, quantum mechanics tells us that in forming the partition function $T \to \int e^{-x/kT}d\beta(x)$, the continuous measure β must be replaced by a discrete measure. In addition, it tells us what discrete measure to choose. Known facts about the relationship of measures on the line to their Laplace transforms and the observation that classical and quantum statistical mechanics agree at high temperatures suggest that β_q and β should be asymptotically equal to constant multiples of one another. This suggestion, formulated as a theorem, turns out to include as special cases various theorems on the asymptotic distribution of eigenvalues proved earlier by Weyl and Courant and later by a number of mathematicians, among whom Titchmarsh (1899-1963) may be especially mentioned.

Group-theoretical ideas were introduced into quantum mechanics in quite different ways in two papers published in 1927 by Weyl and Wigner (1902—) respectively. Part I of Weyl's paper is devoted to a discussion (independent of von Neumann's) of the concept of a mixed state. Parts II and III contain the group theory. Let A be any bounded self-adjoint operator in a Hilbert space $\mathscr{H}$. For all real numbers t, let $U_t = e^{iAt} = I + iAt + \frac{(iAt)^2}{2!} + \frac{(iAt)^3}{3!} \cdots$. The series converges for all t, and it is not hard to verify that each U_t is unitary and that $t \to U_t$ is a unitary representation of the additive group R of the real line. Moreover, this representation is continuous in the sense that $t \to U_t(\varphi)$ is a continuous function from the real line to the Hilbert space for all φ in the Hilbert space. It is not difficult to verify that A is uniquely determined by the representation U so that there is a natural one-

to-one correspondence between bounded self-adjoint operators on the one hand and *certain* continuous unitary representations of R on the other. Without actually formulating a theorem, Weyl suggested that von Neumann's extension of the spectral theorem to unbounded operators should make it possible to extend the correspondence just described to one between all continuous unitary representations of R and all self-adjoint operators. At the same time, he pointed out that diagonalizing A, when this is possible, is equivalent to decomposing the group representation as an infinite direct sum and that the spectral theorem correspondingly must be equivalent to some sort of continuous decomposition theorem for U.

These suggestions of Weyl were given a solid mathematical foundation by Stone (1903—). In a short note published in 1930, Stone sketched a proof of the following theorem: Let $t \to U_t$ be an arbitrary continuous unitary representation of R. Then there exists a unique projection-valued measure P on R such that for all φ in the (separable) Hilbert space $\mathcal{H}(U)$, one has $(U_x(\varphi) \cdot \varphi) = \int e^{ixy} d(P_y(\varphi) \cdot \varphi)$ identically in x. Conversely, every projection-valued measure P on R arises in this way from some continuous unitary representation U of R. Combining the one-to-one correspondence between representations and projection-valued measures produced by Stone's theorem with the one between self-adjoint operators and projection-valued measures produced by von Neumann's extension of Hilbert's spectral theorem, one has a natural one-to-one correspondence between representations and self-adjoint operators, which reduces in the case of bounded operators to that defined above. Actually, one can even make sense of the formula e^{iAt} in the general case by using the "operational calculus" implied by the spectral theorem. Stone's note was the third in a series devoted to the theory of operators in Hilbert space. Von Neumann's paper formulating quantum statics did not actually contain a proof of the generalized spectral theorem, and Stone found a different proof which he sketched in an earlier note. Von Neumann's proof appeared in a long and famous paper published in 1929.

In Schrödinger's equation as formulated by the physicists, the domain of the operator H is left vague. This is an important gap in the theory because, as emphasized in Weyl's paper, the time evolution of the system is determined by the representation $t \to e^{(2\pi i/h)Ht}$. This representation is not known until H is precisely defined as a self-adjoint operator with a definite domain. In the language of the theory of partial differential equations, the initial value problem does not have a unique solution unless appropriate boundary conditions are imposed. It is interesting that the same condition on an operator that makes it suitable for assigning probability distributions to states also makes it suitable for defining a dynamics in which the present state uniquely determines all future states.

In his 1927 paper on group theory and quantum mechanics, Weyl also pointed out that unitary representations of the real line are technically easier to deal with than unbounded self-adjoint operators that are not everywhere defined. Moreover, he showed that the unitary representations associated with the position and momentum operators for a set of n particles form a system with very simple and natural group-theoretical properties. Specifically, let $U_t^j = e^{iQ_j t}$ and $V_s^j = e^{iP_j s}$. Then $(U_t^j)(f)(q_1, \cdots, q_{3n}) = e^{itq_j} f(q_1, \cdots, q_{3n})$ and $V_s^j(f)(q_1, \cdots, q_{3n}) = f(q_1, q_2, \cdots, q_{j-1}, q_j + \frac{sh}{2\pi}, q_{j+1}, \cdots, q_{3n})$. From this one computes easily that the following commutation relations are satisfied: $U_{t_1}^j U_{t_2}^k - U_{t_2}^k U_{t_1}^j = V_{s_1}^j V_{s_2}^k - V_{s_2}^k V_{s_1}^j = 0$ for all j, k, s_1, s_2, and $U_t^j V_s^k = e^{-(ist h\delta_k^j)/2\pi} V_s^k U_t^j$ for all j, k, s, and t. Using the first set of relations, one obtains a continuous unitary representation U of the additive group of $3n$-dimensional vector space of all n-tuples of real numbers by setting $U_{t_1, t_2, \cdots, t_{3n}} = U_{t_1}^1 U_{t_2}^2 \cdots U_{t_{3n}}^{3n}$, and another V by setting $V_{s_1, s_2, \cdots, s_{3n}} = V_{s_1}^1 V_{s_2}^2 \cdots V_{s_{3n}}^{3n}$. Moreover, the second set of relations is equivalent to the following simple commutation relation between U and V: $U_{t_1, \cdots, t_{3n}} V_{s_1, s_2, \cdots, s_{3n}} = e^{-(ih/2\pi)(s_1 t_1 + \cdots + s_{3n} t_{3n})} V_{s_1, \cdots, s_{3n}} U_{t_1, \cdots, t_{3n}}$. If we recall that the most general continuous character (of absolute value one) on the group of $3n$-tuples can be written uniquely in the form $s_1, s_2, \cdots, s_{3n} \to e^{-(ih/2\pi)(s_1 t_1 + \cdots + s_{3n} t_{3n})}$, this last relation can be written more simply and suggestively in the form $U_s V_\chi = \chi(s) V_\chi U_s$, where s stands for the general $3n$-tuple $s_1, \cdots, s_{3n}$ and χ for the general character $s = s_1, s_2, \cdots, s_{3n} \to e^{-(ih/2\pi)(s_1 t_1 + \cdots + s_{3n} t_{3n})}$. Let G now denote the additive group of all $3n$-tuples of real numbers and let $\hat{G}$ denote its (isomorphic) group of all continuous characters of absolute value 1 (unitary characters). If we define $W_{s,\chi} = U_s V_\chi$, we "almost" obtain a continuous unitary representation of the commutative product group $G \times \hat{G}$. However, $W_{(s_1, x_1)(s_2, x_2)} = W_{s_1 s_2, x_1 x_2} = U_{s_1 s_2} V_{x_1 x_2} = U_{s_1} U_{s_2} V_{x_1} V_{x_2} = U_{s_1} \chi_1(s_2) V_{x_1} U_{s_2} V_{x_2} = \chi_1(s_2) W_{s_1, x_1} W_{s_2, x_2}$ so that $W_{(s_1, x_1)(s_2, x_2)}$ is equal to $W_{s_1, x_1} W_{s_2, x_2}$ only up to multiplication by a complex number of modulus one. W is what is known as a *projective* or *ray* representation of $G \times \hat{G}$. Such representations for finite groups were studied by Schur beginning in 1904. Weyl pointed out that they occur naturally in quantum mechanics because two vectors in Hilbert space determine the same state when one is a constant multiple of the other. He also pointed out that while an irreducible ordinary representation of a finite commutative group is necessarily one-dimensional, an irreducible projective representation of such a group can be multi-dimensional.

Weyl regarded it as highly significant that the position and momentum operators for a quantum mechanical system of n interacting particles are related in such a simple way to a projective unitary representation of a $6n$-dimensional vector group, which, as he suggested, turned out to be irreducible. In fact, as stated by Stone in the paper cited above and as proved

shortly afterwards by von Neumann, the projective representation W is, to within unitary equivalence, the only irreducible projective unitary representation of $G \times \hat{G}$ having $s_1,\chi_1, s_2,\chi_2 \to \chi_2(s_2)$ as its corresponding multiplier. Some years later, Weyl's views were shown to be essentially correct in that methods related to his made it possible to go a long way toward deducing Schrödinger's equation and the form of the position and momentum operators from plausible assumptions about invariance and symmetry. Of course, the commutation relations satisfied by U and V are just the celebrated Heisenberg commutation relations for the Q_j and P_j in global form. It is perhaps not too far from the truth to assert that the essential idea of Weyl is that one does not have to assume the truth of the Heisenberg commutation rules — rather that they may be deduced from plausible *a priori* symmetry considerations.

Wigner's paper, as already suggested, applied the theory of group representations in a completely different way. In the first place, it was concerned with non-commutative finite groups rather than with continuous commutative groups, and in the second place, it dealt with the technique of finding approximate eigenvalues of the energy operator rather than with foundational questions. Let H be a self-adjoint operator whose eigenvalues are to be found. Suppose that H can be written in the form $H_0 + J$ where H_0 has known eigenvalues and eigenvectors and J is in some sense "small." One can then attempt to estimate the eigenvalues of H as follows: Replace $H_0 + J$ by $H_0 + \epsilon J$ where ϵ is a variable parameter, *assume* that the eigenvalues of $H_0 + \epsilon J$ vary smoothly with ϵ and can in fact be expanded in power series in ϵ, compute the first few terms and set $\epsilon = 1$. Suppose that λ^0 is an eigenvalue of H_0 whose corresponding eigenspace M is finite-dimensional, and let $\psi_1, \psi_2, \cdots, \psi_n$ be an orthonormal basis for M. One cannot expect that the n occurrences of λ^0 as an eigenvalue will all change in the same way as ϵ varies from zero to one. Instead, one must seek n different functions $\lambda_1^0(\epsilon), \cdots, \lambda_n^0(\epsilon)$, each of which reduces to λ^0 when $\epsilon = 0$ and is an eigenvalue of $H_0 + \epsilon J$ for small ϵ. If these n functions can be expanded in powers of ϵ, $\lambda_j^0(\epsilon) = \lambda^0 + \lambda_j^1\epsilon + \lambda_j^2\epsilon^2 + \cdots$, then a simple analysis allows one to conclude that the $\lambda_1^1, \lambda_2^1, \cdots, \lambda_n^1$ are the n eigenvalues of the matrix $((J(\psi_i) \cdot \psi_j))$. This is the fundamental theorem of so called "first order perturbation theory," and in many problems one gets a useful approximation to the eigenvalues of $H_0 + J$ by using only the first two terms of the series and setting $\epsilon = 1$.

Now, whatever one may think about the validity of the assumptions leading to this approximation, finding the eigenvalues of $((J(\psi_i) \cdot \psi_j))$ is a well-defined mathematical problem. Moreover, it is usually rather non-trivial because eigenvalues of high multiplicity are the rule rather than the exception. They occur whenever the underlying mechanical system possesses symmetries. These are reflected in automorphisms of the state space which are implemented by unitary operators which commute with the energy operator.

On the other hand, let U be a unitary representation of a group G in a Hilbert space $\mathscr{H}(U)$ which decomposes as a direct sum of finite-dimensional irreducible representations $L^1, L^2, \cdots$ in finite-dimensional orthogonal invariant subspaces $\mathscr{H}^1, \mathscr{H}^2, \cdots$. Suppose that T is any self-adjoint operator which commutes with all U_x and has a pure point spectrum. It follows from elementary considerations that each eigenspace of T is an invariant subspace for U and hence that the $\mathscr{H}^j$ may be chosen to be inside the eigenspaces. This implies, however, that for every L^j whose dimension is greater than one there will be an eigenvalue whose multiplicity is at least equal to this dimension.

It would take us too far afield to give all the details, so let it suffice to say that the high order matrices $((J(\psi_i) \cdot \psi_j))$ which one finds it necessary to diagonalize turn out to commute with all the operators L_x of some representation $x \to L_x$ of a finite or compact group G. The matrices of this representation are explicitly known with respect to the basis ψ_j, and one can exploit these facts to simplify rather considerably the problem of diagonalizing $((J_{ij})) = ((J(\psi_i) \cdot \psi_j))$. The easiest case is that in which no irreducible constituent of L occurs more than once. In that case, the decomposition of L into irreducibles is necessarily a decomposition of the operator defined by $((J_{ij}))$ as a direct sum of constants. Moreover, one can compute these constants directly from the characters of G, the J_{ij}, and the matrix elements of the L_x without solving any equations of higher degree. More generally, when multiplicities occur in the decomposition of L, the same methods may be used to reduce the problem to diagonalizing matrices of lower dimension. The dimensions that occur are the multiplicities.

In the generality described in the preceding paragraphs, the method emerged gradually in a sequence of papers by Wigner and by Wigner and von Neumann in collaboration, published in 1927 and 1928. In Wigner's first paper, he considered only the symmetric group on n objects. This arises because of the identity of the electrons in an n-electron atom. In an earlier paper he had managed the three-electron case without using the theory of group representations as such. Moreover, he credits von Neumann with having called his attention to the existence and applicability of the latter theory.

Weyl followed up his 1927 paper with a remarkable book published in 1928. Based on a course of lectures Weyl gave at the Eidgenossische Technische Hochschule in Zurich in the winter semester of 1927-28, this book, *Gruppentheorie und Quantenmechanik,* was destined to become one of the great classics of mathematical physics. In addition to a presentation in developed form of the group-theoretical ideas of Wigner, von Neumann, and himself, it contained an astonishingly complete, coherent account of the conceptual structure of quantum mechanics, together with its application to concrete physical problems. Perhaps for pedagogical reasons Weyl had little

to say about von Neumann's use of the spectral theorem to deal with observables whose operators are not discretely decomposable. Thus for a full appreciation of the extent to which the physicists' discoveries could be incorporated into a beautiful and rigorous mathematical model, one had to read Weyl's book in conjunction with von Neumann's presentation of his own ideas in book form. This book, *Mathematische Grundlagen der Quantenmechanik,* appeared in 1932. On the other hand, it must be emphasized that both in physical content and in the extent to which it integrated group representations with physics, the book went far beyond the brief indications which I have given here. In particular, Weyl added considerably to what could be found in the literature of the time, and more than once it has turned out that some "new" idea in physics could be found hidden away in some little-understood part of Weyl's book. A considerably revised second edition appeared in 1930 and an English translation in 1931.

It is difficult to overemphasize the importance for physics (and chemistry) of the discovery of quantum mechanics. It did much more than explain away the inconsistencies between classical mechanics and the mysterious quantum rules of Planck, Einstein, and Bohr. Now it was possible (at least in principle—and subject to certain qualifications to be indicated below) to deduce all the properties of matter from its atomic constitution and Rutherford's hypothesis of 1911 that an atom of atomic number n consists of n "electrons" of charge $-e$ interacting with one another and with a much heavier nucleus of charge ne according to Coulomb's law. Of course in applying Coulomb's law one has to replace classical mechanics by quantum mechanics. Moreover, one has to determine the fundamental charge e and the relevant masses by suitable experiments. After that, the theory (as modified by the discovery of electron "spin" and the Pauli exclusion principle) provides a mathematically well-defined procedure (which may be extremely difficult to carry out in practice) for computing the free energy function, the electric and magnetic properties, etc., of any piece of matter whose atomic constitution is known. In the same sense, the theory allows one to compute the binding energies of all molecules and in other ways to deduce the laws of chemistry from first principles. As stated by Dirac in the introduction to a paper published in 1929 (in volume 123, series A of the Proceedings of the Royal Society of London),

> The general theory of quantum mechanics is now almost complete, the imperfections that still remain being in connection with the exact fitting in of the theory with relativity ideas. These give rise to difficulties only when high speed particles are involved, and are therefore of no importance in the consideration of atomic and molecular structure and ordinary chemical reactions, in which it is, indeed, usually sufficiently accurate if one neglects relativity variation of mass with velocity and assumes only Coulomb forces between the various electrons and atomic nucleii. The underlying physical laws necessary

> for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble.

It was Dirac himself who showed how to reconcile Einstein's light quanta with Maxwell's equations. In a paper published in 1927 he pointed out that Maxwell's equations could be looked upon as the equations of motion of a dynamical system with an infinite number of degrees of freedom. When the standard procedures of quantum mechanics are applied to this dynamical system, one obtains a quantum mechanical system which can be reinterpreted as a system of particles (photons or light quanta). Dirac was perhaps the first physicist to see the new quantum mechanics as a logically coherent system, and his book on the subject is another great classic. While less satisfactory from the standpoint of a pure mathematician than the books of Weyl and von Neumann, it was more acceptable to physicists and had an enormous influence. First published in 1930, it has gone through many editions.

The key to using quantum mechanics to explain the formation of molecules from atoms was found by Heitler (1904—) and London (1900-1954), who applied it to the hydrogen molecule in a paper published in that magic year 1927. Shortly thereafter they showed in independent papers that in dealing with molecules with more than two electrons, one could apply the representation theory of the symmetric group in much the same way that Wigner had done. The fifth chapter of Weyl's book contains an exposition of their ideas as interpreted by him and includes a beautiful group-theoretical explanation of chemical valence.

In 1931 Wigner published a book on the application of the theory of group representations to the analysis of atomic spectra. He went on in the 1930s to publish a series of fundamental and influential papers showing how to apply that same theory to a wide variety of quantum mechanical problems.

17. THE DEVELOPMENT OF THE THEORY OF UNITARY GROUP REPRESENTATIONS BETWEEN 1930 AND 1945

It follows at once from Stone's theorem of 1930 connecting self-adjoint operators, unitary representations of the real line and projection-valued measures on the real line (see section 16), that the work of Hilbert and his students and coworkers on self-adjoint operators in Hilbert space, as well as the later work of Stone and von Neumann on the unbounded case, can be reinterpreted as work on the problem of analyzing the unitary representations of R, the additive group of the real line. The spectral theorem itself is the analogue of the theorem stating that any unitary representation of a compact Lie group is a direct sum of irreducible representations, and the spectral multiplicity theory of Hahn and Hellinger is the analogue of the

theorem stating that two direct sums of irreducible representations are equivalent if and only if the same irreducibles occur with the same multiplicities. Hahn and Hellinger dealt only with bounded self-adjoint operators, but in 1932 Stone published a now classic book on the theory of linear transformations in Hilbert space, which among other things contains the details of his own approach to the spectral theorem. Chapter VII of Stone's book is devoted to an improved exposition of the Hahn-Hellinger theory generalized to the unbounded case. Although it has a reputation for being complicated and difficult, the Hahn-Hellinger theory is actually quite easy to explain. I shall do so below in a more general context.

As matters stood at the end of 1932, one had all the ingredients of a complete theory of the unitary representations of the compact Lie groups, of certain finite groups, and of two non-compact groups. The two non-compact groups were the additive group R of all real numbers and the additive group Z of all integers. In the case of Z, the most general unitary representation is of course $n \to U^n$ where U is an arbitrary unitary operator. Since $H \to (H + i)(H - i)^{-1}$ can be shown to set up a one-to-one correspondence between all self-adjoint operators and all unitary operators, the spectral theorem, etc., for self-adjoint operators takes care of both R and Z. It must be confessed, however, that although the ingredients were there, their consequences for unitary group representations had not yet been spelled out. In particular, although it is an easy consequence of the Peter-Weyl theorem that every unitary representation of a compact Lie group is a discrete direct sum of irreducibles, this was not stated or proved in the literature until 1943. Moreover, the connection of the Hahn-Hellinger spectral multiplicity theory with the problem of classifying unitary representations of commutative groups was not explicitly pointed out until the 1950s.

An important stimulus to developing the theory of unitary group representations for infinite groups in a framework more general than that of compact Lie groups was a remarkable paper by A. Haar (1885-1933) published in 1933 and already mentioned in section 15. The notion of a topological group had been introduced by Schreier (1901-1929) only a few years before and Haar proved that whenever such a group is separable and locally compact, it admits a measure invariant under right translation which is defined on all Borel sets, is finite on compact sets, and is non-zero on open sets. (That such a measure is unique up to multiplication by a multiplicative constant was shown a few years later by von Neumann). Of course a left invariant measure with the same properties must also exist, but the two need not be equal. Groups for which the left and right invariant measures coincide are called unimodular and include the compact groups. Thus every compact separable group admits a unique left and right invariant Haar measure which assigns the measure one to the whole group. As pointed out by Haar, the arguments of Peter and Weyl can be applied to the most general com-

pact separable groups once one has Haar measure to use to replace the Hurwitz integration process.

Another important stimulus was provided in 1934 when L. Pontrjagin (1908—) (inspired by the needs of duality theorems in topology) published a paper extending the known duality between finite abelian groups and their character groups to one between arbitrary discrete countable commutative groups on the one hand and separable compact commutative groups on the other. Let G be an arbitrary countable commutative group, and let us define a character on G to be a function χ from G to the complex numbers of modulus one such that $\chi(xy) = \chi(x)\chi(y)$ for all x and y in G. Then just as in the finite case, the pointwise product of two characters is again a character, and the set $\hat{G}$ of all characters is a group with multiplication as the composition law. Again as in the finite case, for each x in G the mapping $\chi \to \chi(x)$ is a character f_x of $\hat{G}$. If one gives $\hat{G}$ the weakest topology which makes all characters f_x continuous, it is not difficult to show that G becomes a topological group which moreover is compact and separable. Conversely, let A be any compact separable commutative topological group and let $\hat{A}$ denote the group of all *continuous* homomorphisms from A to the complex numbers of modulus one. Then $\hat{A}$ is countable and discrete, and as before each member a of A defines a character f_a of $\hat{A}$ by way of the definition $f_a(\chi) = \chi(a)$. Using Haar's extension of the Peter-Weyl theorem applied to A and $\hat{G}$, Pontrjagin was able to show that the maps $x \to f_x \to$ and $a \to f_a$ from G to $\hat{\hat{G}}$ and from A to $\hat{\hat{A}}$ are one-to-one and onto, and that in the case of A and $\hat{\hat{A}}$ that $x \to f_x$ is an isomorphism of topological groups. Thus every separable compact commutative group arises as the character group or dual of some countable discrete commutative group and conversely. There is a natural one-to-one correspondence between separable compact commutative groups on the one hand and countable discrete commutative groups on the other.

In the following year (1935), E. R. van Kampen (1908-1942) extended Pontrjagin's duality to a more general and more symmetrical one involving arbitrary locally-compact commutative groups. In particular, he was able to dispense with the hypothesis of separability. If G is an arbitrary locally-compact commutative group, one defines $\hat{G}$ just as before as the set of all continuous functions χ from G to the complex numbers of modulus one such that $\chi(xy) = \chi(x)\chi(y)$ for all x and y. $\hat{G}$ becomes a locally-compact topological group if one declares a set $\mathcal{O}$ of characters to be open whenever for each χ_0 in $\mathcal{O}$ there exists a compact subset C of G and $\epsilon > 0$ such that $|\chi(x) - \chi_0(x)| < \epsilon$ for all x in C implies $\chi \in \mathcal{O}$. Just as before, one defines $f_x(\chi) = \chi(x)$, and van Kampen's duality theorem asserts that $x \to f_x$ is simultaneously a homeomorphism and a group isomorphism of G on $\hat{\hat{G}}$. Every locally-compact commutative group is the dual of its dual. In the par-

ticular case in which G is an n-dimensional real vector group, G and $\hat{G}$ are isomorphic and the theorem is obvious.

Among the many properties of this duality relation, the following are particularly elegant and useful: a) If H is a closed subgroup of the locally-compact commutative group G, and $H^{\perp}$ is the subgroup of all χ in $\hat{G}$ such that $\chi(h) = 1$ for all h in H, then $H^{\perp\perp} = H$ and restricting χ to H sets up a one-to-one map of the quotient group $\hat{G}/H^{\perp}$ onto $\hat{H}$. This map is both a homeomorphism and an algebraic isomorphism—an isomorphism of topological groups. In particular, a character of a closed subgroup can always be extended to a character of the whole group. b) Let ψ be a continuous homomorphism from G_1 to G_2 where G_1 and G_2 are both locally compact commutative groups. Then for each χ in $\hat{G}_2$, the function $x \to \chi(\psi(x))$ is a character χ^* on G_1 and it is obvious that $\chi \to \chi^*$ is a homomorphism. It is actually a continuous homomorphism called the dual of ψ, which we may denote by ψ^*. One can prove that $\psi^{**} \equiv \psi$ and that ψ has a dense range if and only if ψ^* is one-to-one. More generally, if N^* is the subgroup on which ψ^* reduces to the identity, then $(N^*)^{\perp}$ is the closure of the range of ψ.

This last fact is the basis of an interesting application of the duality theorems to a topic in the harmonic analysis of functions on the real line. As already noted, the most general continuous character on R is $x \to e^{ixy} = \chi_y(x)$ so that $\hat{R}$ may be identified with R. However, it will be convenient here not to identify R and $\hat{R}$. Let D be $\hat{R}$ made into a discrete group by ignoring the topology, and let ψ be the identity map of D onto $\hat{R}$. Then ψ is a continuous homomorphism which is one-to-one. Hence ψ^* is a continuous homomorphism of $\hat{\hat{R}} = R$ into a dense subgroup of the *compact* character group $\hat{D}$ of D. Moreover, since ψ has all of $\hat{R}$ for its range, ψ^* is one-to-one. Thus the device of making $\hat{R}$ discrete yields a natural imbedding of the real line onto a dense subgroup of a compact group. Evidently the continuous complex-valued functions on the compact group $\hat{D}$ are in natural one-to-one correspondence with the members of a certain subclass of the bounded continuous complex-valued functions on the real line. These turn out to be precisely the so-called "almost periodic" functions introduced in 1924 by Harald Bohr (1887-1951).

Bohr's doctoral thesis of 1910 was on the summability of Dirichlet series, and his early work was primarily concerned with the further study of such series with particular emphasis on the Riemann zeta function and the location of its zeros. In collaboration with Landau (1877-1938), he published a paper in 1914 showing that the non-trivial zeros, if not on the central line, must at least be clustered about it. He made studies of the value distribution of the zeta function and other Dirichlet series on vertical lines and ultimately was led to ask for a characterization of those complex-valued functions on the line which can be represented in the form $t \to \sum_{n=1}^{\infty} a_n e^{-\lambda_n(s+it)}$, where s

is fixed real number and the a_j and λ_j are suitable sequences of complex and real numbers respectively. Writing $c_n = a_n e^{-\lambda_n s}$, this becomes $t \to \sum_{n=1}^{\infty} c_n e^{-i\lambda_n t}$, which reduces to a Fourier series when the λ_n are integer multiples of a fixed real number. More generally, the function $t \to \sum_{n=1}^{\infty} c_n e^{-i\lambda_n t}$ will not be periodic but will be "almost periodic" in the sense of having many "approximate" periods. Bohr gave a precise definition and in three long papers published between 1924 and 1926 presented a detailed theory of this new class of functions. One key theorem states that a bounded continuous function on the real line is almost periodic (in the sense of having enough approximate periods) if and only if it is a uniform limit of functions of the form $c_1 e^{i\lambda_1 x} + \cdots + c_n e^{i\lambda_n x}$ where the λ_j are real. Another asserts that every almost-periodic function f has a well defined "mean value" $M(f)$. Using $M(f)$ instead of $\frac{1}{2A}\int_{-A}^{A} f(x)dx$, one can develop an analogue of Fourier series expansions assigning to each almost-periodic function f the formal series $\Sigma\, c_n e^{i\lambda_n x}$ where $c_n = M(fe^{-i\lambda_n x})$ and $M(fe^{-i\lambda x}) = 0$, for all but countably many λ.

Bohr's theory aroused considerable interest at first, but, as I have already indicated, it was destined to be subsumed under the rubric of the duality theory of locally-compact commutative groups. In fact, the mean value $M(f)$ turns out to be just the Haar integral of the extension of f to the compact completion $\hat{D}$ of the real line and the Fourier series expansions to be reducible to the Peter-Weyl expansion on $\hat{D}$.

In 1927 Bochner showed that Bohr's definition in terms of approximate periods can be reformulated in a way that makes sense for arbitrary groups. A function is almost periodic in Bochner's sense if it is bounded and continuous, and if the set of all functions which can be uniformly approximated by its translates is compact in the topology of uniform convergence. For any locally-compact commutative group G, one has a theory of almost-periodic functions analogous to that of Bohr, and this theory can be reduced to the theory of functions on a compact group by taking the dual or adjoint of the natural map of the discretization of $\hat{G}$ into $\hat{G}$ just as for R. Von Neumann published a more general paper in 1934 developing a theory of almost-periodic functions on an *arbitrary* non-commutative group G. Of course the theory could be vacuous, and in 1935 A. Weil (1906—) showed that whenever G has sufficiently many almost-periodic functions, one can imbed it densely in a compact group K in such a fashion that the theory reduces to the Peter-Weyl theory on K.

The theory of almost-periodic functions on the group Z of all integers under addition turns out to have interesting connections with number theory. Rather than considering the whole of $\hat{Z}$, let Λ be the dense subgroup of $\hat{Z}$ consisting of all elements of finite order. The same argument as before shows that Z has a natural dense imbedding as a subgroup of the *separable*

compact group $\hat{\Lambda}$. Moreover, if Λ_p is the subgroup of Λ consisting of all elements whose order is p^k for some k, then $\hat{\Lambda}$ is isomorphic to the direct product of all the compact groups $\hat{\Lambda}_p$. The restrictions to Z of the continuous functions on $\hat{\Lambda}$ are just the almost-periodic functions on Z whose non-zero Bohr-Fourier coefficients are those corresponding to the characters in Λ ($n \to e^{2\pi irn}$ where r is rational). Moreover, such a restriction φ is multiplicative in the sense that $\varphi(nm) = \varphi(n)\varphi(m)$ for n and m relatively prime if and only if the original function on $\hat{\Lambda} = \prod_p \hat{\Lambda}_p$ is a product of continuous functions on the $\hat{\Lambda}_p$. Since each Λ_p is dense in $\hat{Z}$, there is a natural dense imbedding of Z in each $\hat{\Lambda}_p$, and the multiplication in Z has a unique continuous extension to a multiplication in $\hat{\Lambda}_p$. The ring which $\hat{\Lambda}_p$ thus becomes is isomorphic (algebraically and topologically) to the ring of all p-adic integers and its "field of quotients" is the field of p-adic numbers (see section 20). As will be explained more fully later, the Hardy-Littlewood results on Waring's problem (see section 14) have illuminating interpretations in terms of almost-periodic functions on Z.

The coherent theory of locally-compact groups which are either compact or commutative made possible by the results of Peter and Weyl, Haar, Pontrjagin, and van Kampen published between 1927 and 1935, inspired the publication of two very influential books a few years later. Pontrjagin's *Topological groups* was published in Russian in 1938, and an English translation appeared in 1939. *L'intégration dans les groupes topologiques et ses applications à l'analyse* by André Weil appeared in 1940. Although there is a large overlap, the two books are quite different in emphasis and to some extent in content. The key to the difference is revealed in the extra words in Weil's title. Pontrjagin is concerned above all with structure theorems for locally-compact groups and barely mentions harmonic analysis. Weil emphasizes harmonic analysis and shows in detail how one can define a Fourier transform for suitably restricted functions on any locally-compact commutative group and so include Fourier series and Fourier transforms in one and several variables in one unified theory. Specifically, if μ is a choice of Haar measure in G, then Weil defines the Fourier transform of a μ-integrable function f to be the function $\hat{f}$ on $\hat{G}$ such that $\hat{f}(\chi) = \int \chi(x)f(x)d\mu(x)$ for all χ in $\hat{G}$. When f is both integrable and square integrable, he shows that $\hat{f}$ is both continuous and square integrable and that the arbitrary constant in the Haar measure $\hat{\mu}$ in $\hat{G}$ can be so chosen that $f \to \hat{f}$ preserves the Hilbert space norms: $\int |f(x)|^2 d\mu(x) = \int |\hat{f}(\chi)|^2 d\hat{\mu}(\chi)$. Since the domain and range can be shown to be dense in $\mathcal{L}^2(G, \mu)$ and $\mathcal{L}^2(\hat{G}, \hat{\mu})$ respectively, it follows that $f \to \hat{f}$ has a unique extension to be a norm-preserving map of $\mathcal{L}^2(G, \mu)$ onto $\mathcal{L}^2(\hat{G}, \hat{\mu})$ just as in Plancherel's theorem about the classical Fourier transform. In addition to a proof of the generalized Plancherel theorem, Weil's book contains a statement and proof of a generalized Bochner-Herglotz

theorem, a study of how Fourier transforms turn convolution and multiplication into one another, and a number of theorems about summability and pointwise convergence of generalized Fourier transforms. Weil's book makes it quite clear that the natural domain of classical commutative Fourier analysis is the study of the Fourier transform on general locally-compact commutative groups. The extra generality so provided was to prove of importance in applications of harmonic analysis to both number theory and probability. We have already hinted at one application to number theory, and there were to be many others. It is perhaps worth noting that Chevalley's "idèles," which were to play a central role in these applications, were introduced in 1936 — almost at the same time as the duality theory itself.

Let μ be a finite measure in the dual $\hat{G}$ of the locally-compact commutative group G, and let us define $\hat{\mu}$, the Fourier transform of μ, to be the continuous function on G defined by the equation $\hat{\mu}(x) = \int \chi(x) d\mu(\chi)$. One verifies at once that this function is "positive definite" in the sense that $\Sigma\, c_i \bar{c}_j \hat{\mu}(x_i x_j^{-1}) \geq 0$ for all pairs $x_1, x_2, \cdots, x_n, c_1, c_2, \cdots, c_n$ of n-tuples where the x_j are group elements and the c_j are complex numbers. The generalized Bochner-Herglotz theorem of Weil asserts conversely that given any continuous positive definite function f on G, there exists a unique finite ("Radon") measure μ on $\hat{G}$ such that $\hat{\mu} = f$. We shall see in the next section that this theorem is important in applications to probability theory. It is also interesting in that it turns out to be equivalent to the spectral theorem for unitary representations of locally-compact commutative groups. In 1929 Wintner (1903-1958) published a paper showing that the Herglotz theorem itself implied the spectral theorem for unitary operators (and so for unitary representations of Z, the additive group of the integers). Then in 1933 Bochner and Khinchin (1894—) independently proved that Stone's spectral theorem for unitary representations of the real line can be derived in the same way from Bochner's real line analogue of Herglotz's theorem. The argument can easily be extended to the general case. This was done in 1943 and 1944 in independent papers of Ambrose (1914—), Godement (1921—), and Naimark (1909—). The key point in connecting unitary representations with positive definite functions can be explained very easily: Let $x \to U_x$ be a continuous unitary representation of any topological group, and let φ be any vector in $\mathcal{H}(U)$, the space of U. Then it is trivial to verify that $x \to (U_x(\varphi) \cdot \varphi)$ is a continuous positive definite function on G. Indeed, $\Sigma\, c_i \bar{c}_j (U_{x_i x_j^{-1}}(\varphi) \cdot \varphi) = (\Sigma_j c_j U_{x_j}(\varphi) \cdot \Sigma_j c_j U_{x_j}(\varphi)) \geq 0$.

As one might guess immediately from the case of the real line, the spectral theorem for locally compact commutative groups G asserts the existence of a one-to-one correspondence between continous unitary representations U of G and projection-valued measures P on G such that $(U_x(\varphi) \cdot \varphi) =$

$\int \chi(x)d(P_x(\varphi) \cdot \varphi)$ for all φ in $\mathcal{H}(U)$. It immediately implies the generalized Bochner-Herglotz theorem for all continuous positive definite functions on G which may be put into the form $x \to (U_x(\varphi) \cdot \varphi)$. We need only take $\mu(E) = (P_E(\varphi) \cdot \varphi)$. Conversely, assuming the truth of the Bochner-Herglotz theorem, one can write $(U_x(\varphi) \cdot \varphi) = \int \chi(x)d\mu_\varphi(x)$ and obtain the spectral theorem by studying the dependence of μ_φ on φ. Actually, as observed by Gelfand and Raikov in 1943, there is a very simple argument showing that every continuous positive definite function on a topological group can indeed be thrown into the form $x \to (U_x(\varphi) \cdot \varphi)$. Thus the spectral theorem and the generalized Bochner-Herglotz theorem are equivalent results.

As remarked earlier, the spectral theorem for locally-compact commutative groups is the analogue of the theorem that any unitary representation of a compact group is a direct sum of irreducible representations. When combined with the essential uniqueness of the decomposition, this last theorem tells us that to classify all unitary representations it suffices to classify the irreducible ones. The uniqueness theorem for unitary representations of locally-compact commutative groups has a more subtle formulation and is less easy to prove but is still not very difficult. Moreover, it is essentially contained in the Hahn-Hellinger spectral multiplicity theory for self-adjoint operators mentioned in section 14 and developed between 1907 and 1911. Because of the spectral theorem, the problem of determining all continuous unitary representations of a locally-compact commutative group G to within equivalence is the same problem as determining all projection-valued measures P on $\hat{G}$ to within equivalence. Moreover, the solution of this second problem is independent of the group structure of $\hat{G}$ and depends only on its measure-theoretic structure—more precisely on the system consisting of the set $\hat{G}$ and its σ Boolean algebra of Borel sets. When G (and hence $\hat{G}$) is separable, then $\hat{G}$ is either finite, countable, or isomorphic as a Borel space to the real line. Hence when the solution of the problem is not trivial it may be deduced at once from the solution given by Hahn and Hellinger for projection-valued measures on the line.

I shall present this solution in a modernized form. Let S be a Borel space which is either countable or such that for each finite measure μ on S there exists a Borel set N of measure zero such that $S - N$ is Borel isomorphic to a Borel subset of a separable complete metric space. Such an S is said to be metrically standard. For each finite measure μ defined on all Borel subsets of S, let us define a projection-valued measure P^μ on S whose values are projections in the Hilbert space $\mathcal{L}^2(S, \mu)$ by setting $P^\mu_E(f)(s) = \varphi_E(s)f(s)$, where $\varphi_E(s)$ is 1 for s in E and zero for s not in E. Using the Radon-Nikodym theorem on the existence of "densities" for measures having the same sets of measure zero, it is almost immediate that there exists a unitary map V of $\mathcal{L}^2(S, \mu_1)$ on $\mathcal{L}^2(S, \mu_2)$ such that $VP^{\mu_1}_E V^{-1} = P^{\mu_2}_E$ for all E if and only if μ_1 and

μ_2 have the same sets of measure zero. To get insight into how general projection-valued measures of the form P^μ can be, it is useful to look at the special case in which P is supported by a countable set $\Lambda \subseteq S$. It is easy to see in that case that P is equivalent to some P^μ if and only if $P_{\{\gamma\}}$ has a zero- or one-dimensional range for all $\gamma \in \Lambda$. It is almost trivial that P is uniformly one-dimensional in this sense if and only if P has a commutative commuting algebra; that is, that $P_E T = TP_E$ and $P_E S = SP_E$ for all E implies $ST = TS$. This result suggests that in the general case the P^μ are precisely the P's with commutative commuting algebras. In fact, this theorem is true and moreover not difficult to prove. Let us say that P is *multiplicity free* if its commuting algebra is commutative. Moreover, let us say that P^1 and P^2 are disjoint if $TP^1_E = P^2_E T$ for all E implies that $T = 0$. If we define direct sums of projection-valued measures in the obvious way, then the rest of the Hahn-Hellinger spectral multiplicity theory can be summed up in the following propositions:

1) P^{μ_1} and P^{μ_2} are disjoint if and only if μ_1 and μ_2 are supported by disjoint Borel subsets of S; i.e., are mutually singular measures.

2) If P and P^1 are multiplicity free and $\underbrace{P \oplus P \oplus \cdots}_{n \text{ times}}$ is equivalent to $\underbrace{P^1 \oplus P^1 \oplus \cdots}_{m \text{ times}}$, then $n = m$ and P and P^1 are equivalent. (Here m and n may take on the value ∞.)

3) An arbitrary projection-valued measure on S in a separable Hilbert space may be written in the form $\infty P^\infty \oplus P^1 \oplus 2P^2 \oplus 3P^3 \oplus \cdots$ where some terms may be missing and the P^j are disjoint, multiplicity free, and uniquely determined up to equivalence.

Let μ be a finite measure defined in the dual $\hat{G}$ of the locally compact commutative group G. For each x in G and each f in $\mathscr{L}^2(\hat{G}, \mu)$ let $U^\mu_x(f)(\chi) = \chi(x)f(\chi)$. The reader can easily check that $x \to U^\mu_x$ is a unitary representation of G and that P^μ is the projection-valued measure on G canonically associated to U^μ by the spectral theorem. On the other hand, U^μ has an obvious interpretation as the "direct integral" with respect to μ of irreducible (one-dimensional) representations of G.

The results described so far have little to say about the unitary representations of groups that are neither compact nor commutative. The systematic theory of the unitary representations of such groups began to be developed rather abruptly in 1946, a year after the end of World War II. Three important contributions were made earlier, however, and I shall conclude this section with brief descriptions of these. Every irreducible unitary representation of a commutative group is one-dimensional and every continuous unitary irreducible unitary representation of a compact group is finite-dimensional unitary representations at all—except for the identity. In order to have a sufficiency of irreducible representations, one has to allow them to

be infinite-dimensional. Also in order to have a reasonable theory, one has to define irreducibility in a topological rather than in an algebraic manner. The continuous unitary representation L of G is said to be irreducible if there exist no *closed* invariant subspaces. That there are in some sense "enough" infinite-dimensional irreducible unitary representations was proved by Gelfand and Raikov in 1943, in the same paper in which they proved that every continuous positive definite function is of the form $x \to (U_x(\varphi) \cdot \varphi)$. They did so by using the fact that the positive definite functions defined by irreducible continuous unitary representations are the extreme points in the convex set of all positive definite functions.

For a large (but far from exhaustive) class of locally compact groups, new phenomena appear which have to do with the possible structures of the commuting algebras of unitary representations. When the representation is a discrete direct sum of irreducibles, their commuting algebras are direct sums of algebras, each of which is isomorphic to the algebra of all bounded operators in a Hilbert space. One might hope in general for a direct integral of such algebras, but as first pointed out by Murray (1911—) and von Neumann in 1936, it is possible to have a commuting algebra with a trivial center that is not isomorphic to the algebra of all bounded operators in any Hilbert space. In a series of four papers published between 1936 and 1943, Murray and von Neumann made a detailed study of these strange new generalizations of full matrix algebras. They called them factors. After 1950 it became important to classify unitary representations according to the nature of the factors associated with their commuting algebras.

The third contribution was a now famous paper published by Wigner in 1939, containing an analysis of the possible irreducible unitary representations of the so-called inhomogeneous Lorentz group. This is the group generated by the translations in space-time and the celebrated Lorentz group of special relativity. According to the latter theory (advanced by Einstein in 1905), this group is the true group of symmetries of space-time. Because of the principles enunciated by Weyl and described in section 16, one expects a close relationship between the possible relativistic Schrödinger equations and the irreducible unitary representations of the inhomogeneous Lorentz group. It will be easier to describe Wigner's results in a later section. Here it will suffice to remark that Wigner's paper, while incomplete in certain respects, was the first to obtain a genuine classification of the irreducible unitary representations of a group having no non-trivial finite-dimensional unitary representations. Moreover, completing Wigner's analysis was to be a major stimulus to the systematic development which began in 1946.

18. HARMONIC ANALYSIS IN PROBABILITY; ERGODIC THEORY AND THE GENERALIZED HARMONIC ANALYSIS OF NORBERT WIENER

One of the several major consequences of the introduction of the Lebesgue integral (see section 14) was to make possible a convenient and rigorous mathematical framework in which to discuss the many mathematical problems that arose as probability theory found more and more applications in science, engineering, and human affairs. This framework emerged gradually between 1909, when E. Borel published a paper emphasizing the importance of countable additivity in probabilistic considerations, and 1933, when Kolmogorov's (1903—) book *Grundbegriffe der Wahrscheinlichkeitsrechnung* appeared showing how naturally all the main ideas of probability theory could be formulated in measure-theoretic terms. Important landmarks along the way were the papers of Wiener (1894-1964) on "Brownian motion" which appeared between 1920 and 1923, and a paper of Steinhaus (1887-1972) published in 1923. Among other things, Steinhaus's paper related the peculiar convergence properties of expansions in terms of the so-called "Rademacher functions" to the fact that these functions are not only orthogonal but are "independent" as "random variables."

The measure-theoretic framework for probability may be described very simply. One starts with a suitably restricted measure space Ω, μ such that $\mu(\Omega) = 1$, and thinks of the points ω in Ω as "events" in the "universe" Ω of all possible events. If E is a measurable subset of Ω, then $\mu(E)$ is the probability that the event that actually occurs is in the set E. The space Ω, μ is given and fixed once and for all, and probability theory concerns itself with measurable functions defined on Ω. These are given the suggestive name of "random variables" and are usually (but not necessarily) real valued. If g is a real-valued random variable, then the probability that g takes on a value in the Borel set F of real numbers is just $\mu(g^{-1}(F))$. Thus the behavior of the random variable g taken in isolation is completely determined by the measure α on the real line $F \to \mu(g^{-1}(F))$. This is a *probability measure* in the sense that $\alpha([-\infty, \infty]) = 1$ and is what is called the *distribution* of the random variable g. When it exists, $\int_\Omega g(x)d\mu(x) = \int_{-\infty}^{\infty} x d\alpha(x)$ is called the *expected value* or *expectation* of the random variable g, and if this is denoted by $\bar{g}$, then $\int (g(x) - \bar{g})^2 d\mu(x) = \int_{-\infty}^{\infty} (x - \bar{g})^2 \, d\alpha(x)$ is called the *variance* of g. It is clear that g has an expectation *and* a finite variance if and only if g is in $\mathscr{L}^2(\Omega, \mu)$. Of course, as long as only one random variable is involved, there is little point in introducing the "universe" Ω. Everything can be expressed in terms of the measure α. It is in dealing with relationships between

different random variables that the usefulness of Ω becomes clear — especially when there are infinitely many.

Let g_1 and g_2 be two real-valued random variables and let α_1 and α_2 be the probability measures in the real line R that define their distributions. Then $\omega \to g_1(\omega), g_2(\omega)$ is a measurable function ψ from Ω to $R \times R$ and setting $\beta(F) = \mu(\psi^{-1}(F))$ defines a probability measure β in $R \times R$ which is far from uniquely determined by α_1 and α_2. It is called the joint distribution of the random variables g_1 and g_2. If $g_2 = g_1^2$, then β is supported by the curve $y = x^2$. On the other hand if g_1 and g_2 are so-called *independent* random variables, then β is the product measure $\alpha_1 \times \alpha_2$, i.e., $\beta(F_1 \times F_2) = \alpha_1(F_1)\alpha_2(F_2)$. In fact, this is the definition of independence, and this definition seems to capture quite completely the intuitive notion of what it means for random variables to be "independent." Of course, there are many possibilities intermediate between independence on the one hand and functional dependence ($g_2 = F \circ g_1$) on the other. Random variables may be more or less "correlated." If g_1 and g_2 are in $\mathcal{L}^2(\Omega, \mu)$ and have expectations a_1 and a_2 respectively, then it is easy to verify that $g_1 - a_1$ and $g_2 - a_2$ are orthogonal functions whenever g_1 and g_2 are independent. The converse is not true, but $\int (g_1(x) - a_1)(g_2(x) - a_2)d\mu(x)$ can be used as a rough index of the extent to which g_1 and g_2 are not independent. It is called the *correlation coefficient* of the two random variables g_1 and g_2.

The idea that one could have a mathematical theory of a dependency relation weaker than strict functional dependency was a new and exciting one in the late nineteenth century. It is due to Sir Francis Galton (1822-1911), who became interested in continuous aspects of human inheritance (arm length, height, etc.) more or less at the same time as his exact contemporary Gregor Mendel (1822-1884) was concerning himself with the discrete aspects of inheritance in plants. He introduced the correlation coefficient in the 1870s (after much reflection), but being no mathematician defined it incorrectly. His book *Natural Inheritance,* published in 1889, caught the imagination of a young applied mathematician named Karl Pearson (1857-1936), who corrected Galton's concept of correlation and devoted the rest of his career to founding mathematical statistics.

Apart from modern refinements, the idea of thinking in terms of random variables and their expectations, variances, and higher moments goes back to work of Tchebycheff (1821-1894), as already mentioned in section 3. One of Tchebycheff's achievements was to use these concepts together with a famous inequality that bears his name to give a very simple proof of a generalized form of Bernoulli's "weak law of large numbers." This proof can be particularly elegantly formulated if one uses the modern definition of a random variable. I shall present it here as an example of the convenience of the modern framework. Let $f_1, f_2, \cdots$ be a sequence of independent ran-

dom variables and let each f_j have the expected value a. Let $\varphi_n(\omega) = \frac{f_1(\omega) + f_2(\omega) + \cdots + f_n(\omega)}{n}$ for $n = 1, 2, \cdots$. Then $\varphi_n(\omega) - a$ has zero as expected value and $\int (\varphi_n(\omega) - a)^2 d\mu(\omega) = \int \frac{(f_1(\omega) - a + f_2(\omega) - a + \cdots + f_n(\omega) - a)^2}{n^2} d\omega = \frac{1}{n^2} \sum_{k=1}^{n} \int (f_k(\omega) - a)^2 d\mu(\omega)$. If each f_k has finite variance less than or equal to δ, it follows that $\int (\varphi_n(\omega) - a)^2 d\mu(\omega) \leq \frac{n\delta}{n^2} = \delta/n$. Now let E_ϵ^n be the set in which $|\varphi_n(\omega) - a| \geq \epsilon$. Then $\int (\varphi_n(\omega) - a)^2 d\mu(\omega) \geq \epsilon^2 \mu(E_\epsilon^n)$ so $\mu(E_\epsilon^n) \leq \frac{\delta}{n\epsilon^2}$. Thus for each $\epsilon > 0$ and each $\eta > 0$, one can find n_0 so that $n > n_0$ implies $\mu(E_\epsilon^n) < \eta$. Now $\mu(E_\epsilon^n)$ is the probability that $\frac{f_1(\omega) + \cdots + f_n(\omega)}{n}$ differs from a by less than ϵ. We have proved that this probability can be made arbitrarily small by choosing n sufficiently large. Bernoulli's weak law of large numbers is the very special case in which the distributions of the f_j are all identical and are concentrated in a finite number of points. His proof was much more complicated.

Let $f_1, f_2, \cdots$ be a sequence of independent random variables with a common distribution α, and suppose that the f_j have zero expected value and finite variance ν. One verifies at once that $\varphi_n = \frac{f_1 + f_2 + \cdots + f_n}{n}$ has zero expected value and variance $\frac{\nu}{n}$. Thus $\varphi_n\sqrt{n}$ has zero expected value and variance ν. The central limit theorem of de Moivre and Laplace, in its simplest form, says that the distribution α_n of $\varphi_n\sqrt{n}$ converges as n tends to ∞ to a probability measure of the form $ce^{-x^2/a^2}dx$. Here a^2 and c are uniquely determined by the fact that the variance must be ν and the measure a probability measure. The first rigorous proof was given by Tchebycheff's student Markov (1856-1922) along lines suggested by Tchebycheff (see section 3). Shortly thereafter, a simpler proof was given by Liapunov (1858-1918), another student of Tchebycheff. Liapunov's proof makes use of Fourier transforms and is an interesting example of the application of harmonic analysis to probability. The idea is very simple. It depends on the easily checked fact that if f_1 and f_2 are independent random variables with distributions α_1 and α_2, and α is the distribution of $f_1 + f_2$, then α is the "convolution" of α_1 and α_2. It then follows from the general properties of Fourier transforms that the continuous positive definite functions $\hat{\alpha}$, $\hat{\alpha}_1$, $\hat{\alpha}_2$ obtained by taking the Fourier transforms of the measures are related by the equation $\hat{\alpha} = \hat{\alpha}_1\hat{\alpha}_2$. (I may mention in passing that specialists in probability theory refer to the Fourier transform of a probability measure as its *characteristic function.*) Now if $f_1, f_2, \cdots, f_n, \cdots$ is a sequence of independent identically distributed random variables, the sum $f_1 + f_2 + \cdots + f_n$ will

have a distribution whose characteristic function is g^n where g is the characteristic function of the common distribution α of the f_j. Thus $\frac{f_1 + \cdots + f_n}{\sqrt{n}}$ will have a distribution whose characteristic function is $y \to (g(y\sqrt{n}))^n$. To prove the central limit theorem one need only investigate the limit of $(g(y\sqrt{n}))^n$ (which turns out to be easy), verify that it is the characteristic function of a distribution of the form $ce^{-x^2/a^2}dx$, and prove an appropriate continuity theorem for the Fourier transform and its inverse. Liapunov's proof is not explicitly of this form, but in 1935 Levy (1886-1971) proved a theorem stating that the one-to-one map from probability measures to positive definite functions defined by the Fourier transform is a homeomorphism when the two spaces are given natural topologies. The central limit theorem in the simple form stated above is essentially a corollary of this theorem in harmonic analysis.

The so-called strong law of large numbers could not even be formulated until the concept of a set of measure zero was available. One says that the strong law of large numbers holds for a sequence $f_1, f_2, \cdots$ of random variables having a common expectation a if $\lim_{n\to\infty} \frac{f_1(\omega) + \cdots + f_n(\omega)}{n} = a$ for almost all ω; that is, if the sequence $\frac{f_1(\omega) + \cdots + f_n(\omega)}{n}$ converges to a with probability one. That this is so for independent random variables with a common distribution and finite variance was proved by Cantelli (1875-1966) in 1917, a less general result having been proved earlier by E. Borel.

Quite generally, a singly or doubly infinite sequence of random variables $f_1, f_2, \cdots$ or $\cdots, f_{-1}, f_0, f_1, f_2, \cdots$ is called a *discrete parameter stochastic process*. The various sequences $f_1(\omega), f_2(\omega), \cdots$ or $\cdots, f_{-1}(\omega), f_0(\omega), \cdots$ arising from the points ω of Ω are called the *sample sequences* of the process. When the random variables are not independent, one classifies processes by the nature of the dependency relations that exist. These are determined by the joint probability distributions in $R \times \overset{n \text{ times}}{R \times \cdots \times R} = R^n$ associated with the finite subsequences $f_k, f_{k+1}, \cdots, f_{k+n-1}$. For example, if the distribution in R^n is independent of k, the process is said to be *stationary*. Perhaps the most important and widely studied class of all is the class of *Markov processes*. To define this class one needs the notion of "conditional probability." Let α be a probability measure in $R \times R$ and let $\tilde{\alpha}$ be the projection on the first factor, i.e., $\tilde{\alpha}(F) = \alpha(F \times R)$ for each Borel subset F of R. By a simple and fundamental theorem in measure theory, there exists for each x in R a probability measure β_x in R such that $\alpha(E \times F) = \int_E \beta_x(F)d\tilde{\alpha}(x)$ for all Borel sets E and F. The measures β_x are uniquely determined (mod $\tilde{\alpha}$ sets of measure zero). If two random variables f and g have α as joint probability distribution, one says that $\beta_x(E)$ is the *conditional prob-*

ability that g is in E given that f takes the value x. The generalization to several variables is obvious. Thus if $f_k, f_{k+1}, \cdots, f_{k+n-1}$ are random variables in a process, one can introduce $\beta_{x_k, x_{k+1}, \cdots, x_{k+n-2}}$ the conditional probability distribution for f_{k+n-1} given that $f_k = x_k, f_{k+1} = x_{k+1}, \cdots, f_{n+n-2} = x_{k+n-2}$. The process is said to be a *Markov process* if $\beta_{x_k, \cdots, x_{k+n-2}}$ is independent of n and of $x_k, \cdots, x_{k+n-3}$ and depends only on x_{k+n-2}. It is said to be a Markov process with *stationary transition probabilities* if $\beta_{x_k, \cdots, x_{k+n-2}} = \beta_{x_{k+n-2}}$ is also independent of $k+n$; that is, if there exists a map $x \to \beta'_x$ of real numbers into probability distributions such that $\beta_{x_k, \cdots, x_{k+n-2}} = \beta'_{x_{k+n-2}}$. One can think of the sample sequences of such a process as generated by a "random walk." Choose an arbitrary starting point x_1. Move to x_2 "at random" using the probability measure β'_{x_1}. Then move to x_3 "at random" using β'_{x_2}, etc. In the special case in which the probability distribution of the f_j are all supported by a fixed finite set — say $\{1, 2, \cdots, k\}$—one says that there is a "finite state space" and $x \to \beta'_x$ is defined by a $k \times k$ matrix. Markov began the theory in 1906 with an analysis of the finite state case.

Instead of a sequence of random variables $f_1, f_2, \cdots$, it is convenient in many problems to consider a family $t \to f_t$ of random variables parameterized by a real number t (interpreted as the time in most applications). While there are additional technical complications (of a highly non-trivial nature) and one brand new problem, in broad outline the theory of continuous parameter Markov processes is similar to the theory of the discrete parameter case. Of course, one has sample functions instead of sample sequences, and the measure μ in the universe Ω leads to a measure in the space of all sample functions. The celebrated Wiener measure is of this character. In fact, Wiener's theory of Brownian motion, developed in the early 1920s, is the theory of the most important special case of a continuous parameter Markov process with stationary probabilities and continuous state space. The general theory of such processes was inaugurated by Kolmogorov in an important paper published in 1931. The new problem arises out of a difference between discrete and continuous parameter Markov processes that is analogous to that between unitary representations of the groups Z and R. In the first case, the representation $n \to U_n$ is uniquely determined by the single unitary operator U_1. In the second case, since R has no least element, one must differentiate and express $t \to U_t = e^{iHt}$ in terms of its "infinitesimal generator" iH in order to describe the representation by a single operator. In a continuous parameter Markov process, the transition probabilities are described by a probability measure β'_{x, t_1, t_2} which depends in general on *three* real variables and has the following interpretation: $\beta'_{x, t_1, t_2}(F)$ is the probability that f_{t_2} will have a value in F given that f_{t_1} has the value x ($t_1 < t_2$). In the stationary case $\beta'_{x, t_1, t_2} = \beta'_{x, 0, t_2 - t_1}$ so that there are effectively only two variables. Quite generally, let $x \to \gamma^1_x$ and $x \to \gamma^2_x$ be two mappings of R into probability measures in R, and suppose that $x \to \gamma^j_x(E)$ is

measurable in x for $j = 1, 2$, and all E. Then the two step random walk using first $x \to \gamma_x^1$ and then $x \to \gamma_x^2$ leads to a mapping $x \to \gamma_x^3$ of R into probability measures which one can think of as the composite $\gamma^2 \circ \gamma^1$ of γ^1 and γ^2. It follows from the definition of a continuous parameter Markov process that $\beta'_{t_3,t_2} \circ \beta'_{t_2,t_1} = \beta'_{t_3,t_1}$ when $t_1 < t_2 < t_3$, a relationship known as the Chapman-Kolmogorov equation. When the transition probabilities are stationary so that $\beta'_{t_1,t_2} = \beta'_{0,t_2-t_1}$, this reduces to $\beta'_{0,t_1+t_2} = \beta'_{0,t_1} \circ \beta'_{0,t_2}$ so that the $\beta'_{0,t}$ constitute a one-parameter semi-group under composition. This has an infinitesimal generator that is defined by a linear operator in a suitable function space. The new problem lies in choosing this function space in such a way that one can recover the $\beta'_{0,t}$ from the infinitesimal generator. In the case of an infinite discrete state space, one has a one parameter semi-group of infinite matrices and one can differentiate to get an infinite matrix. This would seem to be the infinitesimal generator, but unfortunately it does not determine the semi-group. One needs "boundary conditions" in addition. Papers published by Doob (1910—) in 1942 and 1945 and by Lévy in 1951 did much to clarify the situation, but there has been room for much subsequent work by other mathematicians. In the case of a continuous state space, the problems are even more difficult. In Wiener's theory of Brownian motion, the formal infinitesimal generator is a constant times d^2/dx^2, and more generally one defines a diffusion process in such a way that its formal infinitesimal generator is a second order ordinary differential operator (with rather general coefficients). W. Feller (1906-1970) devoted a large fraction of a distinguished career to a study of the determination of the process from its (suitably defined) infinitesimal generator in this case. Many applications of probability theory involve Markov processes, and as in physics one starts by knowing the infinitesimal generator.

The other much studied and extensively applied class of stochastic processes is the class of stationary processes — both discrete and continuous parameter. This class is also the most relevant to our main theme because its theory is related in an intimate way both to harmonic analysis and to the branch of mathematical analysis known as ergodic theory. Ergodic theory is one of the newer branches of mathematics, essentially non-existent before 1931. Moreover, it turns out to have a number of significant connections with the theory of unitary group representations and increases the scope of harmonic analysis in a manner which is only now beginning to be explored. Let us begin with a brief account of the nature and history of ergodic theory.

Let Γ denote the phase space of a classical dynamical system and let U_t be the one-to-one transformation of Γ into itself which is such that $U_t(\gamma)$ is the point of Γ that describes the state of the system t time units after it was described by γ. (We consider only systems that are "reversible" in the sense that the U_t exist.) Then $U_{t_1+t_2} = U_{t_1}U_{t_2}$ and we have an action of the additive

group of the real line on Γ. Let H be the real-valued function on Γ that defines the energy, and for each real number E for which $H^{-1}(E)$ is not empty let $\Gamma_E = H^{-1}(E)$. Then the sets Γ_E are carried into themselves by the U_t, and moreover, each of them admits a natural measure ϱ_E which is U_t invariant. Furthermore, for the systems of interest in statistical mechanics (see section 13), $\varrho_E(\Gamma_E)$ is finite. As part of the program for deducing the fundamental algorithm of statistical mechanics, Boltzmann wanted to prove the following theorem: Let g be a bounded continuous real-valued function on Γ_E. Then the "time average" $\lim_{T\to\infty} \frac{1}{T}\int_0^T g(U_t(\gamma))dt$ is equal to the "space average" $\frac{1}{\varrho_E(\Gamma_E)}\int_{\Gamma_E} g(\gamma)d\varrho_E(\gamma)$ for all choices of γ. Boltzmann naïvely hoped to base a proof of this theorem on the hypothesis that each trajectory or path $t \to U_t(\gamma)$ goes through every point of the constant energy hypersurface Γ_E. Putting together the Greek words *ergon* 'work' and *odos* 'path,' he called his hypothesis the *ergodic hypothesis.* The validity of this hypothesis was much debated, and it was ultimately shown to be untenable for topological reasons that now seem obvious. While various alternative hypotheses were proposed, nothing much could be deduced from them, and the subject remained in a state of uncertainty and confusion for over half a century.

Ergodic theory in its modern sense came into being in 1932 as a consequence of the following sequence of events: In May 1931, B. O. Koopman (1900—), inspired by Stone's paper of 1930 on unitary representations of the real line and by von Neumann's 1929 work on the spectral theorem, published a short note making an observation that today seems quite obvious. He pointed out that if one forms the Hilbert space $\mathscr{L}^2(\Gamma_E, \varrho_E)$, one can obtain a unitary representation V of the real line by setting $V_t(f)(\gamma) = f(U_t(\gamma))$. He also pointed out that Stone's theorem permits one to assign a projection-valued measure to the system, and suggested that it might be fruitful to relate the properties of this projection-valued measure to the properties of the system. Koopman discussed his work with von Neumann before it was published, and this conversation suggested to von Neumann the possibility of applying operator-theoretic methods to prove the equality of space and time averages in statistical mechanics. That Koopman's work might be so applied was also suggested to von Neumann by Weil[6] a short time after Koopman's paper appeared. These facts are stated in the introduction to a short note by von Neumann published in early 1932. In this note von Neumann proves what is now called the mean ergodic theorem. This asserts that for any f in $\mathscr{L}^2(\Gamma_E, \varrho_E)$, the time averages $\frac{1}{T}\int_0^T f(U_t(\gamma))dt = f_T$ converge in the Hilbert space metric to a function $\bar{f}$ which is a constant on the U_t orbits. If the action of U_t on Γ_E is "metrically transitive" in the sense

that there are no measurable invariant subsets (except sets of measure zero and their complements), then it is easy to show that $\bar{f}$ is almost everywhere equal to the constant $\frac{1}{\varrho_E(\Gamma_E)} \int f(\gamma) d\varrho_E(\gamma)$. Except for replacing transitivity by metric transitivity and using convergence "in the mean" instead of pointwise convergence, this is just what Boltzmann had wanted to do. In fact, whenever the hypothesis of metric transitivity can be verified, the physical conclusions desired by Boltzmann follow. Pointwise convergence is not really needed, but it too can be proved. Shortly after learning of von Neumann's result, G. D. Birkhoff (1884-1944) proved the much more difficult pointwise ergodic theorem. For f integrable with respect to ϱ_E, the time averages $\frac{1}{T}\int_0^T f(U_t(\gamma))dt$ converge for almost all γ to an integrable function $\bar{f}$ which is constant on the U_t orbits. As before, if the action is metrically transitive, $\bar{f}$ is almost everywhere equal to $\frac{\int f(\gamma) d\varrho_E(\gamma)}{\varrho_E(\Gamma_E)}$.

Let us look more closely at the hypothesis of metric transitivity. It is very close to the original ergodic hypothesis of Boltzmann, but differs from it in one important respect. The ergodic hypothesis may be reformulated as the hypothesis that Γ_E has no proper subsets that are invariant under the U_t. The hypothesis of metric transitivity seems to be only slightly weaker. It excludes proper invariant subsets, but only those which are measurable and not measure-theoretically improper. One might suppose at first that this weakening made little difference — that under reasonable regularity conditions metric transitivity could be reduced to ordinary transitivity by discarding an invariant set of measure zero. This naïve supposition is false. Consider the action of the infinite cyclic group Z on the circle $|z| = 1$ defined by setting $(z)n = ze^{in\theta}$ where θ is an irrational multiple of π. The ordinary arc length measure in $|z| = 1$ is preserved, and every Z trajectory is countable and hence of measure zero. Cannot some of these trajectories be gathered together into a measurable set other than by taking almost all of them or almost none? A simple argument using Fourier analysis shows that they cannot. One need only study the Fourier coefficients of the characteristic function of a Z-invariant measurable set to see that every such set is either of measure zero or the complement of a set of measure zero. Metric transitivity, though analogous to ordinary transitivity, is *not* reducible to it and in fact turns out to be much more general.

That non-transitive metric transitivity can exist at all was noted for the first time only a few years earlier. Following the lead of G. W. Hill (1838-1914) and Poincaré (1854-1912), G. D. Birkhoff had been studying the qualitative properties of low-dimensional dynamical systems whose equations of motion could not be integrated. Using Poincaré's idea of "surfaces of section," he could reduce problems about actions of the real line on

three-dimensional manifolds to problems about actions of the integers on two-dimensional surfaces. In this connection and in collaboration with Paul Smith (1900—), he published a long paper in 1928 on the structure of such surface transformation groups. This paper contains a definition of metric transitivity and the example sketched above.

It seems to have been von Neumann, however, who first realized the far-reaching significance of the new concept. A slight modification of the example of Birkhoff and Smith shows that it is possible for the real line to act in a metrically transitive manner on a compact manifold of arbitrarily high dimension. Thus (though the question seems difficult to settle in concrete cases) it is at least conceivable that many, and even most, dynamical systems have the property that the action defined by the time evolution of the system is metrically transitive on the constant energy hypersurfaces. As von Neumann emphasized, the hypothesis of metric transitivity is exactly the right substitute for Boltzmann's untenable ergodic hypothesis. Because of this and because "ergodic" is shorter than "metric transitivity," it has become customary to follow von Neumann's lead and call a metrically transitive action an *ergodic* action. Of course, a transitive action or even one reducible to a transitive action by neglecting an invariant set of measure zero is automatically ergodic. It will be convenient to distinguish between the two possibilities by using the terms *essentially transitive* and *properly ergodic* depending on whether there is or is not an orbit of positive measure.

The existence of proper ergodicity raises a host of interesting questions, and modern ergodic theory may be loosely defined as the branch of mathematics that attempts to answer them. The first detailed paper on the subject was published by von Neumann in 1932 and is entitled "Zur Operatorenmethode in der klassischen Mechanik." Theorem 1 of the paper is von Neumann's mean ergodic theorem, which is proved in detail. Theorem 2 is a fundamental decomposition theorem underlining the fact that the ergodic actions are the fundamental building blocks out of which all measure-preserving actions can be constructed. Consider the action of the infinite cyclic group Z on the unit disk $|z| \leq 1$ in the complex plane defined by setting $(z)n = ze^{in\theta}$ where θ/π is irrational. This action preserves the area measure in the disk but is far from ergodic. Indeed, for $0 < a < 1$, the set of all z with $|z| \leq a$ is both invariant and measure-theoretically proper. On the other hand, the circles $|z| = r$ fiber the disk, and the area measures may be recovered from the arc measures in the fibers by an integral over $0 \leq r \leq 1$. Moreover, these fibers are invariant so that the given action is in an obvious sense a "direct integral" of the actions in the fibers. Finally, by an argument already sketched, the fiber actions are all ergodic. Thus the given measure-preserving action decomposes as a "direct integral" of ergodic actions. Von Neumann's Theorem 2 asserts (with mild regularity assumptions) that a fibering (not necessarily with good topological properties) per-

mitting such a decomposition always exists and is essentially unique. Theorem 5 makes a beginning on the difficult and still unsolved problem of classifying the possible "essentially different" ergodic actions of R by giving a complete classification in a special case. Let $t \to V_t$ be the unitary representation of the real line associated with the action by Koopman's construction. Let $V_t = e^{iHt}$ where H is self-adjoint. If H has a basis of eigenvectors (equivalently if V is a discrete direct sum of irreducibles), then the action is said to have a *pure point spectrum.* Given such an action, let $\lambda_1, \lambda_2, \cdots$ be the eigenvalues of H. It is easy to see that the λ_j all occur with multiplicity one and constitute a *subgroup* of the additive group of the real line. Von Neumann showed that actions with a pure point spectrum are determined to within an obvious equivalence by the countable subgroup of eigenvalues of H, and that *every* countable subgroup occurs. The action is properly ergodic if and only if the subgroup is dense, and this happens if and only if the subgroup is not the set of all integer multiples of a fixed positive real number. Using group duality (which was discovered two years later), it is very easy to describe the properly ergodic action associated with a dense countable subgroup D of the real line. Think of D as a subgroup of $\hat{R}$ and let ψ be the natural isomorphism of D into $\hat{R}$. Then ψ^* (see section 17) is a dense imbedding of R in $\hat{D}$. The action of R on $\hat{D}$ defined by setting $(\chi)t = \psi^*(t)\chi$ preserves Haar measure in $\hat{D}$ and is the required properly ergodic action. Theorem 6 is concerned with the properties of an interesting class of ergodic actions of R constructed out of ergodic actions of Z and real-valued functions. A main conclusion is that not only do there exist ergodic actions with point spectrum $\{0\}$, but that these are the rule rather than the exception — at least in the class considered by von Neumann. Von Neumann's paper stimulated many others and ergodic theory developed rapidly. By 1937 E. Hopf (1902—), one of the more active workers, was able to publish a short book on the subject. For a while there was hope that an ergodic action would be completely determined by its spectrum (as in the pure point spectrum case), but that hope turned out to be illusory. Even in the case of ergodic actions of the integers there is no immediate hope of obtaining either a complete classification or a characterization of the spectra that can arise.

Although ergodic theory originated in statistical mechanics, it had little impact on that subject until rather recently. In the first few decades of its existence, its most important applications by far were to the theory of stationary stochastic processes. Consider first the discrete parameter case and let $\cdots, f_{-1}, f_0, f_1, \cdots$ be the random variables of our stationary process. There is no essential loss in generality in assuming that the functions f_j separate the points of the universe Ω in that $\omega \to \cdots, f_{-1}(\omega), f_0(\omega), f_1(\omega), \cdots$ is a one-to-one map of Ω into the product space of countably many replicas of R. If μ' is the image of μ in the product space, the complement of the image

of Ω will be of measure zero. Hence there is no loss in generality in identifying Ω with the space R^Z of *all* real-valued functions on the integers. When this is done, $f_n(\omega)$ is just $\omega(n)$, and translation provides a natural action of Z on Ω such that $f_n(\omega) = f_0([\omega]n)$. Moreover, the definition of stationarity translates into the statement that μ is invariant under the Z action. In other words, a discrete parameter stationary stochastic process is uniquely determined by the system consisting of a measure-preserving action of Z on a probability measure space Ω, μ and a single measurable function f on Ω. The random variables of the process are of course the functions $\omega \to f([\omega]n)$ for $n = \cdots, -1, 0, 1, 2, \cdots$. Similarly, although the argument is less straightforward, a continuous parameter stationary stochastic process is defined by the system consisting of a measure-preserving action of R on a measure space Ω, μ and a measurable function f on Ω, the random variable f_t being $\omega \to f([\omega]t)$. While the measure-preserving action of R or Z canonically associated with a stationary stochastic process need not be ergodic, the underlying universe Ω can always be measure-theoretically fibered into ergodic parts by the theorem of von Neumann stated above. Moreover, any actual event $\omega \in \Omega$ will belong to some ergodic part, and for most applications of the theory one can forget the rest of Ω. For these reasons there is little loss in generality in considering only stationary stochastic processes in which the action is ergodic. Actually, it is quite easy to show that when the random variables are independent (which can happen only in the discrete parameter case) then the underlying action is ergodic!

Suppose then that our stationary stochastic process is defined by an ergodic action of R or Z on a probability measure space Ω, μ and a real-valued measurable function f. If the random variables have expectations, so that $\int(f(\omega))d\mu(\omega) < \infty$, we may apply the pointwise ergodic theorem of G. D. Birkhoff. There is a discrete version as well as a continuous one, and it asserts that for almost all ω, $\lim_{n\to\infty} \dfrac{f(\omega) + f(\omega\bullet 1) + f(\omega\bullet 2) + \cdots + f(\omega\bullet n)}{n+1}$ exists and equals $\int f(\omega)d\mu(\omega)$. But $f((\omega)n) = f_n(\omega)$, and $\int f(\omega)d\mu(\omega)$ is the common expectation of the random variables f_n. Thus the discrete version of the pointwise ergodic theorem is nothing more or less than the strong law of large numbers for arbitrary (ergodic) stationary stochastic processes. As such it is a considerable generalization of the Borel-Cantelli theorem, which states the strong law for independent identically-distributed random variables. It is curious that this interpretation of the pointwise ergodic theorem was not immediately recognized. While G. D. Birkhoff's proof was announced and sketched in late 1931, and Kolmogorov's book on measure theoretic foundations appeared in 1933, the book does not mention the ergodic theorem. The connection between the ergodic theorem and the strong law of large numbers was not mentioned in print until 1934. It was pointed out in three independent papers published in that year by Doob, E. Hopf,

and Khintchine respectively.

Suppose now that the random variables have finite variances so that f is in $\mathscr{L}^2(\Omega, \mu)$. Let $G = Z$ or R depending upon whether we are in the discrete case or the continuous case. Let $x \to V_x$ be the unitary representation of G defined by the Koopman construction. Then $(V_x(f) \cdot f)$ is a positive definite function canonically associated with the process, and by the Bochner (or Herglotz) theorem $V_x(f) \cdot f = \int \chi(x) d\nu(\chi)$ where ν is a finite measure on $\hat{G}$, i.e., on the circle or the line. This measure is called the *spectrum* of the process and plays an important role in its theory. Another interesting consequence of the pointwise ergodic theorem is that the function $x \to (V_x(f) \cdot f)$, and hence the spectrum ν can be determined from the "complete past" of almost any sample function. For definiteness let us look at the continuous case. Then $V_x(f)(\omega) = f([\omega]x)$, and since f and $V_x(f)$ are both square integrable, their product $(V_x(f))f = g$ is integrable. Hence, by the ergodic theorem, $\lim_{T \to \infty} \frac{1}{T} \int_0^T g([\omega](t))dt$ exists for almost all ω and equals $\int g(\omega) d\mu(\omega) = V_x(f) \cdot f$. Now $g([\omega](t)) = f([\omega](t))f([\omega](x - t)) = f_\omega(t) f_\omega(x - t)$ where $t \to f_\omega(t) = f([\omega]t)$ is the sample function attached to the point ω of the universe Ω. Thus, for almost all ω, $\lim_{T \to \infty} \frac{1}{T} \int_0^T f_\omega(-t) f_\omega(x - t)dt$ exists and equals $V_x(f) \cdot f$. It can be computed for negative x when one knows $f_\omega(t)$ for $t \leq 0$, and since it is an even function of x, this determines it for all x. The significance of this result lies in the fact that the theory of stationary stochastic processes is mainly applied to the statistical analysis of so-called "time series" — such as occur for example in economics and meteorology. One thinks of the sample functions $t \to f_\omega(t)$ as "possible" functions describing the temperature, say, or the price of wheat as a function of the time. One supposes that the variation of temperature or wheat prices actually observed is given by a function f_{ω_0} chosen "at random" from Ω according to the probability measure μ. Looking at how $f_{\omega_0}(t)$ has behaved for $t \leq 0$ (i.e., in the past), one can compute the Fourier transform of the spectrum of the whole stochastic process. It is just the so-called "autocorrelation function" $\lim_{T \to \infty} \frac{1}{T}$ $\int_0^T f_{\omega_0}(-t) f_{\omega_0}(x - t)dt$ of the sample function f_{ω_0}.

Some appreciation of the significance of the spectrum of a process may be had by looking at its discrete components, if any. Suppose that the measure ν of which $V_x(f) \cdot f$ is the Fourier transform has an "atom," i.e., that $\nu(\{\lambda\}) \neq 0$ for some λ in $\hat{R}$. This happens precisely when there exists $\varphi \not\equiv 0$ in $\mathscr{L}^2(\Omega, \mu)$ such that $V_x(\varphi) \equiv e^{i\lambda x}\varphi$, and then φ is uniquely determined up to a multiplicative constant and f may be written uniquely in the form $f = c\varphi + f^\perp$ where φ and $f^\perp$ are orthogonal in $\mathscr{L}^2(\Omega, \mu)$. Now $f_\omega(t) = f([\omega]t) = c\varphi([\omega]t) + f^\perp([\omega]t) = ce^{i\lambda t}\varphi(\omega) + f^\perp([\omega]t)$. Thus every sample function has a canonical decomposition as the sum of a constant multiple of the periodic

function $e^{i\lambda t}$ and a sample function of a new process defined by $f^{\perp}$. The spectrum of the new process is the same as that of the original except for the removal of the atom at λ. Other atoms can be removed similarly, and in fact one can write $f = \Sigma\, c_j\varphi_j + g$ where the φ_j and g are mutually orthogonal, $|\varphi_j(\omega)| \equiv 1$, $V_t(\varphi_j) = e^{i\lambda_j t}\varphi_j$, and g defines a process whose spectrum has no atoms. Correspondingly, the sample function in f_ω may be written $f_\omega(t) = \Sigma\, c_j\varphi_j(\omega)e^{i\lambda_j t} + g_\omega(t)$. The terms $c_j\varphi_j(\omega)e^{i\lambda_j t}$ are called the "hidden periodicities" of the sample function. Using the ergodic theorem just as before (but applied to $f\overline{\varphi_j}$), one finds that the coefficient $c_j\varphi_j(\omega)$ may be computed from the past of almost any sample function by the formula

$$c_j\varphi_j(\omega) = \lim_{T\to\infty} \frac{1}{T}\int_0^T f_\omega(-t)e^{-i\lambda_j t}dt.$$

This formula is essentially identical with that used by Bohr (see section 17) in computing the expansion coefficients of almost-periodic functions. Indeed, when the c_j do not go to zero too slowly, the difference $f_\omega(t) - g_\omega(t)$ is an almost-periodic function in Bohr's sense. In any case, it is a generalized almost-periodic function whose values for $t > 0$ are determined by its values for $t < 0$. There can be no true randomness in our stochastic process if the underlying ergodic action has a pure point spectrum.

The consequences of ergodic theory for the statistical study of time series just described were pointed out in the 1930s, the major publications being papers by Khintchine and Wold published in 1934 and 1938 respectively. They provided a justification for and a conceptual clarification of methods already in use by scientists and statisticians in studying specific time series. The determination of hidden periods goes back to work of the physicist Schuster (1851-1934) beginning in 1898. The use of the autocorrelation functions goes back to before 1920. A much-cited paper by Yule (1871-1951), published in 1927, makes use of autocorrelations in studying sunspot data.

In 1930 Norbert Wiener published a long memoir entitled "Generalized harmonic analysis," which though conceived in a different spirit was in effect a remarkable anticipation of the theory of stationary time series. He had been engaged in trying to help electrical communication engineers cope with some of the problems of circuit design, and these problems seemed to require applying harmonic analysis to functions which were neither square summable nor periodic and, moreover, were more general than the almost-periodic functions of Bohr. For reasons which I shall not attempt to describe, Wiener decided to study measurable functions for which the autocorrelation function exists and is continuous. As we have seen, it is a consequence of the ergodic theorem that the sample functions of stationary stochastic processes constitute a rich source of examples. While the ergodic theorem was not stated or proved until a year or so after Wiener's paper appeared, practically all the examples offered by Wiener were essentially sam-

ple functions and were defined using randomness. On the other hand, Wiener did not at the time think of himself as studying a whole statistical ensemble of functions at once as in the case of the f_ω, but as studying a single function. His chief concern was to define the spectrum of the function, which he had to do without using Bochner's theorem (Bochner's theorem was published in 1932).

In studying carefully the relationship between a function and its spectrum, Wiener found himself in need of a more powerful Tauberian theorem than any that existed (see section 14), and was thereby led to write his prize-winning paper "Tauberian theorems," which was published in 1932.

Later Wiener came to realize the merits of thinking of his functions as sample functions. This led him to the important insight that the coded messages with which the communication engineers had to deal were close analogues of time series. Modern communication engineering makes heavy use of his discovery.

From the point of view of this article the relationship between Fourier integrals, Fourier series, almost-periodic function expansions, and Wiener's generalized harmonic analysis is best viewed as follows: Let S, μ be a suitably restricted measure space and let the real line R act upon S to preserve μ. Harmonic analysis on R is concerned with decomposing the unitary representation V of R defined in $\mathscr{L}^2(S, \mu)$ by the Koopman construction. Now by von Neumann's decomposition theorem, the given action can be decomposed into ergodic actions, and correspondingly V decomposes as a direct integral of Koopman representations — one for each ergodic component. This part of the decomposition is basically geometry, and one may think of harmonic analysis proper as the decomposition of V when the action is ergodic. We now divide into four cases according to whether ergodic action is properly ergodic or essentially transitive, and also according to whether $\mu(S)$ is finite or infinite. When the action is essentially transitive, S may be taken to be R modulo a closed subgroup which is necessarily either $\{0\}$ or the subgroup of all integer multiples of a, and the action is then just translation on the quotient group. Depending upon whether the subgroup is $\{0\}$ or not, that is upon whether $\mu(S) = \infty$ or $\mu(S) < \infty$, one is reduced to the Fourier transform or to Fourier series. When the action is properly ergodic, one can no longer identify S with the real line or one of its quotient groups. However, the decomposition of the Koopman representation V may still be regarded as defining a decomposition of real-valued functions on the real line, namely the functions $t \to f([s]t)$ for each s in S and each f in $\mathscr{L}^2(S, \mu)$. Indeed, for each E and each s the function $t \to P_E^V(f)([s]t)$ may be regarded as the component of $t \to f([s]t)$ whose spectrum is in the set E. When $\mu(S) < \infty$, this analysis is essentially the generalized harmonic analysis of Wiener — carried somewhat further than Wiener carried it. When in addition V is a discrete direct sum of irreducible (and hence one-dimensional) representa-

tions, one recovers a slight generalization of Bohr's theory of almost-periodic functions (see section 17). The case in which $\mu(S) = \infty$ does not seem to have been investigated.

19. EARLY APPLICATIONS OF GROUP REPRESENTATIONS TO NUMBER THEORY—THE WORK OF ARTIN AND HECKE

In the first quarter-century of its existence, the theory of representations of finite groups had many applications to group theory itself. I have already mentioned the theorem on the structure of groups of order $p^\alpha q^\beta$. To my knowledge, however, there were no applications to other fields such as physics, number theory, or probability until the 1920s. The extensive applications to quantum mechanics, which began in 1927 (see section 16) have had the most publicity, but they were not the first. Appropriately, in view of Dedekind's role in inspiring Frobenius, the first application outside group theory seems to have been to Dedekind's own creation — the theory of algebraic number fields. It was made in a celebrated paper of Artin (1898-1962) entitled "Über eine neue Art von L-Reihen," published in 1923.

Before attempting to explain the nature and significance of what Artin did in this paper, I must devote a few paragraphs to sketching the origins of the theory of algebraic number fields and the course of its development between its beginnings in the 1870s and the publication of Artin's paper in 1923.

As observed by Gauss in 1830, the problem of finding the integer solutions of the equation $x^2 + y^2 = n$ may be usefully approached by factoring $x^2 + y^2$ as the product of $x + iy$ and $x - iy$. The set of all complex numbers of the form $x + iy$ where x and y are integers is a ring called the ring of *Gaussian integers.* Like the ordinary integers, the Gaussian integers admit "unique" factorization into "primes." One calls a Gaussian integer a *unit* if it has a multiplicative inverse, and a *prime* if it cannot be written as a product of Gaussian integers other than units and products of units with itself. It is then easy to prove that every Gaussian integer is a product of primes and that this factorization is unique up to reordering and multiplication by units. The units are $1, -1, i$, and $-i$. Now since $(a + ib)(a - ib) = a^2 + b^2$, it follows at once that every Gaussian prime is a factor of some ordinary integer and hence of some ordinary prime. To determine the Gaussian primes, it thus suffices to factor the ordinary primes, and it is obvious that an ordinary prime p which is not already a Gaussian prime must be of the form $x^2 + y^2$ and factor into $x + iy$ and $x - iy$. It is not difficult to show that $x + iy$ and $x - iy$ are always Gaussian primes and are equal mod units only when $p = 2$. In that case $(1 + i) = (1 - i)(i)$. Thus once one knows for which primes p one can solve $x^2 + y^2 = p$, one knows all Gaussian primes. To solve $x^2 + y^2 = n$, one then has only to factor n into Gaussian primes

and then divide the factors into two classes in such a way that each class contains just one of each pair of conjugates.

The advantage of this approach is that it may be applied not only to quadratic Diophantine equations other than $x^2 + y^2 = n$, but to higher order equations as well. For example, when m is odd, one can study the equations $x^m + y^m = n$ by factoring $x^m + y^m$ as the product $(x + \omega y)(x + \omega^2 y) \cdots (x + \omega^{m-1} y)(x + y)$ where ω is a primitive mth root of 1. There is a difficulty, however, in that the ring generated by ω need not have unique factorization. If it did, one could obtain enough information about the solvability of $x^m + y^m = n$ to prove Fermat's famous conjecture about the nonexistence of integer solutions of $x^m + y^m = z^m$ when $m > 2$. Indeed, in the early 1840s, Kummer (1810-1893), overlooking the possible failure of unique factorization, thought he had a proof of the Fermat conjecture in the general case. This mistake led him to attack the problem of finding a substitute for the unique factorization law and to solve the problem for the particular case of the ring generated by the mth roots of 1. The "ideal numbers" he introduced as substitutes for primes sufficed to deal with various special cases of the Fermat problem, but the general case remains open to this day.

If one passes from $x^m + y^m$ to a general homogeneous form of the mth degree $a_m x^m + a_{m-1} x^{m-1} y + \cdots + a_0 y^m$, one can still approach the problem of solving Diophantine equations of the form $a_m x^m + a_{m-1} x^{m-1} y + \cdots + a_0 y^m = n$ by factoring the left-hand side. One can write it as $a_m y^m((x/y)^m + a_{m-1}(x/y)^{m-1} + \cdots + a_0) = a_m y^m((x/y) - \alpha_1)((x/y) - \alpha_2) \cdots ((x/y) - \alpha_m) = a_m(x - \alpha_1 y)(x - \alpha_2 y) \cdots (x - \alpha_m y)$, where the α_j are the roots of $a_m x^m + a_{m-1} x^{m-1} + \cdots + a_0 = 0$, and attempt to generalize Kummer's ideas to the ring generated by the α_j. It turned out to be not at all obvious how to do this. The problem remained unsolved until attacked in different ways by two younger mathematicians, Kronecker (1823-1891) and Dedekind (1831-1916). Although Kronecker was Kummer's pupil, his solution was published a decade after that of Dedekind, and proved the less popular. Dedekind's theory appeared in 1871 as supplement X to the second edition of Dirichlet's lectures on number theory. Various revised forms appeared in later editions.

Given an equation $a_m x^m + \cdots + a_0 = 0$ with integer coefficients, let $\mathcal{F}$ be the smallest set of complex numbers containing all the roots of the equation and closed under addition, multiplication, and division. In other words, let $\mathcal{F}$ be the so-called *algebraic number field* generated by the roots in question. Every member of $\mathcal{F}$ satisfies some polynomial equation with integer coefficients, and those that satisfy such an equation with leading coefficient 1 are said to be *algebraic integers*. The set R of all algebraic integers in $\mathcal{F}$ is a ring, and Dedekind concerned himself with factorization in this ring. For each non- zero algebraic integer x in R, one can form the set I_x of all xy

for y in R and prove that $I_{x_1} = I_{x_2}$ if and only if $x_1 = ux_2$ where u is a unit. Moreover, it is easy to see that for all x_1 and x_2, $I_{x_1x_2} = I_{x_1}I_{x_2}$ where the latter is defined to be the set of all sums $y_1z_1 + y_2z_2 + \cdots y_rz_r$ where the y_j are in I_{x_1} and the z_j are in I_{x_2}. Thus factorization can be translated into properties of the sets I_x and then one doesn't have to be concerned about the arbitrariness produced by units. Dedekind's key observation is that in those rings for which unique factorization does not hold, one can augment the subrings I_x with other subrings in such a way that unique factorization is restored. Each I_x in addition to being a subring has the property that $z \in I_x$, and $y \in R$ implies $zy \in I_x$. Dedekind defined a subring to be an *ideal* whenever it has this property. When R has unique factorization, the ideals correspond one-to-one to the elements mod units. Otherwise there are *always* ideals which are not of the form I_x, and Dedekind thought of them as defining "virtual" or "ideal" elements — hence the word *ideal*. Defining an ideal I in R to be prime when it cannot be written in the form I_1I_2 where I_1 and I_2 are (not necessarily distinct) ideals, Dedekind was able to prove that every ideal may be written in the form $I_1^{k_1} I_2^{k_2} \cdots I_r^{k_r}$ where the I_j are distinct prime ideals and are uniquely determined up to a rearrangement.

An ideal of the form I_x is said to be principal, and one can show that the factorization of principal ideals leads to all prime ideals — indeed that every prime ideal "lies over" one and only one ordinary prime p in the sense that it occurs in the factorization of I_p. For any ideal I other than $\{0\}$, one can introduce an equivalence relation in R by setting $x \equiv y \bmod I$ if $x - y$ is in I, and show that there are only a finite number of equivalence classes. The number of these is denoted by $N(I)$ and called the *norm* of the ideal. When R is the ring of ordinary integers $N(I_x)$ is just $|x|$. More generally, $N(I_1I_2) = N(I_1)N(I_2)$, and when p is an ordinary prime $N(I_p) = p^m$ where m is the dimension of the field considered as a vector space over the rationals. Thus every prime ideal lying over p has a norm which is a power of p. Actually it turns out that with the exception of a finite number of so called "ramified" primes, all the prime ideals lying over p are distinct and have the same norm p^k. If there are ℓ of these, then $p^{k\ell} = p^m$, so k and ℓ are divisors of m. Knowing ℓ for each prime p and knowing how the ramified primes behave tells us $N(I)$ for all possible ideals I. This information is summed up in the Dedekind zeta function of the field, which is defined by the equation $\zeta_{\mathfrak{F}}(s) = \sum_I \frac{1}{N(I)^s}$ — the sum being over all non-zero ideals. Of course one has also $\zeta_{\mathfrak{F}}(s) = \sum_{n=1}^{\infty} \frac{\varphi_{\mathfrak{F}}(n)}{n^s}$ where $\varphi_{\mathfrak{F}}(n)$ is the number of ideals of norm n.

In the special case in which the field $\mathfrak{F}$ is a two-dimensional vector space over the rational numbers, a so-called quadratic extension, then $\zeta_{\mathfrak{F}}(s)$ coincides with the function $\sum_{n=1}^{\infty} \frac{\varphi_{\mathfrak{F}}(n)}{n^s}$ which Dirichlet attached to a binary quad-

ratic form a third of a century earlier (see section 6) in his proof of the existence of an infinity of primes in an arithmetic progression. In fact, the theory of the binary quadratic forms of a fixed discriminant D is more or less equivalent to the ideal theory in the algebraic number field generated by $\sqrt{D}$, and the theory of algebraic number fields as worked out by Dedekind and his successors may be looked upon as a far-reaching generalization of the theory of binary quadratic forms of Gauss and Dirichlet. In this generalization, the algebraic number fields whose integers admit unique factorization correspond to quadratic forms of class number one. For more general algebraic number fields there is a finite commutative group that generalizes the group formed by the inequivalent classes of quadratic forms of a given discriminant under Gauss's composition law (see section 6). This is the group of ideal classes, which may be defined as follows: One declares the ideals I_1 and I_2 to be in the same class if there exist elements x and y such that $I_1 I_x = I_2 I_y$. This equivalence relation divides all ideals into a finite number of classes and the number h is called the *class number* of the field. It is obvious that the class of II' depends only on the classes to which I and I' belong, and that the resulting composition law makes the classes into a commutative group. In addition to the problem of determining $\varphi_{\mathfrak{F}}(n)$ (the number of ideals of norm n), one has also the more delicate problem of determining for each ideal class c the number $\varphi^c_{\mathfrak{F}}(n)$ of ideals of norm n in that class. This latter problem is the analogue of the problem of finding the number of representations of n by a particular quadratic form, and as with that problem $n \to \varphi^c_{\mathfrak{F}}(n)$ is not a multiplicative function of n (see section 6), but a linear combination of such functions with one term for each character of the ideal class group.

Generally speaking, the theory of algebraic number fields parallels the theory of binary quadratic forms except for the extra complications produced by passing from a second to a higher degree equation. Of course, these added complications can be quite serious, and one has as complete a knowledge of the function $n \to \varphi_{\mathfrak{F}}(n)$ as in the quadratic form case only when the Galois group (i.e., the group of automorphisms of $\mathfrak{F}$) is a commutative group. In that case, one has a generalization of the quadratic reciprocity law in that whether or not I_p is a prime ideal — and more generally the numbers of prime ideals into which I_p decomposes — depends (for the unramified primes) only on the congruence class of p relative to some fixed modulus m. That this is so is by no means obvious. It is a consequence of Kummer's results on the field generated by the mth roots of unity and of a remarkable theorem conjectured by Kronecker in 1886 and first completely proved by Weber (1842-1913) in 1887. This theorem asserts that every $\mathfrak{F}$ with a commutative Galois group is contained for some m in the field generated by the mth roots of unity. Very little is known about the dependence on

p of the prime decomposition law when the Galois group is not commutative.

Let $\mathcal{F}$ be as above with Galois group G, and for each subgroup H of G let $\mathcal{F}_H$ be the subfield of all x in $\mathcal{F}$ such that $\alpha(x) = x$ for all automorphisms α of H. In its modern form, one of the main results of the Galois theory of equations (see section 11) asserts that *every* subfield of $\mathcal{F}$ is an $\mathcal{F}_H$ and that $H \to \mathcal{F}_H$ sets up a one-to-one inclusion inverting correspondence between the subgroups of G and the subfields of $\mathcal{F}$. Now each $\mathcal{F}_H$ has a subring of integers R_H and one can consider generalizing the problem of factoring the ideals I_p to that of factoring the ideals in R generated by the prime ideals in R_H. In other words, one can study ideal factorization in $\mathcal{F}$ *relative* to an arbitrary subfield $\mathcal{F}_H$. Moreover, in view of the facts stated above about the case in which $H = G$ and $\mathcal{F}_H$ is the rational subfield Q, it is natural to hope for immediate results only when H is Abelian; in other words, to study first those extensions $\mathcal{F}|\mathcal{F}_H$ in which the *relative* Galois group is Abelian.

Extending the classical results of Gauss, Dirichlet, and Kummer to relatively Abelian extension fields turned out to be far from easy. Important preliminary results were obtained by Kronecker in 1882 and by Weber in 1891, 1897, and 1898. Hilbert is usually credited with having begun the systematic general theory, however, in a series of papers published between 1898 and 1902. Even the case of a relative quadratic extension proved to be difficult; it was the only one that Hilbert worked out in full detail. However, he outlined how the theory should look for a more general Abelian relative Galois group and made a number of conjectures which were established in the next two decades by his student Furtwängler (1864-1939) and by Takagi (1875-1966). Takagi, who brought the subject to a certain degree of completion in two important papers published in 1920 and 1922 respectively, not only proved Hilbert's conjectures, but made important conceptual advances as well. For a more complete account of the relationship between the work of Kronecker, Weber, Hilbert, Furtwängler, and Takagi the reader is referred to Hasse's article "History of class field theory," published in the proceedings of the 1965 Brighton Conference on algebraic number theory.

In generalizing the quadratic reciprocity law, Hilbert was led to an elegant new formulation in the classical case. It is based on a concept introduced by him and called the *norm residue symbol.* Let $\mathcal{F}$ be a quadratic extension field of the rational field Q. Then the norm residue symbol assigns a character $\chi_p^{\mathcal{F}}$ of the multiplicative group Q^* of Q to each prime p and to ∞. This assignment is such that $\chi_p^{\mathcal{F}}(r) = \pm 1$, and for each r and $\mathcal{F}$, $\chi_p^{\mathcal{F}}(r) = -1$ for only finitely many values of p. Thus $\prod_p \chi_p^{\mathcal{F}}(r)$ makes sense. Hilbert showed that the classical quadratic reciprocity law (see section 6), together with its two supplements, is completely equivalent to the assertion

that $\prod_p \chi_p^{\mathcal{F}}(r) = 1$ for all non- zero rational numbers r (p of course ranges over all primes *and* ∞). Hilbert's definition of the character $\chi_p^{\mathcal{F}}$ depended on the factorization of I_p in the ring of integers of $\mathcal{F}$ in a way which it will be easier to describe in section 20 below.

For each prime, the characters $\chi_p^{\mathcal{F}}$ that arise as $\mathcal{F}$ varies over the quadratic number fields turn out to form a subgroup A_p of the group of all characters of order 2. Hilbert also showed that a system $p \to \chi_p$ of characters of order 2 arises as $p \to \chi_p^{\mathcal{F}}$ for some $\mathcal{F}$ if and only if the following conditions are satisfied: 1) $\chi_p \in A_p$ for all p including ∞; 2) for all but a finite number of values of p, $\chi_p(n) = 1$ whenever p does not divide n; 3) $\prod_p \chi_p(r) = 1$ for all non-zero rationals r. In this sense, Hilbert described all possible quadratic extensions of Q in terms of the character groups A_p. In his theory of relative quadratic extensions, he found generalizations of both the quadratic reciprocity law and the theorem about the possible quadratic extensions of Q.

When one goes beyond the quadratic case to more general Abelian extensions, the system $p \to \chi_p$ of characters of order 2 must be replaced by a finite set of systems $p \to \chi_p$, where the χ_p are now only of finite order and where the finite set forms a group under pointwise multiplication. This finite group turns out to be isomorphic to the Galois group of the field $\mathcal{F}$. (Of course when the Galois group is of order two, it suffices to specify the unique system $p \to \chi_p$ which does not correspond to the identity.) In the generalization to relative fields, the prime ideals for the base field $\mathcal{F}_H$ replace the primes, and the symbol ∞ gets replaced by several such symbols — one for each possible dense imbedding of $\mathcal{F}_H$ in the real or complex number fields.

With this background it is possible to describe Artin's 1923 application of group representations to number theory. It is based on Artin's discovery that for each $\mathcal{F}$ there is a canonical way of assigning a Dirichlet series $s \to L(s, \chi, \mathcal{F}_H)$ to every pair $\mathcal{F}_H$, χ consisting of a subfield $\mathcal{F}_H$ of $\mathcal{F}$ and a character χ of the Galois group H of $\mathcal{F}$ relative to $\mathcal{F}_H$. Artin's definition of $L(s, \chi, \mathcal{F}_H)$ depends in turn upon certain facts about the action of H on the prime ideals in R which lie over a fixed prime ideal $\mathfrak{P}$ in R_H. It turns out that (with the exception of a finite number of "ramified" prime ideals $\mathfrak{P}$) H acts transitively on the prime ideals of R lying over $\mathfrak{P}$ and that the subgroup of H leaving a given one of these fixed is always cyclic and, moreover, has a canonical generator. Since the conjugacy class of this generator is evidently the same for all prime ideals on R lying over $\mathfrak{P}$, we have a natural map $\mathfrak{P} \to C_{\mathfrak{P}}$ of (unramified) prime ideals in R_H into the conjugacy classes in H. Now for each $\mathfrak{P}$ and each representation Γ of H, the determinant of $(I - \Gamma_x N(\mathfrak{P})^{-s})$ where I is the identity is easily seen to depend only on the

conjugacy class to which x belongs — and of course on s and $\mathfrak{P}$. Choosing x to be any element in C this determinant becomes a well-defined function $f_{\mathfrak{P}}(s)$ of s and $\mathfrak{P}$. The Artin "L function" $L(s, \chi, \mathscr{F}_H)$ is $\prod_{\mathfrak{P}} \frac{1}{f_{\mathfrak{P}}(s)}$ where Γ is any representation of character χ.

Although it is possible to define $f_{\mathfrak{P}}(s)$ for the ramified prime ideals as well, it suffices for many purposes to look only at the unramified ones and to identify two Dirichlet series when one can be transformed into the other by multiplying or dividing by a finite number of factors of the form $\frac{1}{Q(p^{-s})}$ where Q is a polynomial and p is an ordinary prime. Modulo such factors one can then verify the truth of the following relationships:

(1) When χ is the identity character, $L(s, \chi, \mathscr{F}_H)$ is the zeta function of $\mathscr{F}$ relative to $\mathscr{F}_H$.

(2) $L(s, \chi_1 + \chi_2, \mathscr{F}_H) = L(s, \chi_1, \mathscr{F}_H)L(s, \chi_2, \mathscr{F}_H)$ for all characters χ_1 and χ_2 of H.

(3) If χ^* is the character of H *induced* by a character χ of some subgroup H_1 of H (see section 15), then $L(s, \chi^*, \mathscr{F}_H) = L(s, \chi, \mathscr{F}_{H_1})$.

Applying (3) to the case in which H_1 contains only the identity and χ is one-dimensional, and using the fact that χ^* is then the sum of *all* irreducible characters $\chi_1, \chi_2, \cdots, \chi_\ell$ each occurring with multiplicity $\chi_j(e)$, one finds that $L(s, \Sigma\chi_j(e)\chi_j, \mathscr{F}_H) = L(s, 1, \mathscr{F})$. It now follows on applying (2) and (1) to the left- and right-hand sides respectively that $\prod_j L(s, \chi_j, \mathscr{F}_H)^{\chi_j(e)} = \zeta_{\mathscr{F}}(s)$. In other words, the zeta function of any $\mathscr{F}$ (modulo the equivalence relation mentioned above) can be factored as a product of the Artin L functions attached to the irreducible characters of any fixed subgroup H of the Galois group G of $\mathscr{F}$. This factorization is significant because when H is commutative, it follows from the work of Takagi that $\zeta_{\mathscr{F}}(s)$ factors as a product of generalized Dirichlet L functions and that Artin's L functions for the characters of H coincide (modulo equivalence) with the generalized Dirichlet L functions. Moreover, this coincidence of the two kinds of L functions is not an immediate consequence of the definitions, but is equivalent to one of the main theorems of class field theory. In fact, a few years later Artin was able to give a new proof of this theorem and reorganize the whole subject in a conceptually advantageous way by starting with a direct proof of the identity of the two kinds of L functions.

The identity of the two kinds of L functions in the commutative case provided for the first time a natural generalization of the Dirichlet L functions for fields with a non-commutative Galois group, and therefore a tool with which to attack "non-commutative class field theory." Using the relationship $L(s, \chi^*, \mathscr{F}_H) = L(s, \chi, \mathscr{F}_{H_1})$, one can express Artin L functions based on non-commutative characters in terms of generalized Dirichlet L functions

to the extent that general characters on G can be expressed in terms of characters induced by one-dimensional characters of subgroups. Using this device, Artin was able to establish various important properties of his new L functions and to show that others would follow if one could prove that every character of an arbitrary finite group can be written as a linear combination with positive and negative integer coefficients of characters induced by one-dimensional characters of subgroups. A proof of this difficult theorem about finite groups was found by Brauer (1901-1977) and published in 1947. Brauer was a student of Schur and his most important immediate successor in developing the representation theory of finite groups.

Another early application of the theory of group representations to number theory was provided by Hecke (1887-1947) in 1928. It is perhaps more accurate to say that it was an application to the theory of modular forms, but the connection of the latter theory to questions in number theory — especially the theory of n-ary quadratic forms (see sections 10 and 22) — is so close that it seems appropriate to speak of an application to number theory. Let Γ_0 denote the subgroup of the group $SL(2, R)$ of all 2×2 real matrices of determinant 1 consisting of the matrices with integer coefficients. Let Γ_N be the normal subgroup of Γ_0 consisting of all $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ in Γ_0 for which $a - 1$, $d - 1$, b, and c are divisible by N. Then (see section 10) a modular form of weight k and level N is an entire function on the upper half-plane satisfying the identity $f\left(\frac{az+b}{cz+d}\right) = (cz + d)^{2k} f(z)$ for all $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_N$ (and certain growth conditions as well). It was known that for each fixed k and N, the modular forms of weight k and level N form a finite-dimensional vector space. Hecke was concerned with the problem of finding an explicit basis for this space. For forms of level 1, such a basis was described in section 10. The basic idea of Hecke's paper was to break this problem down into subproblems as follows: For each modular form f of level N and weight k, consider the set of all functions $z \to f\left(\frac{az+b}{cz+d}\right)(cz + d)^{-2k}$ for $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0$. When $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_N$, these all reduce to f itself. More generally, since Γ_N has finite index in Γ_0, these functions span a finite-dimensional vector space $\mathcal{M}_f$. For each $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0$, then $g \to g\left(\frac{az+b}{cz+d}\right)(cz + d)^{-2k}$ is a linear transformation $L_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}$ of $\mathcal{M}_f$ into itself, and the mapping $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \to L_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}$ is a representation of Γ_0. Since $L_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}$ is the identity whenever $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ is in the normal subgroup Γ_N, this representation is in effect a representation of the finite quotient group Γ_0/Γ_N. Let $\mathcal{M}_f = \mathcal{M}_1 \oplus \mathcal{M}_2 \oplus \cdots \oplus \mathcal{M}_e$ be a decomposition of $\mathcal{M}_f$ as a direct sum of irreducible L-invariant subspaces, and let f_j be the component of f in $\mathcal{M}_j$. It is evident that each f_j is a modular form of weight k and level N but is special in being intrinsically associated with a particular irreducible representation of Γ_0/Γ_N. Hecke suggested the strategy of looking at the irreducible representations of Γ_0/Γ_N one at a time and for each representation W seeking a basis for those particular modular forms intrinsically associated with

W. It is not difficult to see that this problem is more or less equivalent to the following: For each irreducible representation W of Γ_0/Γ_N, let us define an entire function g from the upper half-plane to the vector space of W to be a modular W form of weight k if it satisfies the identity

$$g\left(\tfrac{az+b}{cz+d}\right) = (cz + d)^{2k} W_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)} g(z) \text{ for all } \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0$$

as well as an appropriate growth condition. The equivalent problem then is to find a basis for the modular W forms of weight k.

In order to carry out this program, one has first to determine the irreducible representations of Γ_0/Γ_N. When $N = N_1N_2$ with N_1 and N_2 relatively prime, one has $\Gamma_0/\Gamma_N \simeq \Gamma_0/\Gamma_{N_1} \times \Gamma_0/\Gamma_{N_2}$. Hence it suffices to consider the case in which N is a prime power. In the case in which N is actually a prime p, Γ_0/Γ_N is isomorphic to the group of all 2×2 matrices of determinant one with coefficients in the field of p elements. Its representations are relatively easy to find and were described by Frobenius in 1896. For higher powers of p, the problem is more difficult and was not completely solved until very recently. Hecke confined himself to the case of prime N. The representations of Γ_0/Γ_{p^2} were determined in the 1933 thesis of Hecke's student Praetorius and independently by Rohrbach a year earlier.

20. IDÈLES, ADÈLES, AND APPLICATIONS OF PONTRJAGIN-VAN KAMPEN DUALITY TO NUMBER THEORY, CONNECTIONS WITH ALMOST-PERIODIC FUNCTIONS, AND THE WORK OF HARDY AND LITTLEWOOD

The representation theory of finite groups and compact Lie groups on the one hand, and the Pontrjagin-van Kampen duality theory on the other, constitute two rather different generalizations of the harmonic analysis of the nineteenth century. The first of these began to have applications to number theory and to physics in the middle 1920s;these applications have been discussed in sections 16 and 19. Applications of the Pontrjagin-van Kampen duality theorem began in 1936 with the introduction by Chevalley (1909—) of the concept of an *idèle* and the idèle group of an algebraic number field. The idèle concept is based in turn on the notion of a p-adic number introduced in 1901 by Hensel (1861-1941). If p is any prime, the p-*adic distance* $\varrho_p(r_1, r_2)$ between two rational numbers r_1 and r_2 is defined to be p^{-k} where k is determined by the relationship $r_1 - r_2 = \frac{m}{n} p^k$, m and n being integers not divisible by p. The p-adic numbers are then the elements of the field Q_p obtained by completing the rational field Q with respect to the p-adic distance just as the real field is obtained by completing Q with respect to the distance $\varrho_\infty(r_1, r_2) = |r_1 - r_2|$. It turns out that every p-adic number can be written uniquely in the form $p^n(a_0 + a_1p + a_2p^2 + \cdots)$, where each $a_j = 0, 1, 2, \cdots, p-1$, $a_0 \neq 0$, and that every sequence of such a_j's occurs. Those p-adic numbers for which $n \geq 0$ are called p-*adic integers*. They form

a compact open subring of the field of all p-adic numbers which is accordingly locally compact. The additive group of Q_p and the multiplicative group Q_p^* of all non-zero elements of Q_p are both totally-disconnected, locally-compact commutative groups. Moreover, the additive group is isomorphic to its own dual. In terms of Q_p^* it is quite easy to complete the definition of the norm residue symbol of Hilbert mentioned in the last section. If the relevant quadratic extension field $\mathcal{F}$ of Q is generated by $\sqrt{D}$, it can be shown that the set of all $x^2 - y^2D$ is a subgroup of Q_p^* of index 2 and one defines $\chi_p^{\mathcal{F}}$ to be the restriction to Q^* of the unique character of Q_p^* which is 1 on the subgroup and -1 otherwise. For fixed p, the $\chi_p^{\mathcal{F}}$ for different quadratic extension fields $\mathcal{F}$ are precisely those characters of Q^* of order two which are continuous in the p-adic topology. Those p-adic integers which are units in the ring of all p-adic integers, that is those of the form $a_0 + a_1p + \cdots$ with $a_0 \neq 0$, are called the p-*adic units*. They form a compact open subgroup U_p of the group Q_p^*, and the quotient group is the infinite cyclic group.

For the special case of the rational number field Q, Chevalley's idèles are the members of a certain subgroup I of the infinite product $\left(\prod_p Q_p^*\right) \times R^*$. This subgroup consists of all members $\{x_p\}, x$ of this product group such that $x_p \in U_p$ for all but a finite number of primes p. While this entire infinite product group is not itself a locally-compact group in any simple or natural way, the subgroup of idèles can be given a simple locally-compact topology. Quite generally, let $G_1, G_2, \cdots$ be any sequence of locally-compact groups, each member G_j of which admits a compact open subgroup K_j. Let G be the subgroup of $\prod_j G_j$ consisting of all sequences $x_1, x_2, \cdots$ with $x_j \in K_j$ for all but finitely many indices j. Then G contains the compact product group $K = \prod_j K_j$ as a subgroup and there are only countably many K cosets. Defining a subset $\mathcal{O}$ of G to be open whenever its intersection with each right K coset is open, one obtains a topology in G which converts it into a locally-compact topological group in which K is a compact open subgroup. The group G is called the restricted direct product of the G_j with respect to the K_j. The idèle group is the direct product of R^* and the restricted direct product of the Q_p^* with respect to the U_p, and as such has a locally compact topology. The subgroup $\prod_p U_p$ is compact in this topology and $\prod_p U_p \times R^*$ is open.

Since Q^* has a natural dense imbedding in each Q_p^* as well as in R^*, one has a natural imbedding of Q^* into the full product group $\Pi\, Q_p^* \times R^*$. Moreover, it is easy to see that for each $r \in Q^*$, r as an element of Q_p^* is in U_p for all but finitely many p. Hence the image of r in $\Pi\, Q_p^* \times R^*$ is actually in the idele group I. In other words, one has a natural imbedding of Q^* as

a subgroup I_0 of the idèle group I, and the members of this subgroup I_0 are called the *principal idèles*. It turns out to be possible to prove that the subgroup I_0 of all principal ideles is a closed subgroup so that the quotient group I/I_0 — the so-called idele class group — also has a natural locally-compact topology.

The significance of the idèle class group can be most easily appreciated by considering its characters of order two and confronting their determination with Hilbert's formulation of the quadratic reciprocity law in terms of his norm residue symbol. A character of the idèle class group is of course a character of the idèle group which is identically one on I_0. A character of the idèle group is uniquely determined by a system $\{\chi_p\},\chi_\infty$ where χ_p is a character on Q_p^* and χ_∞ is a character on R^*. Not every system occurs, however. One sees easily that a system $\{\chi_p\}$, χ_∞ arises from some character of I if and only if for all but finitely many p, $\chi_p(u) = 1$ for all $u \in U_p$. Remembering that every character of Q_p^* defines a character on Q^* which determines it uniquely, we see that the characters on I correspond one-to-one to certain systems $\{\chi_p\}$, χ_∞ of characters on Q^*, and that the condition that such a system satisfies Hilbert's criteria (see section 19) for being associated with a quadratic extension field is precisely that it defines a character of order two on I which is identically one on I_0. In other words, Hilbert's description of all possible quadratic extension fields in terms of systems of characters of order 2 on Q^* may be reformulated as the statement that they correspond one-to-one in a natural way to the characters of order two on the idèle class group I/I_0. Of course, characters of order two correspond one-to-one to closed subgroups of index two, and more generally the Abelian extension fields of Q of finite degree correspond one-to-one to the closed subgroups of I/I_0 of finite index.

Now consider the dual $\widehat{(I/I_0)}$ of I/I_0. Its subgroups of finite order correspond one-to-one to the closed subgroups of finite index of I/I_0 and in fact are the duals of the corresponding quotient groups. Moreover, it follows from the Artin reciprocity law that each quotient group is canonically isomorphic to the Galois group of the corresponding extension field. Thus the subgroups of finite order of $\widehat{(I/I_0)}$ are the duals of the Galois groups of the finite Abelian extensions of Q. More generally, one can consider the infinite extension fields generated by countable sets of finite Abelian extensions of Q, including the maximal one consisting of all algebraic numbers. They correspond one-to-one to the infinite subgroups of the group $\widehat{(I/I_0)}_f$ of all elements of finite order of $\widehat{(I/I_0)}$. Their Galois groups may be identified with the duals of these infinite subgroups and so given a compact, totally disconnected topology. The group $\widehat{(I/I_0)}_f$ itself is the dual of the quotient of I/I_0 by its connected component. Thus the quotient of I/I_0 by its connected component is a totally disconnected compact commutative group which can be identified with the Galois group of the maximal Abelian extension of Q

and whose closed subgroups correspond one-to-one to the finite and infinite Abelian extension fields of Q. As indicated by the title of his 1936 paper, *"Généralisations de la théorie du corps de classes pour les extensions infinies,"* Chevalley's original motivation in introducing idèle groups was to have a method of describing infinite Abelian extension fields analogous to that of Hilbert and Takagi for the finite ones. For this purpose, idèles were indispensible. It turned out, however, that the idèle group notion simplified the finite theory as well, and in 1940 Chevalley published a paper redoing the whole of class field theory in terms of idèles. In his thesis of 1933, Chevalley had simplified class field theory in other ways, and his 1940 paper blended the two kinds of simplification. In particular, he replaced many complicated arguments using Dirichlet series and complex analysis with simpler arguments involving the theory of topological groups. For the sake of simplicity we have defined the idèle group and the idèle class group only for the rational field Q. However, Chevalley dealt with the general case in which Q is replaced by an arbitrary algebraic number field $\mathfrak{F}$ and the Q_p by the completions of $\mathfrak{F}$ with respect to metrics defined by the prime ideals in the ring of integers of $\mathfrak{F}$.

An additive analogue of the idèle group of an algebraic number field was introduced in 1945 by Artin and Whaples (1914—). It is defined as a restricted direct product group over the prime ideals and the "infinite primes" just as in the idele case. But now the additive groups of the completed fields are used in place of the multiplicative ones, and the compact open subgroups with respect to which the restricted product is taken are not the unit groups but the closures in the $\mathfrak{P}$-adic topologies of the integers of the field. The members of this additive infinite "product" group were originally called *valuation vectors*, but are now usually called adèles. Adèles can be multiplied together as well as added; they form a ring (the adèle ring of the field) under these two operations. The idèle group is precisely the group of units of the adèle ring, but its topology as a subset of the adèle ring is not the same as its topology as an idèle group.

Artin and Whaples introduced adèles as a tool in giving an axiomatic characterization of algebraic number fields. In the course of doing so, they went further than Chevalley had in demonstrating the utility of idèles in formulating and proving the facts of algebraic number theory. Actually, their characterization of algebraic number fields was included in a characterization of a parallel class of fields having prime characteristic. Let p be a prime and let Z_p denote the finite field of p elements. Then the field $Z_p(x)$ of all rational functions with coefficients in Z_p is a countable field, which is in an obvious sense the simplest infinite field of characteristic p. As such, it is a characteristic p analogue of the rational field Q, and one can develop a theory of the finite algebraic extensions of $Z_p(x)$ which is quite analogous to the theory of algebraic number fields. Artin laid the foundations for such a the-

ory in his 1921 Ph.D. thesis. This thesis, published in 1924, dealt with the quadratic extensions of the fields $Z_p(x)$. Artin and Whaples showed that any field satisfying certain simple axioms was necessarily a finite algebraic extension of either Q or $Z_p(x)$.

In his 1950 Ph.D. thesis, Tate (1925—), carrying out a suggestion of Artin, showed how to use harmonic analysis in adèle groups to prove a vast generalization of the well-known functional equation for the Riemann zeta function. An abstract announcing similar results was published by Iwasawa (1917—) in the Proceedings of the 1950 International Mathematical Congress. Tate's thesis was not published until 1967, when it appeared in the Proceedings of the 1965 Brighton Conference on algebraic number theory. However, copies of it were privately circulated long before this and it also appeared in rewritten form as a chapter in a book by Lang. The fact that the zeta function of an *arbitrary* algebraic number field has an analytic continuation and satisfies a functional equation analogous to that satisfied by the Riemann zeta function was first proved by Hecke in a paper published in 1917. Various special cases had been treated earlier by other authors. In the same year Hecke published a second paper doing the same thing for general Dirichlet L functions. A bit later, he introduced a more general kind of L function determined by an algebraic number field and a so-called *Grössencharakter* for the field. He studied these L functions in papers published in 1918 and 1920, and proved that they too satisfy (rather complicated) Riemann-type functional equations. The method of Tate and Iwasawa made it possible to obtain all of these results of Hecke at one stroke by applying a generalization of the Poisson summation formula. The classical Poisson summation formula asserts that, for reasonably general complex-valued functions on the line, $\sum_{n=-\infty}^{\infty} f(n) = \sum_{n=-\infty}^{\infty} \hat{f}(n)$, where $\hat{f}(x) = \int_{-\infty}^{\infty} f(y)e^{2\pi ixy}dy$. More generally, if G is any separable locally-compact commutative group and Γ is any countable closed subgroup such that G/Γ is compact, then $\Gamma^{\perp}$, the subgroup of the character group $\hat{G}$ consisting of all characters χ with $\chi(\gamma) = 1$ for all $\gamma \in \Gamma$, is also closed and countable, and when the Haar measure μ in G is suitably normalized, one has

$$\sum_{\gamma\in\Gamma} f(\gamma) = \sum_{\chi\in\Gamma^{\perp}} \hat{f}(\chi)$$

for all suitably restricted functions f on G. Here $\hat{f}(\chi) = \int_G f(x)\chi(x)d\mu(x)$. Tate and Iwasawa take the adèle group of the number field for G and the subgroup of principal adèles for Γ. Hecke's *Grössencharaktere*, whose original definition was rather complicated, can be defined much more simply using idèles. They are just the characters of the idèle class group I/I_0 that are not of finite order. Hecke's proof also hinged on the Poisson sum-

mation formula, which he used to prove a generalized form of Jacobi's inversion formula. The difference between Hecke and Tate is that Hecke does his computations "at infinity," i.e., over the Archimedean primes, whereas Tate works over all primes simultaneously.

There is another way of defining the adèle group of a number field which is based directly on group duality and makes no use of p-adic metrics. Let R be the ring of all algebraic integers in the algebraic number field $\mathcal{F}$ and let R^+ be the additive group of R. R^+ is countable, and we make it into a locally-compact commutative group by giving it the discrete topology. The dual $\hat{R}^+$ is then compact and in fact is isomorphic to the direct product of n replicas of the circle group T where n is the degree of $\mathcal{F}$ over the rationals. Let $\hat{R}^{+f}$ denote the subgroup of $\hat{R}^+$ consisting of all elements of finite order. Then $\hat{R}^{+f}$ is a dense countable subgroup of $\hat{R}^+$ and we may regard its natural imbedding as an injective homomorphism θ of the discrete group $\hat{R}^{+f}$ into $\hat{R}^+$. Its dual θ^* (see section 19) is then an injective homomorphism of $\hat{\hat{R}}^+ = R^+$ onto a dense subgroup of the compact totally disconnected dual $\hat{\hat{R}}^{+f}$ of $\hat{R}^{+f}$. One of the easy general theorems about group duality asserts that a compact group is infinitely divisible if and only if its dual has no elements of finite order, and that it has no elements of finite order if and only if its dual is infinitely divisible. Since $\hat{R}^{+f}$ is infinitely divisible, it follows at once that $\hat{\hat{R}}^{+f}$ has no elements of finite order. Since it has no elements of finite order, it has a unique minimal completely divisible extension in which it has countable index $(\hat{\hat{R}}^{+f})^{\sim}$. This extension may be made into a locally-compact group by giving each $\hat{\hat{R}}^{+f}$ coset the topology of $\hat{\hat{R}}^{+f}$ and declaring a subset of the divisible extension to be open if its intersection with each coset is open. This locally-compact group has the additive group $\mathcal{F}^+$ of $\mathcal{F}$ densely imbedded, and is the so-called non-Archimedean component of the adèle group of $\mathcal{F}$. The actual adèle group is the direct product of this locally-compact group with an n-dimensional vector space over the real numbers called the Archimedean component of the adèle group. The latter can be defined in a manner vaguely analogous to that used in defining the non-Archimedean component. Let $\overline{R^+}$ denote the vector space of all homomorphisms of R^+ into the multiplicative group of all positive real numbers. Then the Archimedean component of the adèle group is the vector space dual of $\overline{R^+}$. It is an n-dimensional real vector space containing R^+ as a lattice subgroup and $\mathcal{F}^+$ as a dense subgroup. The dense φ_1 and φ_2 imbeddings of $\mathcal{F}^+$ into $(\hat{\hat{R}}^{+f})^{\sim}$ and $\overline{\hat{R}^+}$ may be combined to give an imbedding $f \to \varphi_1(f), \varphi_2(f)$ of $\mathcal{F}^+$ into the product group, i.e., into the adele group of $\mathcal{F}$. The range of this imbedding can be shown to be closed and is the group of all principal adeles.

It is interesting to examine the results of Hardy and Littlewood on Waring's problem (see section 14) in the light of the theory of almost-periodic functions (see section 19) and its connection with group duality. Choose fixed positive integers k and r. For each positive integer n, let $f(n)$ denote the

number of integer solutions of $x_1^k + \cdots + x_r^k = n$. Then $f(1) + f(2) + \cdots + f(n)$ is the number of points with integer coordinates inside and on the hypersurface $x_1^k + x_2^k + \cdots x_r^k = n$. Elementary arguments show accordingly that $(f(1) + f(2) + \cdots + f(n))$ is asymptotic to a constant multiple of $n^{r/k}$. It follows easily from this that if $f_0(n) = \frac{f(n)}{n^{r/k-1}}$, then $f_0(1) + f_0(2) + \cdots f_0(n)$ is asymptotic to a constant multiple of n; that is, that

$$\frac{f_0(1) + \cdots + f_0(n)}{n}$$

has a limit as n tends to ∞. In other words, the function $n \to f_0(n)$ behaves like an almost-periodic function on the integers to the extent that it has a mean value. Of course, the properties of sample functions of stationary stochastic processes (see section 18) warn us that having a mean value is far from implying almost periodicity. On the other hand, one does not expect f_0 to be like a random function, and even if it were, one could still compute its hidden periods (if any). All of this suggests investigating the existence of the mean of $n \to f(n)e^{-in\lambda}$ for the various real values of λ, and in these terms the main results of Hardy and Littlewood on Waring's problem may be summed up as follows: Suppose that $r \geq 2^k(2k + 1)$. Then

(1) $c_\lambda = \lim_{n\to\infty} \frac{1}{n}(f(1)e^{-i\lambda} + f(2)e^{-2i\lambda} + \cdots + f(n)e^{-ni\lambda})$ exists for all real λ, and $c_\lambda = 0$ whenever λ/π is irrational.

(2) If $a_\lambda = c_{2\pi\lambda}$, then the series $\sum_\lambda a_\lambda e^{2\pi i n\lambda}$ (where the sum is over all rational numbers λ) converges for all n to a function $S(n)$ and the convergence is uniform.

(3) $\lim_{n\to\infty} S(n) - f_0(n) = 0$.

(4) If $M(f_0) = \lim_{n\to\infty} \frac{f_0(1) + \cdots + f_0(n)}{n}$ and $\lambda = p/q$ where p and q are relatively prime, then $a_\lambda = a_{p/q} = \frac{1}{M(f_0)} \frac{1}{q^r} \left(\sum_{m=0}^{q-1} e^{2\pi i (m^k)(p/q)^r} \right)$.

One sees in particular that $f(n)$ is asymptotic to $\frac{n^{(r/k-1)}}{M(f_0)} S_0(n)$ where $S_0(n) = S(n)M(f_0)$, and is an almost-periodic function of n with explicitly known Bohr-Fourier coefficients. The function $S_0(n)$, or rather the Bohr-Fourier expansion of it, is what Hardy and Littlewood call the "singular series."

Every almost-periodic function on the integers Z (see section 17) may be extended to be continuous on a certain compactification of Z. This compactification is obtained by starting with a countable discrete subgroup Γ of $\hat{Z}$ and taking the dual θ^* of the isomorphic imbedding θ of Γ in $\hat{Z}$. θ^* imbeds Z as a dense subgroup of the compact dual $\hat{\Gamma}$ of Γ. In the case at hand, the

fact that $c_\lambda = 0$ except when λ/π is rational implies that the subgroup Γ of $\hat{Z}$ is precisely the subgroup of all elements of finite order. Thus $\hat{\Gamma}$ is an open compact subgroup in the non-Archimedean component of the adèle group of the rational field. The discrete group Γ has a natural direct product decomposition over the primes. Indeed, for each prime p, the subset Γ_p of all elements γ in Γ with $\gamma^{p^k} = e$ for some k is a subgroup and every element is uniquely a product of members of a finite number of the Γ_p. Correspondingly, the compact dual $\hat{\Gamma}$ is the full direct product of the compact groups $\hat{\Gamma}_p$, and it turns out to be easy to verify that $S_0(n)$ regarded as a function on $\hat{\Gamma}$ factors as a product of functions on the various $\hat{\Gamma}_p$. As mentioned earlier, $\hat{\Gamma}_p$ is isomorphic to the group of all p-adic integers. Moreover, it is not hard to interpret these p-adic components of S_0 in terms of p-adic solutions of the equation $x_1^k + \cdots + x_r^k = n$. Similarly, $\frac{n^{(r/k-1)}}{M(f_0)}$ has an interpretation in terms of real solutions. Quite apart from these interpretations, the product $\frac{n^{(r/k-1)}}{M(f_0)} S_0(n)$ may be looked upon as a function defined on the whole adèle group which factors according to the natural factorization of the adèle group.

The fact that there is a connection between almost-periodic functions and the Hardy-Littlewood results was pointed out (in the special case $k = 2$) by Kac (1914—) in 1940.

21. THE DEVELOPMENT OF THE THEORY OF UNITARY GROUP REPRESENTATIONS AFTER 1945 — A BRIEF SKETCH WITH EMPHASIS ON THE FIRST DECADE

With the exceptions mentioned at the end of section 19, the theory of unitary group representations was until 1946 exclusively concerned with groups that were either compact or both locally-compact and commutative. A more general theory encompassing all locally-compact groups began rather suddenly with the publication in 1947 of four long papers and half a dozen or so short notes and announcements. (Two of the long papers were preceded by short announcements published in 1946.) Since then, there has been an enormous development, which cannot begin to be summarized within the compass of this article. I shall content myself instead with some brief indications and refer the reader at the appropriate time to some lengthy survey articles for a more adequate account.

In attempting to generalize from compact groups to locally-compact groups, one is confronted with two major problems. In the first place, one can no longer decompose representations as discrete direct sums except in very special cases; and in the second place, one has to deal with irreducible unitary representations which are infinite-dimensional. The latter circumstance brings with it a further difficulty in that the trace of an infinite-di-

mensional unitary operator is undefined. This means that the character of an infinite-dimensional irreducible representation must be defined in a roundabout way when it can be defined at all. When the group is commutative, the lack of compactness is partially compensated for by the fact that the irreducible unitary representations are not only finite-dimensional but *one*-dimensional. In this case, one can combine Pontrjagin-van Kampen duality with the ideas of spectral theory to obtain an entirely adequate substitute for the Peter-Weyl theorem. Just how this works has already been explained in section 17.

Thus far I have said relatively little about the problem of actually finding the possible irreducible representations of our groups. This is because the problem is relatively easy in the commutative case, was solved more or less completely for the important compact Lie groups by Weyl in the 1920s, and for finite groups is more a problem in algebra than analysis. However, for groups which are neither compact nor commutative, the problem of finding the possible (usually infinite-dimensional) unitary irreducible representations is one of the main problems of the theory. It involves a heavy use of analysis and is by no means readily solved. All four of the long papers published in 1947 dealt with important special cases of it and so did the majority of the short notes. Another case was the subject of Wigner's 1939 paper mentioned at the end of section 17. Before the results of any of these papers are described, it will be convenient to discuss a method used in most of them which was first discussed in an abstract setting in a paper that I published in 1949. This is a method for constructing unitary representations of locally-compact groups out of unitary representations of closed subgroups which generalizes the Frobenius construction $\chi \to \chi^*$ mapping characters of subgroups of finite groups into characters of the whole group (see section 15). Frobenius found his construction to be a very useful tool in producing irreducible representations and characters, and its generalization has turned out to be equally useful.

Let G be an arbitrary separable locally-compact group. (We may restrict ourselves to the separable case because most if not all of the important examples are separable, and because in so doing we avoid various distracting technical complications.) Let H be a closed subgroup of G and suppose for the time being that the right coset space G/H admits a measure μ which is invariant under the natural action $(Hx)y = Hxy$ of G on G/H. It is not difficult to see that μ, if it exists, is uniquely determined up to a multiplicative constant. Now let L, $x \to L_x$ be any unitary representation of H in a separable Hilbert space $\mathscr{H}(L)$. (We always suppose that $(L_x(\varphi) \cdot \psi)$ is continuous for all φ and ψ in $\mathscr{H}(L)$; this implies that $x \to L_x(\varphi)$ is continuous for all φ in $\mathscr{H}(L)$.) Consider the set $\mathscr{F}_L$ of all Borel functions $x \to f(x)$ from G to the Hilbert space $\mathscr{H}(L)$ which satisfy the identity $f(hx) = L_h(f(x))$ for all h in H and all x in G. Now for each f in $\mathscr{F}_L$, the function $x \to (f(x) \cdot f(x))$ is a

non-negative Borel function on G. Moreover, since $(f(hx) \cdot f(hx)) = (L_h f(x) \cdot L_h f(x)) = (f(x) \cdot f(x))$, it follows that $x \to (f(x) \cdot f(x))$ is a constant on the right H cosets and so may be regarded as a function on G/H. Let $\mathfrak{F}_L^0$ denote the subset of $\mathfrak{F}_L$ consisting of all f in $\mathfrak{F}_L$ for which $\int_{G/H}(f(x) \cdot f(x))d\mu(\overline{x}) < \infty$ where $\overline{x}$ is the image of x on G/H and $(f(x) \cdot f(x))$ is thought of as a function on G/H. It is not hard to show that $\mathfrak{F}_L^0$ becomes a Hilbert space if we define $||f|| = \sqrt{\int_{G/H}(f(x) \cdot f(x))d\mu(\overline{x})}$ and identify two members when they are almost everywhere equal. Moreover, for each x, the mapping $f \to f_x$, where $f_x(y) = f(yx)$, is easily seen to be a unitary operator in the Hilbert space $\mathfrak{F}_L^0$. If we denote this unitary operator by U_x^L, we verify that $x \to U_x^L$ is a unitary representation of G. It is called the *representation of* G *induced by the representation* L *of* H. When G is finite and χ is the character of L, one verifies without difficulty that the character of U^L is precisely the induced character χ^* of G defined by Frobenius. For the case in which G is compact, the definition given here is essentially to be found in Weil's book, cited in section 17. When G/H fails to have an invariant measure, a slightly more complicated definition involving quasi-invariant measures has to be given. I shall not repeat it here, but I assure the reader that U^L is a well-defined unitary representation of G for all unitary representations L of all closed subgroups H of G.

In terms of this definition, it is possible to state a general theorem of which the results of Wigner's 1939 paper as well as those of one of the 1947 notes are both special cases. Let the separable locally-compact group G admit a closed commutative normal subgroup N, and suppose that there exists a second closed subgroup H such that $N \cap H$ contains only the identity and $NH = G$. Then every element of G can be written in one and only one way, as a product nh where n is in N and h is in H. One says that G is a *semi-direct product* of N and H. Each h in H defines an automorphism $n \to hnh^{-1} = \alpha_h$ of N, and the mapping $h \to \alpha_h$ is a homomorphism of H into the group of automorphisms of N. Evidently one can reconstruct G knowing only N, H, and the mapping $h \to \alpha_h$. The general theorem I propose to state reduces the problem of finding the irreducible unitary representations of G to that of finding the irreducible unitary representations of certain subgroups of H — at least when the "adjoint" action of H on the dual of N has a certain regularity property. To explain this property, notice that for each automorphism α_h of N, there is a well-defined adjoint automorphism α_h^* of $\hat{N}$. Indeed, if $\chi \in \hat{N}$ and $h \in H$, then $n \to \chi(\alpha_h(n))$ is also a member of $\hat{N}$ which may be denoted by $[\chi]\alpha_h^*$. It is obvious that $\chi \to [\chi]\alpha_h^*$ is an automorphism and $h \to \alpha_h^*$ is a homomorphism. Let us say that the semi-direct product is *regular* if there exists a Borel set C in $\hat{N}$ which meets each H orbit in $\hat{N}$ in one and only one point; that is, if for each χ in $\hat{N}$ there is one and only one χ^1 in C such that $[\chi]\alpha_h^* = \chi^1$ for some h in H.

Now let G be a regular semi-direct product and let C be a Borel set which

meets each H orbit in C in one point. The general theorem alluded to above states that the equivalence classes of irreducible unitary representations of G may all be obtained as follows: Choose $\chi \in C$. Let H_χ denote the closed subgroup of H consisting of all h in H for which $[\chi]\alpha_h^* = \chi$. Choose an irreducible unitary representation L of H_χ. Then $n, h \to \chi(n)L_h$ is an irreducible unitary representation χL of NH_χ. Form $U^{\chi L}$, the unitary representation of G induced by χL. It can be shown that $U^{\chi L}$ is irreducible, that $U^{\chi_1 L^1}$ and $U^{\chi_2 L^2}$ are equivalent if and only if $\chi_1 = \chi_2$ and $L^1 = L^2$ are equivalent representations of H_χ and that every irreducible unitary representation of G is equivalent to some $U^{\chi L}$. When the semi-direct product is not regular, one can use the axiom of choice to find a subset C which meets each orbit just once, but C will not be a Borel set. In this case, the $U^{\chi L}$ can still be formed and proved to be irreducible and inequivalent as indicated. However, it is no longer true that every irreducible unitary representation of G is equivalent to one of the $U^{\chi L}$.

The inhomogeneous Lorentz group considered by Wigner in 1939 is a regular semi-direct product of a four-dimensional real vector group and the Lorentz group, the latter being isomorphic to the quotient of $SL(2, C)$ by its two-element center. Actually, Wigner studied the irreducible unitary representations of the two-fold covering group one gets by using the whole of $SL(2, C)$. His results are a restatement of what one finds by applying the theorem above. It turns out that, depending on the position of the character χ with respect to the "light cone," there are four possibilities for the subgroup H_χ. It is conjugate either to a) the compact subgroup of all unitary matrices, b) a non-compact subgroup isomorphic to the group generated by the translations and rotations in the plane, c) the subgroup $SL(2, R)$ of all matrices in $SL(2, C)$ with real coefficients, or d) the whole of $SL(2, C)$. In case a, the irreducible representations of H_χ are known from the work of Schur and Weyl. In case b, H_χ is a semi-direct product of two commutative groups, and the general theorem just cited can be applied. In cases c and d, H_χ is a non-compact semi-simple Lie group. At the time Wigner's paper was written, nothing was known about their irreducible representations. Wigner, in fact, determined only those irreducible unitary representations of the inhomogeneous Lorentz group falling under cases a and b. However, he gave cogent arguments suggesting that the others were not relevant to the physical applications he had in mind.

A much simpler example of a regular semi-direct product was dealt with in one of the short notes published in 1947. In this note, Gelfand (1913—) and Naimark (1909—) determined all irreducible unitary representations of the group of all one-to-one transformations of the real line into itself of the form $x \to ax + b$ where $a > 0$. This group, often referred to as the "$ax + b$ group," is a semi-direct product of the additive group of all real numbers with the multiplicative group of all positive real numbers. Here N is the ad-

ditive group of the real line and there are just three orbits. These are $\{0\}$ and the positive and negative real axes. H is the multiplicative group of all positive real numbers and H_x is respectively H, $\{1\}$, and $\{1\}$. It follows that the "$ax + b$ group" has (to within equivalence) just two irreducible unitary representations in addition to the obvious one-dimensional representations defined by the characters of H. They are the representations induced by the characters $b \to e^{ib}$ and $b \to e^{-ib}$ of N. Both are infinite-dimensional.

Three of the four long papers published in 1947 were written respectively by Gelfand and Naimark, Bargmann (1908—), and Harish-Chandra (1923—). All of them were concerned with the determination of the unitary representations of $SL(2, C)$ and hence of the Lorentz group. Bargmann and Gelfand and Naimark also treated $SL(2, R)$, but only Bargmann gave a detailed analysis in that case. As a matter of fact, the papers of Bargmann and Gelfand and Naimark are complementary, in that Bargmann gave details only for $SL(2, R)$ and Gelfand and Naimark only for $SL(2, C)$. Harish-Chandra contented himself with determining the representations of the Lie algebra of $SL(2, C)$, while Bargmann and Gelfand and Naimark found the integrated form of the representations and also discussed the decomposition of the regular representation. The facts about the irreducible unitary representations of $SL(2, C)$ are easily stated in terms of the general concept of an induced representation. Let T be the subgroup of $SL(2, C)$ consisting of all matrices of the form $\begin{pmatrix} \lambda & 0 \\ a & 1/\lambda \end{pmatrix}$. Notice that $\begin{pmatrix} \lambda & 0 \\ a & 1/\lambda \end{pmatrix} \to \begin{pmatrix} \lambda & 0 \\ 0 & 1/\lambda \end{pmatrix}$ is a homomorphism of T onto the subgroup D of all diagonal matrices. Thus each (one-dimensional) character χ of the commutative group D may be lifted to define a one-dimensional representation χ^1 of T. Gelfand and Naimark showed that the induced representations U^{χ^1} are all irreducible and that $U^{\chi_1^1}$ and $U^{\chi_2^1}$ are equivalent if and only if $\chi_1 = \chi_2$ or $\chi_1 = \chi_2^{-1}$. They showed also that the representations U^{χ^1} constitute "almost all" irreducible unitary representations in the sense that they suffice for the decomposition of the regular representation (see below). They constitute what is known as the *principal series* of representations of $SL(2, C)$. In addition to the principal series, they described a second infinite series of irreducible unitary representations called the *supplementary series*. It turns out that one can modify the inducing process[7] in such a manner that some non-unitary representations induce unitary representations of the whole group. This is true of certain non-unitary one-dimensional representations of T, and the members of the supplementary series can all be so described. Gelfand and Naimark were able to prove that every irreducible unitary representation (except of course the trivial representation) is equivalent either to a member of the principal series or to a member of the supplementary series.

In the study of $SL(2, R)$, an interesting new phenomenon arises. One has an obvious analogue of the subgroup T of the principal series and of the

supplementary series, but the irreducible unitary representations of $SL(2, R)$ so defined do not exhaust all equivalence classes and do not even suffice to decompose the regular representation. They have to be supplemented by the members of a third series. It is called the *discrete series* because the representations which belong to it occur discretely in the decomposition of the regular representation. Let K be the compact commutative subgroup of $SL(2, R)$ consisting of all matrices of the form $\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$, and for each integer k let χ_k denote the character $\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \to e^{ik\theta}$. Then the induced representations U^{χ_k} are reducible but contain $|k|$ inequivalent discrete irreducible subrepresentations. Moreover, if $k > 0$, U^{χ_k} contains exactly one discrete component which is not contained in $U^{\chi_{k-1}}$ and if $k < 0$, U^{χ_k} contains exactly one discrete component which is not contained in $U^{\chi_{k+1}}$. In either case, let us denote this irreducible representation by W^k. Then the W^k are mutually inequivalent and constitute the discrete series of irreducible unitary representations of $SL(2, R)$. Notice that there is a natural one-to-one correspondence between the members of the discrete series and the non-trivial characters of K, and also a natural one-to-one correspondence between the members of the principal series and pairs χ, χ^{-1} of characters of the diagonal subgroup D (excluding the two cases in which $\chi = \chi^{-1}$). This, combined with the fact that D and K are both maximal Abelian subgroups, turns out to be highly significant. I shall say more about this significance below.

There is an alternative description of the discrete series more closely related to that originally given by Bargmann and important for applications to the theory of modular forms. Consider the action of $SL(2, R)$ on the upper half H^+ of the complex plane defined by $[z]\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \frac{az+b}{cz+d}$, and let $(V^k_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)} f)(z) = f(\frac{az+b}{cz+d})(cz + d)^{-k}$ where $k = \pm 1, \pm 2, \cdots$. Then there is a measure μ on the upper half-plane — unique up to multiplication by a positive constant such that the $V^k_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)}$ are unitary in $\mathcal{L}^2(H^+, \mu)$. The unitary representation of $SL(2, R)$ in $\mathcal{L}^2(H^+, \mu)$ so defined is easily seen to be equivalent to the induced representation U^{χ_k}. The subspace of $\mathcal{L}^2(H^+, \mu)$ on which U^{χ_k} restricts to W^k is the subspace of all analytic functions or the subspace of all conjugate analytic functions, according as $k > 0$ or $k < 0$.

The determination of the irreducible unitary representations of $SL(2, C)$ and $SL(2, R)$ completed Wigner's 1939 paper very nicely. However, it was only Bargmann whose work was inspired by that of Wigner. Neither Harish-Chandra nor Gelfand and Naimark cite Wigner and presumably were unaware of the relevance of his work to theirs. Instead they cite a paper of Dirac, published in 1945. In this paper, Dirac had pointed out that the Lorentz group (and hence $SL(2, C)$) had infinite-dimensional unitary representations and had suggested that they might have physical relevance. Harish-

Chandra was a student of Dirac and made his investigations at Dirac's suggestion.

In view of the role played by physics in inspiring this work on the unitary irreducible representations of $SL(2, R)$ and $SL(2, C)$, it is interesting to note that I proved the theorem on the unitary irreducible representations of regular semi-direct products as a corollary to a much more general theorem obtained as the end product of a three-stage generalization of the Stone-von Neumann theorem on the uniqueness of the solutions of the Heisenberg commutation relations (see section 16). Let G be a separable locally-compact commutative group and let $\hat{G}$ denote its dual. Let μ be Haar measure in G and let A be the regular representation of $G : A_x(f)(y) = f(yx)$ for all $f \in \mathcal{L}^2(G, \mu)$. For each χ in $\hat{G}$ let B_χ denote the unitary operator $f \to \chi f$. Then $\chi \to B_\chi$ is a unitary representation of $\hat{G}$ (equivalent in fact to the regular representation of $\hat{G}$). Moreover, an obvious calculation shows that A and B satisfy the simple commutation relation $A_x B_\chi = \chi(x) B_\chi A_x$. But when G is the additive group of an n-dimensional real vector space, then so is $\hat{G}$, and these commutation relations reduce precisely to the Heisenberg commutation relations in the "integrated" or Weyl form as described in section 16. It is natural to ask whether there is an analogous uniqueness theorem in the general case and to approach the question by applying the spectral theorem to the unitary representation B. The spectral theorem says that B is uniquely determined by a projection-valued measure P on $\hat{\hat{G}} = G$. A simple calculation shows that A and B satisfy the generalized Heisenberg commutation relations written down above if and only if P and A satisfy the commutation relations $A_x P_E = P_{[E]x^{-1}} A_x$ for all x in G and all Borel subsets E of G. These transformed commutation relations have the interesting property that they refer only to G and not to $\hat{G}$. Moreover, they make sense whether or not G is commutative. Indeed, if G is any separable locally-compact group and μ is a right invariant Haar measure, one obtains a system satisfying the transformed commutation relations by defining A^0_x and P^0_E in $\mathcal{L}^2(G, \mu)$ by the equations $A^0_x(f)(y) = f(yx)$ and $P^0_E(f)(y) = \varphi_E(y) f(y)$. Here $\varphi_E(y) = 1$ if $y \in E$ and $\varphi_E(y) = 0$ if $y \notin E$. The question then arises, whether every irreducible pair A, P consisting of a unitary representation A of G and a projection-valued measure P on G and which satisfies $A_x P_E = P_{[E]x^{-1}} A_x$ is necessarily equivalent to the pair A^0, P^0 defined above. It is, and I published a proof of this generalization of the Stone-von Neumann uniqueness theorem in 1949. Still a further generalization is possible, however. The commutation relation $A_x P_E = P_{[E]x^{-1}} A_x$ makes sense even when P is not defined on G. It is only necessary that P be defined on some Borel space on which G acts as a group of one-to-one Borel set-preserving transformations. In this generality, while uniqueness fails, one has a complete analysis of all possibilities — at least in the case in which S is a coset space G/H. The irreducible solutions of the commutation relations $A_x P_E =$

$P_{[E]x^{-1}}A_x$ (more precisely their equivalence classes) correspond one-to-one in a natural way to the equivalence classes of irreducible unitary representations of H. Of course when $H = \{e\}$ so that $S = G$, H has only one equivalence class of irreducible unitary representations and the solutions of the commutation relations are unique. The correspondence between solutions of the commutation relations and unitary representations of H is set up by the inducing construction and does not involve irreducibility. Given any unitary representation L of H, the Hilbert space $\mathscr{H}(U^L)$ consists of functions f in G which satisfy the identity $f(hx) = L_h f(x)$ for all h in H and all x in G. If E is any Borel subset of G/H, the function $P^L_E(f)$, which one obtains by reducing f to zero on all right cosets in E and leaving it unchanged outside of these cosets, is also clearly in $\mathscr{H}(U^L)$ and the operator $f \to P^L_E(f)$ is a projection operator. It is easy to check that $E \to P^L_E$ is a projection-valued measure on G/H and that U^L and P^L satisfy the commutation relations $U^L_x P^L_E = P^L_{[E]x^{-1}} U^L_x$ for all x in G and all Borel sets $E \subseteq G/H$. In a short note published in 1949, I sketched a proof of the converse theorem. Given any pair A, P satisfying the commutation relations in question, there exists a unitary representation L of H (uniquely determined up to equivalence) such that the pair A, P is equivalent to the pair U^L, P^L. Moreover, the algebra of all bounded operators which commute with all L_h is isomorphic to the algebra of all bounded operators which commute with all U^L_x and all P^L_E.

To gain some insight into why such a theorem might be true and also to understand why it is called "the imprimitivity theorem" it is useful to consider the special case in which H is an open subgroup of G. In that case, the projection-valued measure P defines a discrete direct sum decomposition of $\mathscr{H}(A)$ whose summands are parameterized by the points of $G/H = S$. These summands are not invariant under the A_x. This would be true if and only if $A_x P_E = P_E A_x$ for all x and E. On the other hand, the condition $A_x P_E = P_{[E]x^{-1}} A_x$, which does hold, is precisely equivalent to the assertion that the A_x permute the summands among themselves and in particular that A_x carries the summand whose parameter is s onto that whose parameter is $[s]x$. A direct sum decomposition of a Hilbert space having this property with respect to a unitary representation A taking place therein is called a *system of imprimitivity* for the representation. When the group acts transitively on the subspaces and H is the subgroup leaving a subspace $\mathscr{M}_0$ fixed, the original representation defines a representation L of H in $\mathscr{M}_0$ which evidently determines A. The discrete case of the imprimitivity theorem thus has a trivial proof. For finite groups it was known to Frobenius.

To see how the semi-direct product theorem might be a consequence of the imprimitivity theorem one has only to note a) that a unitary representation V of a semi-direct product NH is uniquely determined by its restrictions B and A to N and H respectively, and b) that the condition that $n, h \to B_n A_h$

be a representation of G, given that B and A are representations of N and H respectively, is easily computed to be the commutation relation $B_n A_h = A_h B_{\alpha_h(n)}$. When B is replaced by the projection-valued measure P on $\hat{N}$ which determines it, this commutation relation reduces to the statement that P is a system of imprimitivity for A. When V is irreducible and the semi-direct product is regular, one shows that P is supported by an H orbit in $\hat{N}$. The rest is calculation.

The work of Harish-Chandra, Gelfand and Naimark, and Bargmann on the irreducible unitary representations of $SL(2, C)$ and $SL(2, R)$ has implications beyond its possible relevance to physics and to completing the work of Wigner on the inhomogeneous Lorentz group. The group $SL(2, R)$ is the non-compact semi-simple Lie group of lowest possible dimension, and the group $SL(2, C)$ is the non-compact semi-simple Lie group of lowest possible dimension which also has the structure of a complex manifold. Moreover, while the semi-direct product theorem described above can be generalized to a theorem dealing with groups having non-commutative normal subgroups, this method of analysis fails completely in dealing with simple groups, that is, with groups having no closed normal subgroups. A different approach must be used and $SL(2, R)$ and $SL(2, C)$ are in different ways the most elementary examples. Modulo its two element center, each of these groups is simple.

With these two groups under control, it was natural to go on to more complicated cases, and Gelfand and Naimark began a study of $SL(n, C)$ almost immediately. In fact, the fourth long paper of 1947 and several of the short notes published in 1946 and 1947 are papers by them concerned with the irreducible unitary representations of $SL(n, C)$. The results are analogous to those for $SL(2, C)$ but are less complete. Let T be the subgroup of all matrices that are zero above the main diagonal, and let D be the subgroup of all diagonal matrices. Then $D \subseteq T$, and just as in the case of $SL(n, C)$ there is a natural homomorphism of T onto D so that every $\chi \in \hat{D}$ can be lifted to be a one-dimensional unitary representation χ^1 of T. The induced representations U^{χ^1} are all irreducible and constitute what Gelfand and Naimark call the *principal series*. As in the case of $SL(2, C)$, the members of the principal series suffice to decompose the regular representation. The theory of $SL(n, C)$ differs from that of $SL(2, C)$ chiefly in that finding the remaining irreducible unitary representations is much more difficult. In fact, except when $n \leq 3$, the problem of determining all equivalence classes of irreducible unitary representations of $SL(n, C)$ is still an open one in which there is considerable current interest. When $n > 2$ there exist proper closed subgroups of $SL(n, C)$ which properly contain T. For example, when $n = 3$, one has the subgroup of all matrices of the form $\begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$, and

when $n = 4$, the subgroup of all matrices of the form $\begin{pmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{22} & a_{23} & 0 & 0 \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}$. The subgroups admit non-trivial one-dimensional characters which can be induced up to $SL(n, C)$ to define new infinite-dimensional irreducible unitary representations — the members of the so-called *degenerate series.* As with T in $SL(2, C)$, certain non-unitary characters of both T and the proper closed subgroups containing it can be "induced" to form irreducible unitary representations of $SL(n, C)$. The chief difficulty in determining *all* irreducible unitary representations of $SL(n, C)$ lies in deciding just which non-unitary characters lead to unitary representations.

Gelfand and Naimark found little difficulty in extending their analysis to the other classical complex groups — the complex orthogonal groups and the complex symplectic groups of all dimensions. All of these groups admit analogues of the groups T and D, and one defines the principal and other series in a strictly analogous manner. While it has only recently been proved that *all* members of the principal series for the complex groups are irreducible, Gelfand and Naimark could prove that almost all of them are, and these suffice to decompose the regular representation. Gelfand and Naimark published a book in 1950 giving a systematic account of their work on all the complex classical semi-simple Lie groups including $SL(n, C)$.

Once one has found the irreducible unitary representations of a group G, the problem arises whether or not they are adequate for harmonic analysis on the measure spaces on which G acts. Given a space S, with a measure μ invariant under the G action, one can form a unitary representation U of G whose space is $\mathscr{L}^2(S, \mu)$ by setting $U_x(f)(s) = f([s])x)$. One can then ask whether there is a sense in which this representation can be decomposed as a discrete or continuous direct sum of irreducibles and whether this decomposition is unique in any useful sense. When G is commutative as well as separable and locally compact, these questions are taken care of very nicely by the spectral theorem of Stone, Ambrose, Godement, and Naimark and the spectral multiplicity theory of Hahn, Hellinger, and Stone (see section 17). To go further, one needed a theory of direct integrals or continous direct sums of Hilbert spaces. Such a theory had been worked out by von Neumann in the 1930s for use in his work with Murray on algebras of operators (see the end of section 17), and a more or less complete typed paper on the subject was in von Neumann's possession in 1938. However, this paper did not get published until 1949 and apparently no one interested in the theory of unitary group representations knew its contents until 1947. In that year von Neumann made the typescript available to Mautner (1921—), who based his 1948 Ph.D. thesis on applying direct integral decompositions to unitary group representations. I used some of the ideas

in von Neumann's typescript in my proof of the generalized Stone-von Neumann uniqueness theorem. Mautner's first main result (announced in a short note published in 1948) was in essence as follows: Let U be a (continuous) unitary representation of a separable locally-compact group G in a separable Hilbert space $\mathscr{H}(U)$. Let $R(U, U)$ denote the commuting algebra of U; that is, the algebra of all bounded linear operators in $\mathscr{H}(U)$ that commute with all U_x. Then corresponding to every *maximal* commutative subalgebra of $R(U, U)$ there exists an essentially unique decomposition of U as a direct integral of *irreducible* unitary representations of G. The meaning of direct integral decomposition can be most quickly explained in the case in which all of the component representations are infinite-dimensional and so have isomorphic Hilbert spaces. Let there be given a (suitably restricted) measure space S, μ, and for each $s \in S$ a unitary irreducible representation L^s of G in a fixed Hilbert space $\mathscr{H}$. Suppose that $L^s_x(\varphi) \cdot \psi$ is measurable in $S \times G$ for each φ and ψ in $\mathscr{H}$. Form the Hilbert space $\mathscr{L}^2(S, \mu, \mathscr{H})$ of all square integrable measurable functions from S to $\mathscr{H}$. For each x in G let M_x denote the operator in $\mathscr{L}^2(S, \mu, \mathscr{H})$ which takes $s \to f(s)$ into $s \to L^s_x(f(s))$. Then each M_x is unitary and $x \to M_x$ is a (continuous) unitary representation of G. It is called the *direct integral* or *continuous* direct sum of the representations L^s with respect to the measure μ. For each measurable subset E of S, one can associate the operator $f \to \varphi_E f$ where $\varphi_E(s) = 1$ if $s \in E$ and is zero otherwise. Denoting this operator by P_E, one verifies that $E \to P_E$ is a projection-valued measure. When all (or almost all) the L^s are irreducible, it turns out that the P_E constitute all the projection operators in a maximal commuting subalgebra of $R(M, M)$. Conversely, according to Mautner's theorem, every maximal commuting subalgebra of every $R(U, U)$ arises in this way from some direct integral of irreducibles that is equivalent to U. Since maximal commutative subalgebras always exist (by Zorn's lemma), so do direct integral decompositions into irreducibles. Mautner's theorem says nothing about the uniqueness of the decomposition, and in fact different choices of the maximal commuting subalgebra of $R(U, U)$ can lead in some cases to radically different decompositions into irreducibles. The situation was clarified in the early 1950s using ideas derived from the von Neumann-Murray theory of "factors." Further details will be given below.

As far as harmonic analysis on certain groups and homogeneous spaces is concerned, the results may be expressed without considering direct integral decompositions as such. This was done by Gelfand and Naimark for $SL(2, C)$ (and later for the classical groups) before they knew the results of von Neumann and Mautner. Bargmann also had results in this direction for $SL(2, R)$. Consider what the Peter-Weyl theorem tells us about expansions of square integrable functions on a separable compact group G. For each irreducible unitary representation L of G, the functions $x \to L_x(\varphi) \cdot \psi$ generate a finite-dimensional two-sided invariant vector space $\mathscr{M}^L$ of continuous

functions on G which depends only on the equivalence class of L and is called the space of matrix coefficients for L. These spaces for the various possible L's are mutually orthogonal, and every square integrable f on G may be written as a sum $f = \sum_L f_L$ where $f_L \in \mathscr{M}^L$. Moreover, f_L may be computed from f and the character χ^L of L by the simple formula $f_L(x) = \chi^L(e) \int_G f(xy^{-1})\chi^L(y)d\mu(y)$, which reduces in the commutative case to the formula for computing Fourier coefficients. There is an obvious possible generalization of this formula to any separable unimodular locally-compact group having only finite-dimensional unitary irreducible representations — for example, a semi-direct product of a commutative group and a finite group. One could define $f_L^1(x)$ as $\int_G f(xy^{-1})\chi^L(y)d\mu(y)$ for all f in $\mathscr{L}^1(G, \mu)$ and hope to find a measure $\hat{\mu}$ in the space of all irreducible characters (depending of course on the choice of μ) such that $f(x) = \int f_L^1(x)d\hat{\mu}(L)$ for all f in some dense subspace of $\mathscr{L}^1(G, \mu)$. To do the same for groups with infinite-dimensional irreducible unitary representations seems impossible at first because of the fact that Trace L_x never exists when L is infinite-dimensional. But there is a way out. For each f in $\mathscr{L}^1(G, \mu)$ there exists a unique operator L_f such that $(L_f(\varphi) \cdot \psi) = \int f(x)L_x(\varphi) \cdot \psi)d\mu(x)$ for all φ and ψ in $\mathscr{H}(L)$, and it often turns out that L_f does have a trace. $f \to$ Trace (L_f) is then a linear functional, which may often be shown to be of the form $f \to \int \chi(x)f(x)d\mu(x)$ where $\chi(x)$ is a measurable complex-valued function and is uniquely determined by L (up to changes on sets of measure zero). When L is finite-dimensional, χ always exists and is just χ^L. Thus whenever the above conditions are satisfied it is natural to extend the definition and say that L has a character equal to χ. First for $SL(2, C)$ and later for $SL(n, C)$ and the classical complex groups, Gelfand and Naimark showed a) that characters in this generalized sense exist for all principal series members and b) that there is an expansion formula as indicated. They also found explicit formulae for the characters and the measure $\hat{\mu}$. It is important to notice that the expansion formula does not involve *all* irreducible unitary representations of G. Members outside of the principal series do not occur. On the other hand, one can put them in and simply assign measure zero to the set consisting of all of them. In order to have an expansion formula it suffices to know "almost all" irreducible unitary representations.

More generally, let G be an arbitrary separable locally-compact group whose left and right Haar measures are the same. One might hope to prove the existence of a family $\mathscr{F}$ of irreducible unitary representations L of G having characters χ^L in the above sense together with a measure $\hat{\mu}$ defined on suitable subsets of $\mathscr{F}$ such that $f(x) = \int_{\mathscr{F}}[\int_G f(xy^{-1})\chi^L(y)d\mu(y)]d\hat{\mu}(L)$ for all f in a dense subspace of $L^1(G)$. While such a theorem can indeed be proved for a large and important class of groups, it is true only in a modified form for another large class. Its failure to hold as stated in general is intimately

related to the failure of uniqueness in direct integral decompositions — both failures are due to the existence of the new infinite-dimensional generalizations of full matrix algebras discovered in 1936 by Murray and von Neumann (see the end of section 17).

Let U be a unitary representation of G and suppose that U is discretely decomposable so that $U \simeq L^1 \oplus L^2 \oplus \cdots$ where the L^j are irreducible. Let $L^{i_1}, L^{i_2}, L^{i_3}, \cdots$ be a subset of the L^j including one and only one member of each equivalence class that occurs. Let M^j be the direct sum of all L^k, with L^k equivalent to L^{i_j}. Then $U \simeq M^1 \oplus M^2 \oplus \cdots$ and each M^j is a direct sum of mutually equivalent irreducible representations. The two decompositions, first into the M^j and then into irreducibles, are unique in different senses. The first decomposition is absolutely unique in that for each j, the invariant subspace on which U reduces to M^j is uniquely determined. The further reduction of M^j into irreducibles is unique only up to equivalence. There are very many quite different direct sum decompositions of $\mathcal{H}(M^j)$ into invariant irreducible subspaces. The non-uniqueness in the general case occurs only at the second stage and only in certain instances. Specifically, let G be an arbitrary separable locally-compact group and let U be an arbitrary unitary representation of G in a separable Hilbert space $\mathcal{H}(U)$. Let $CR(U, U)$ be the *center* of the commuting algebra of U. Then Mautner's argument associating a direct integral decomposition of U to every *maximal* commutative subalgebra of $R(U, U)$ associates a direct integral decomposition to $CR(U, U)$ whose components U^λ are not necessarily irreducible but are so-called *factor representations.* A factor representation is by definition a representation whose commuting algebra has a *one-dimensional center,* i.e., is a factor in the sense of Murray and von Neumann. In the discrete case considered above, this decomposition is the decomposition into the M^j. Now whenever a factor representation is discretely decomposable, it is easy to show that all components are equivalent and that the commuting algebra is isomorphic to the algebra of all bounded operators on a finite- or infinite-dimensional separable Hilbert space. Conversely, if M is a factor representation whose commuting algebra is isomorphic to the algebra of all bounded operators on a finite- or infinite-dimensional separable Hilbert space, then M is a direct sum of equivalent irreducibles whose number and equivalence class is uniquely determined. In this case, one speaks of factors and factor representations of type I. When all (or almost all) the factor representations U^λ of U are of type I, one says that U is of type I, and then the decomposition of U into irreducible representations is as unique as in the classical case. As shown by Murray and von Neumann in 1936, however, there exist factors that are not of type I. When these occur among the commuting algebras of the U^λ (on a set of λ of positive measure), the situation is quite different and a further analysis is required.

Murray and von Neumann classified factors according to the behavior of

the lattice of projection operators. The projection operators in $R(U, U)$ correspond one-to-one to the subrepresentations of U, and in the present context it is perhaps more illuminating to present the results of Murray and von Neumann in the language of group representations. Let us define two representations U and V to be *disjoint* if no subrepresentation of U is equivalent to any subrepresentation of V. It is then easy to see that a projection operator P in $R(U, U)$ is such that the subrepresentations U^P and U^{1-P} are disjoint if and only if P is in the center $CR(U, U)$ of $R(U, U)$. Hence $R(U, U)$ is a factor if and only if it is impossible to write U as a direct sum of two disjoint subrepresentations. Let U and V both be factor representations. If they are both of type I, then it is trivial to prove that they fail to be disjoint if and only if each is equivalent to a multiple of the same irreducible representation, and then one is clearly equivalent to a subrepresentation of the other. Even if they are not of type I, it is possible to prove that whenever U and V are not disjoint, then either U is equivalent to a subrepresentation of V or V is equivalent to a subrepresentation of U. Let us define U and V to be *quasi-equivalent* if they are not disjoint. One can then prove that quasi-equivalence is an equivalence relation, that the direct sum of any finite or countably infinite family of quasi-equivalent factor representations is a factor representation in the same quasi-equivalence class and that every subrepresentation of a factor representation is a factor representation in the same quasi-equivalence class. The equivalence classes of factor representations in a fixed quasi-equivalence class thus form an ordered semi-group. When the factor representations in the quasi-equivalence class are of type I, this ordered semi-group is clearly isomorphic to that formed by the positive integers and ∞. One of the main results of Murray and von Neumann is equivalent to the assertion that there are just two other possibilities: either this ordered semi-group is isomorphic to that formed by the positive real numbers and ∞, or it consists of only one element. In the first case, the factor representation is said to be of type II, and in the second case to be of type III. Every factor is the commuting algebra of some representation of some group and has the same type. Let U be a factor representation of type II which is finite in the sense that $U \oplus U$ and U are not equivalent. It can be shown that there exists a subrepresentation V of U, unique to within equivalence such that $V \oplus V$ is equivalent to U. Let $\frac{1}{2}U = V$. Then $\frac{1}{2^n}U$ is defined for all n. Given any positive real number λ, let $\lambda = n + \Sigma\, 1/2^{n_j}$ where n and the n_j are non-negative integers and $n_1 < n_2 < n_3 < \cdots$. One may consistently define λU as $U \overset{n \text{ times}}{\oplus \cdots \oplus} U + \Sigma\, 1/2^{n_j} U$ and ∞U as $U \oplus U \oplus U \oplus \cdots$ and show that every member of the quasi-equivalence class of U is equivalent to λU for one and only one value of λ. If we define λ to be the multiplicity of λU (relative to U), it is clear that quasi-equivalence classes of type II factor representations behave as though some "ideal" or

"virtual" irreducible were being repeated with a continuum of (relative) multiplicities. While type II factor representations can be decomposed as direct integrals of irreducibles, the irreducibles that occur are far from being uniquely determined and the decompositions seem to be of little if any use. It seems best to stop the decomposition at the first level. When U is a type II factor representation but is not necessarily finite, the different projections in $R(U, U)$ define subrepresentations of U, and their multiplicities (relative to some finite subrepresentation of U) define a so-called *relative dimension function* $P \to d(P)$ on the set of all projections P in $R(U, U)$. If A is any bounded self-adjoint operator in $R(U, U)$ and $E \to P_E^A$ is the projection-valued measure on the line assigned to A by the spectral theorem, then the P_E^A are all in $R(U, U)$ and $E \to d(P_E^A)$ is a measure on the real line. When x is integrable with respect to this measure (and it always is when U is finite), the integral is called the *relative trace* of A. In 1937 Murray and von Neumann published a second paper whose chief purpose was to prove the surprisingly difficult theorem that the relative trace, when it exists, is a linear function of the operator. Since every bounded linear operator is uniquely of the form $A + iB$ where A and B are self-adjoint, this linear functional has a well defined extension to all bounded linear operators whenever $d(P_I^A) < \infty$, and to a large subset in any case.

Using the relative trace concept in conjunction with the theory of direct integral decompositions, it is possible to prove an expansion theorem for reasonably general functions on any separable locally-compact group whose left and right Haar measures coincide. This was done by Segal (1918—) and Mautner in independent papers published in 1950. When U is a factor representation, the operator $U_f = \int f(x)U_x d\mu(x)$ will lie in the commutator of $R(U, U)$ and this is always a factor of the same type as $R(U, U)$. If $T^R(U_f)$ denotes the (suitably normalized) relative trace of U_f, then $f \to T^R(U_f)$ will be a linear functional, which one can hope to write in the form $f \to \int f(x)\chi(x)d\mu(x)$. When this is possible, one can think of χ as the character of U (with respect to the normalization chosen). Now consider the regular representation of the group G under consideration and its decomposition into factor representations defined by the center of the commuting algebra. When almost all of these factor representations are of type I *and* the irreducibles that generate them have characters as indicated above, the Segal-Mautner theorem implies the obvious generalization of the expansion theorem of Gelfand and Naimark. However, it goes further in two directions. First of all, it is not necessary that the linear functionals $f \to \text{Trace } L_f$ be of the form $f \to \int f(x)\chi(x)d\mu(x)$. One can think of the linear functions themselves as characters and replace the expression $\int_G f(xy^{-1})\chi^L(y)d\mu(y) = \int_G f(xy)\chi^{\bar{L}}(y)d\mu(y)$ by Trace $\bar{L}_{f_x}$, where f_x is the left translate of f and $\bar{L}$ is the complex conjugate of L. Mautner and Segal do this and also include the case in which type II factor representations occur by replacing the trace of

the irreducible generator with the Murray-von Neumann relative trace. It can be shown that type III representations do not occur in the decomposition of the regular representation of a group whose left and right Haar measures coincide. The final result is called the Plancherel theorem for the group in question, and the measure $\hat{\mu}$ in the appropriate set of irreducible and factor representations is called the *Plancherel measure.* When the group is commutative, the Plancherel measure is Haar measure in the dual group. It is important to notice that the Plancherel theorem of Mautner and Segal is an abstract existence theorem. In specific cases, the problem remains of finding the irreducible and factor representations that are needed and of specifying the particular measure on this set which serves as the Plancherel measure.

When non-type I factor representations exist for a group G, its representation theory is considerably more complicated in several ways, and it no longer suffices to know the irreducible unitary representations in order to know all unitary representations. One must also know the factor representations of type II and type III. Since the latter are always difficult (if not impossible) to find in their totality, it is helpful that many of the most interesting groups can be shown to have no non-type I factor representations at all. Such groups are called type I groups and obviously include the compact groups and the locally-compact commutative groups. Mautner, who was the first to note the effect on uniqueness in decomposition theory of the existence of non-type I factor representations, was also the first to attempt to determine which groups were type I groups and which were not. In a paper published in 1950, he showed that both $SL(2, R)$ and $SL(2, C)$ are type I groups and conjectured that all semi-simple Lie groups are type I. He also found a five-dimensional solvable Lie group that is not of type I and gave examples showing that discrete groups tend never to be type I groups unless they are very close to being finite or commutative. Over a decade later it was shown by Thoma that a countable discrete group is a type I group if and only if it has a commutative normal subgroup with a finite quotient. Irregular semi-direct products are a rich source of groups that are not of type I. Mautner's non-type I five-dimensional solvable Lie group is an example in which the normal subgroup is a four-dimensional vector group.

A Lie algebra version of Mautner's conjecture about the type I-ness of semi-simple Lie groups was proved by Harish-Chandra in the course of a long paper published in 1951. (The original conjecture was proved in a second long paper by the same author published two years later.) Harish-Chandra combined his attack on the type I-ness question with a powerful and original attack on the problem of finding the irreducible Banach space representations of general semi-simple Lie groups. A complete classification is difficult to derive, and Harish-Chandra soon concentrated his efforts on

finding enough irreducible unitary representations to decompose the regular representation and on finding an explicit Plancherel formula. In 1952 he found the Plancherel formula for $SL(2, R)$, and in 1954 published a long paper on the general case, including complete results for the complex semi-simple Lie groups. His results in the complex case go beyond those found earlier by Gelfand and Naimark in that the five exceptional groups are included and in that all cases are treated at once in a uniform manner. The non-complex case turned out to be much more difficult; in fact, it demanded Harish-Chandra's best efforts for over a quarter of a century. The final details have been written down only recently. The fundamental simplification that takes place in the case of a complex semi-simple Lie group is that there is a single commutative subgroup the conjugates of whose elements form a set whose complement has Haar measure zero.

Already in the case of $SL(2, R)$ no such subgroup exists. Instead there are two commutative subgroups, the conjugates of whose elements constitute two sets intersecting in a two element set, whose union has a complement of Haar measure zero. In more complicated real semi-simple Lie groups, many such non-conjugate "Cartan subgroups" exist. By 1952 Harish-Chandra had already suggested that to find enough irreducible unitary representations for a Plancherel formula it would be necessary to associate a different family to each conjugacy class of Cartan subgroups. In $SL(2, R)$ the principal series and the discrete series correspond respectively to the diagonal subgroup and the compact subgroup of all $\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$, and their members are parameterized by finite subsets of characters of these two commutative groups. Work of Gelfand and Graev published a year later showed that this was definitely the case for $SL(n, R)$. Here the number of Cartan subgroups is $\frac{n+2}{2}$ or $\frac{n+1}{2}$ according to whether n is even or odd. In 1954 Harish-Chandra described a method[8] for assigning a family of irreducible (or nearly irreducible) unitary representations to each *non-compact* Cartan subgroup A, its members being parameterized by certain finite subsets of the character group $\hat{A}$ of A. The method involved using the inducing process $L \to U^L$ and forming L out of irreducible unitary representations of lower-dimensional semi-simple Lie groups. The method failed when A was compact, and Harish-Chandra saw the central problem as that of finding some other method for constructing the family that he felt sure existed in this case. This is the celebrated problem of the "discrete series." Harish-Chandra announced a solution in 1963—at least as far as finding the *characters* of the representations was concerned.

I have already explained how some irreducible unitary representations may be said to have "characters" either in the sense of actual functions on the group or at least as linear functions $f \to \ell(f)$ defined for f in some

reasonably large vector space of complex-valued functions on G. In the early 1950s Godement was particularly active in attempting to construct a general theory of characters along such lines. He published a long memoir on the subject in 1951 and another (in two installments) in 1954. As far as semi-simple Lie groups are concerned, however, the most useful theory is due to Harish-Chandra. Let $\mathcal{D}$ be the space of all infinitely differentiable functions with compact support on the semi-simple Lie group G. In 1954 Harish-Chandra showed that for f in $\mathcal{D}$, Trace L_f exists for all irreducible unitary L (and some Banach space representations as well) and that $f \to$ Trace L_f is a distribution in the sense of L. Schwartz (1915—). Two years later he showed that this distribution reduced to an analytic function on an open set with a complement of measure zero. Finally in 1965 he showed that his distribution characters are actually locally summable functions on the whole group.

My aim in this section has been to give some rough idea of the beginnings and principal concepts of the huge theory that has emerged from the successful attempt to extend harmonic analysis to locally-compact groups which are not necessarily compact nor commutative. My account is at most three-quarters complete as far as the first decade is concerned, and the finding of the general Plancherel formula for semi-simple Lie groups is only a fraction of what has been added to the theory in the past two decades. For further details about the period 1946-1961 the reader is referred to my colloquium lectures [*14*], and for a survey of developments between 1955 and 1975 to the second half of my 1976 book [*15*]. (The first half of the book is a reprint of a set of lecture notes for a course given at the University of Chicago in 1955.)

22. APPLICATIONS OF THE GENERAL THEORY

In the earlier sections of this article I have indicated three different origins of the method of harmonic analysis: in probability theory, in mathematical physics, and in number theory. In this section we shall see that the general point of view that began to evolve with Frobenius's introduction of group characters in 1896, and which reached a certain level of completeness and coherence in the 1950s, has significant applications to all three subjects.

The applications to probability theory are still relatively undeveloped and will be dealt with quite briefly. The basic idea is that families of random variables occur in many contexts (not just strung out in time) and that the parameter space may have symmetry properties. In other words, there is a natural generalization of a stationary stochastic process (see section 18) in which the ergodic action of the real line or the integers in a probability measure space Ω, μ is replaced by an ergodic action of some other group.

For example, in considering the statistical mechanics of a gas (considered for convenience as occupying all of space), the number of molecules in a finite subset V of space is a random variable, and the set of random variables obtained by varying V constitutes a generalization of a stochastic process. Assuming the gas to have properties invariant under translation and rotation considerations analogous to those given in section 18 leads to a natural measure-preserving action of the Euclidean group $\mathscr{E}$ on the underlying probability measure space Ω, μ. The Koopman construction then yields a unitary representation of $\mathscr{E}$, and the decomposition of this representation can be used as in section 18 to contribute to the analysis of the statistical mechanical problem. A preliminary study of this kind of application was made in a note published in 1960 by A. M. Yaglom.

The applications of the theory of unitary group representations to quantum physics are by now so extensive that anything like a complete summary would require a book-length article. I shall content myself here with a few remarks and references. In section 21 it was explained how one could be led to the imprimitivity theorem through three successive generalizations of the Stone-von Neumann theorem on the uniqueness of the solutions of the Heisenberg commutation relations. The final results seem to have nothing to do with the original physical problem. It is thus of some interest that it turns out a) to be possible to deduce the Schrödinger equation for a single free particle from the general principles of quantum mechanics and group theoretical invariance postulates, and b) that the imprimitivity theorem is the chief tool used in carrying out the derivation. Indeed, using the imprimitivity theorem one can show that the Heisenberg commutation rules are essentially consequences of Euclidean invariance — and that the existence and properties of spin come along as a bonus.

The argument proceeds as follows: Let S denote physical space and let $\mathscr{H}$ be the Hilbert space of states for our one-particle system (see section 16). Let $\mathscr{E}$ denote the group of all isometries of S so that $\mathscr{E}$ acts transitively on S and preserves the volume measure ν. For each Borel subset E of S, let P_E be the self-adjoint operator in $\mathscr{H}$ corresponding to the observable, which is one when the particle is observed to be in E and zero otherwise. Then P_E must be a projection operator, and it is not difficult to accept hypotheses implying that $E \to P_E$ must be a projection-valued measure on S. This projection-valued measure determines all position observables in the following manner: If g is any real coordinate, i.e., any real-valued (Borel) function on S, then $E \to P_{g^{-1}(E)}$ is a projection-valued measure on the line, and the self-adjoint operator associated with it by the spectral theorem is the self-adjoint operator in $\mathscr{H}$ associated with the coordinate observable g. Let $t \to V_t$ denote the unitary representation of the additive group of the real line, defining the "dynamics" or time evolution of our system. Once P and V have been given in some concrete fashion, our system is completely de-

termined and all questions about what happens can be reduced to mathematical calculation. However, without further assumptions the pair P, V could be an arbitrary pair consisting of a projection-valued measure on S and a unitary representation V of the additive group of the line. We now introduce the assumption that the whole system is invariant under $\mathscr{E}$. This means in the first instance that each $\alpha \in \mathscr{E}$ is intrinsically associated with an automorphism of our quantum model, i.e., with a unitary (or anti-unitary) operator U, and that $U_{\alpha\beta} = U_\alpha U_\beta \sigma(\alpha, \beta)$ where σ is some projective multiplier for the group $\mathscr{E}$. It means in addition that P_E and $P_{[E]\alpha}$ are transforms of one another by the operator U_α; that is, that $U_\alpha^{-1} P_E U_\alpha = P_{[E]\alpha}$ for all E and α. Now when every element in $\mathscr{E}$ is the square of another, all the operators U_α must be unitary rather than anti-unitary. Moreover, by replacing $\mathscr{E}$ by a covering $\mathscr{E}$ one can get rid of the factor σ. Supposing these things done, one recognizes that P is a system of imprimitivity for U based on S. Since $S = \tilde{\mathscr{E}}/K$ where K is the closed subgroup of $\tilde{\mathscr{E}}$ leaving some origin s_0 in S fixed, the imprimitivity theorem implies that the pair U, P is equivalent to the pair U^L, P^L where L is some unitary representation of K. When K is compact, as it is for the usual Euclidean model for space, there is only a discrete countable set of possibilities for L and hence for the system P, U. The simplest case is that in which L is the one-dimensional identity. In that case, we at once reach the conclusion that $\mathscr{H}$ is isomorphic to $\mathscr{L}^2(S, \mu)$ in such a way that $U^L_\alpha(f)(s) = f(s\alpha)$ and $P^L_E(f)(s) = \varphi_E(s) f(s)$ where $\varphi_E(s) = 1$ or 0 according as $s \in E$ or $s \notin E$. We shall not pursue the analysis further here. Let it suffice to say that in most of the particles occurring in physics the representation L is an irreducible unitary representation of $SU(2)$ and that $\frac{1}{2}(\dim L - 1)$ is called the *spin* of the particle. The representation $t \to V_t$ is limited by the requirement that $V_t U^L_\alpha = U^L_\alpha V_t$, and the possibilities under this limitation can be analyzed using the theory of unitary group representations.

The connection between systems of imprimitivity and particle position observables seems to have been first noted by Wightman (1922—). Wigner, in collaboration with T. D. Newton, published a paper on position observables for relativistic particles in 1949 — the same year that my statement and abbreviated proof of the imprimitivity theorem were published. Not long thereafter, Wightman read both papers and noted that the contents of one were just what was needed to make the other rigorous. However, he did not publish his results until 1962. I worked out the axiomatics of a particle discussed above after hearing a vague account of what Wightman had done.

Starting with the group representational model for a single particle described above, one can discuss systems of interacting particles in a similar spirit and finally fit almost the whole of quantum physics into the

framework of the theory of unitary group representations. Further details will be found on pp. 328–357 of my book [*15*].

In some ways the applications to number theory are the most interesting of all, in part perhaps because they differ more from applications of the compact and commutative theory, and in part because they are still very incompletely understood and are at the center of a rapidly developing area of mathematics. I shall describe them at somewhat greater length than I did the applications to probability theory and physics.

Let F be a function analytic in the upper half-plane and let N be a positive integer. Let Γ_N denote the subgroup of $SL(2, R)$ consisting of all $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ in which $a-1$, b, c, and $d-1$ are integer multiples of N. Let k be a positive integer. One defines an (unrestricted) modular form of level N and weight $k/2$ (or dimension $-k$) to be a complex-valued function F analytic in the upper half of the complex plane such that $F\left(\frac{az+b}{cz+d}\right) = (cz + d)^k F(z)$ for all z in the upper half-plane and all $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_N$. Setting $a = 1$, $b = N$, $c = 0$, $d = 1$, one sees in particular that $F(z + N) = F(z)$, so that F has the Fourier expansion $F(z) = \sum_{j=-\infty}^{\infty} \varphi(j) e^{(2\pi i j z)/N}$. When $\varphi(j) = 0$ for $j < 0$ and (when $N > 1$) certain other growth conditions are satisfied, one says that F is a *modular form* of level N and weight $k/2$. As already indicated in sections 10 and 19, the theory of modular forms has close connections with number theory because of the interesting number-theoretical properties of the functions $n \to \varphi(n)$, which occur as coefficient sequences. In particular, for every positive definite quadratic form with an even number of variables, let $\varphi_Q(n)$ denote the number of integer points on the hypersurface $Q(x_1, x_2, \cdots, x_\ell) = n$. Then $n \to \varphi_Q(n)$ occurs as the coefficient sequence for a modular form of weight $\ell/4$ and some level depending on the form.

As explained in some detail in sections 10 and 11, a fairly complete theory of modular forms of level one was given by Hurwitz in his thesis (published in 1881). Extending the theory to forms of higher level turned out to be quite difficult. Although Fricke and Klein made a certain amount of progress (see section 11), the situation in 1925 was that complete results were available for only a few small values of k and N. Then, beginning with an announcement in 1925, Hecke published a series of papers in which several important new ideas were used to carry the theory a great deal further. Since Hecke's ideas and results are vital in the application of unitary group representations to number theory, it will be necessary to devote some space to describing them.

The material announced in 1925 was worked out in detail in a paper published in 1926. The main idea was to try to construct previously unknown modular forms using coefficient functions from other parts of

number theory. Hecke found it possible to get new modular forms of weight $\ell/2$ and various levels from the coefficients of the Dirichlet series expansions for certain zeta functions associated with *real* quadratic number fields. Let $n \to \varphi(n)$ be a complex-valued function defined on the non-negative integers which satisfies suitable growth conditions at ∞. Then $\sum_{n=0}^{\infty} \varphi(n)e^{2\pi inz}$ will converge in the upper half-plane and define an analytic function F_φ such that $F_\varphi(z + 1) \equiv F_\varphi(z)$. Since $\left(\begin{smallmatrix}1 & 1\\ 0 & 1\end{smallmatrix}\right)$ and $\left(\begin{smallmatrix}0 & 1\\ -1 & 0\end{smallmatrix}\right)$ generate the modular group, and since $F_\varphi\left(\frac{0 \cdot z - 1}{1 \cdot z + 0}\right) = F_\varphi(-1/z)$ and $F_\varphi\left(\frac{1 \cdot z + 1}{0 \cdot z + 1}\right) = F_\varphi(z + 1)$, it follows that F_φ is a modular form of weight k and level ℓ if and only if $F_\varphi(-1/z) \equiv z^{2k}F_\varphi(z)$. On the other hand, consider the restriction $y \to F_\varphi(iy)$ of F_φ to the positive real axis. Since the positive real axis is a commutative group under multiplication, we may form the Fourier transform $\int_0^\infty (F_\varphi(iy) - \varphi(0))y^\sigma \frac{dy}{y} = \int_0^\infty (F_\varphi(iy) - \varphi(0))y^{\sigma-1}dy$ and consider its extension to complex values $s = \sigma + i\tau \to \int_0^\infty (F_\varphi(iy) - \varphi(0))y^{s-1}dy$. (When so written, the Fourier transform is usually called the Mellin transform.) Writing $F_\varphi(iy) - \varphi(0) = \sum_{n=1}^{\infty} \varphi(n)e^{-2\pi ny}$ and integrating term by term, one obtains a series expansion $\sum_{n=1}^{\infty} \varphi(n) \int_0^\infty e^{-2\pi ny}y^{s-1}dy$ for this Mellin transform. Making the change of variable $y \to \frac{t}{2\pi n}$, the nth term becomes $\varphi(n)\frac{1}{(2\pi n)^s}$ $\int_0^\infty e^{-t}t^{s-1}dt = \frac{\Gamma(s)}{(2\pi n)^s}\varphi(n)$, where $s \to \Gamma(s) = \int_0^\infty e^{-t}t^{s-1}dt$ is the classical gamma function. Thus the Mellin transform becomes $s \to \frac{\Gamma(s)}{(2\pi)^s}D_\varphi(s)$ where $D_\varphi(s)$ has the Dirichlet series expansion $D_\varphi(s) = \sum_{n=1}^{\infty} \frac{\varphi(n)}{n^s}$. A straightforward calculation shows that F_φ satisfies the identity $F_\varphi(-1/z) = z^{2k}F_\varphi(z)$ if and only if D_φ has an analytic continuation to the whole complex plane and satisfies the functional equation

$$\frac{\Gamma(2k - s)}{(2\pi)^{2k-s}}D_\varphi(2k - s) = \frac{\Gamma(s)}{(2\pi)^s}D_\varphi(s).$$

(The argument is simplest if $\varphi(0) = 0$, for then D_φ is entire. Otherwise D_φ has a pole at $2k$ and the residue at this pole determines $\varphi(0)$.) Thus F_φ is a modular form of weight k if and only if D_φ satisfies the functional equation in question. In his 1926 paper, Hecke used a more complicated variant of the correspondence just described to construct modular forms of weight 1/2 and higher level from Dirichlet series D satisfying functional equations of the form $\frac{\Gamma(1 - s)}{a^{1-s}}D(1 - s) = \frac{\Gamma(s)}{a^s}D(s)$. Here a is a positive constant

which depends upon the level. Certain two-fold products of Dirichlet L functions and products of Dirichlet L functions with the Riemann zeta function have functional equations of this form.

In 1927 and 1928 Hecke published two further and rather different contributions to the problem of determining the modular forms of higher level. The 1928 contribution has already been described in the last part of section 19. In the 1927 paper Hecke showed how to extend the theory of Eisenstein series and cusp forms (see section 10) to forms of higher level. The Riemann surface whose points are the orbits in the action of Γ_N on the upper half-plane can be compactified by adding a finite number of points (the cusps). For each weight $k = 1/2, 1, 3/2, 2, \cdots$, one has a generalized Eisenstein series for each cusp which "takes the value" 1 at that cusp and "the value" 0 at all other cusps. Every modular form of level N and weight k is uniquely a sum of a form which vanishes at all the cusps (a cusp form) and a linear combination of the Eisenstein series attached to the cusps. The Fourier coefficients of the Eisenstein series have simple number-theoretical properties analogous to those in the level one case. The Fourier coefficients of the cusp forms are of a lower order of magnitude than the Fourier coefficients of the Eisenstein series. Using the facts just stated, Hecke was able to obtain the Hardy-Littlewood asymptotic formula (see section 20) for representations of integers as sums of squares (but not higher powers) as a corollary of his theory. The formulae are exact rather than asymptotic whenever there are no cusp forms.

The connection between modular forms and Dirichlet series satisfying a Riemann-type functional equation, which Hecke had exploited in a special case in the 1926 paper mentioned above, became the basis of a far-reaching theory presented by Hecke in four long papers published in 1936 and 1937. Hecke applied his results to the theory of n-ary quadratic forms in a very long paper published in 1940. Detailed summaries of part of this work were published in 1935 and 1936. Instead of continuing to find new modular forms by transforming known Dirichlet series, Hecke reversed things and studied the Dirichlet series obtained by applying the Mellin transform to an arbitrary modular form. This study led to two kinds of results. On the one hand, by using known facts about the general theory of modular forms and by proving precise theorems about the bijectivity of the Mellin transform between well-defined spaces of Dirichlet series and modular forms, Hecke was able to deduce theorems characterizing various zeta and L functions by their functional equation and certain additional properties. In doing so, he found an extensive generalization of a characterization of the Riemann zeta function given by Hamburger (1889-1956) in 1921 and 1922. On the other hand, he found a method for showing that the Dirichlet series one gets by taking the Mellin transforms of a modular form are like those occurring in algebraic number theory not only in that a) they satisfy a Riemann-type

functional equation, but also in that b) they can be written as finite linear combinations of Dirichlet series which both satisfy the functional equation *and have an "Euler product factorization"* with one factor for each prime.

Perhaps a word is in order here about the definition and significance of an Euler product factorization. Let $n \to \varphi(n)$ be a complex-valued function on the integers of such a character that $\sum_{n=1}^{\infty} \frac{\varphi(n)}{n^s}$ converges whenever the real part of s is sufficiently large, i.e., in some half-plane $\sigma > \sigma_0$ where $s = \sigma + i\tau$. Typically $\varphi(n)$ will be the unknown solution to some number-theoretical problem containing n as a parameter. If φ is multiplicative in the sense that $\varphi(nm) = \varphi(n)\varphi(m)$ for n and m relatively prime, then φ is completely known when it is known at the prime powers. Moreover, if φ is multiplicative, one verifies at once that $\sum_{n=1}^{\infty} \frac{\varphi(n)}{n^s} = \prod_p E_p(s)$ where $E_p(s) = \sum_{k=0}^{\infty} \frac{\varphi(p^k)}{p^{ks}}$. Conversely, if $\sum_{n=1}^{\infty} \frac{\varphi(n)}{n^s}$ factors in the indicated way, one says that the Dirichlet series $\sum_{n=1}^{\infty} \frac{\varphi(n)}{n^s}$ has an Euler product factorization and one can prove easily that φ is multiplicative. In the original case considered by Euler (see section 5), $\varphi(n) = 1$ and $E_p(s) = \sum_{k=0}^{\infty} \left(\frac{1}{p^s}\right)^k = \frac{1}{1 - p^{-s}}$. In the Euler products which occur in number theory it is usually (if not always) true that $E_p(s)$ is a *rational* function of p^{-s}, whose numerator and denominator have degrees that are bounded as a function of p. Thus, for each prime p, one need only know $\varphi(p^k)$ for a finite number of values of k to know $\varphi(n)$ for all n. Finally, suppose that one has a system $\varphi_1, \varphi_2, \cdots, \varphi_h$ of linearly independent functions of n which are not multiplicative, but that the vector space they span has a basis $\psi_1, \psi_2, \cdots, \psi_h$ where the ψ_j are multiplicative. Then, as one can easily check, it suffices to know $\varphi_i(p^k)$ for all i, p, and k to know $\varphi_i(n)$ for all i and n. Moreover, if the Euler products which occur in the factorization of the Dirichlet series have *rational* factors as indicated above, it suffices for each p and i to know $\varphi_i(p^k)$ for a finite number of values of k.

Hecke's theory is considerably simpler for modular forms of level 1 than for forms of higher level, and I shall attempt to describe only this simple case. As already mentioned in section 10, the Eisenstein series of level 1 and weight k has a constant multiple whose Fourier coefficients $n \to \varphi_k(n)$ are given by the simple formula $\varphi_k(n) = \sum_{d|n} d^{2k-1}$. It follows easily that $n \to \varphi_k(n)$ is multiplicative, and that an Euler product factorization for $\sum_{n=1}^{\infty} \frac{\varphi_k(n)}{n^s}$ not only exists but takes the special form

$$\prod_p \left(\frac{1}{1 - p^{-s}}\right)\left(\frac{1}{1 - p^{2k-1}p^{-s}}\right) = \prod_p \left(\frac{1}{1 - (1 + p^{2k-1})p^{-s} + p^{2k-1}p^{-2s}}\right).$$

Hecke's principal result was that for each $k = 1, 2, 3, \cdots$, there is a basis for the space of cusp forms of weight k such that if $n \to \psi_\ell(n)$ is the sequence of Fourier coefficients for the ℓth basis element, then the Dirichlet series $\sum_{n=1}^{\infty} \frac{\psi_\ell(n)}{n^s}$ has an Euler product expansion of the form $\sum_{n=1}^{\infty} \frac{\psi_\ell(n)}{n^s} = \prod_p \frac{1}{1 - \lambda_\ell^k(p)p^{-s} + p^{2k-1}p^{-2s}}$ where the coefficients $\lambda_\ell^k(p)$ remain unknown. This has exactly the same form as in the case of Eisenstein series where $\lambda_\ell^k(p)$ is known and equal to $1 + p^{2k-1}$. The lowest weight k for which cusp forms exist is $k = 6$, and in this case the space of cusp forms is one-dimensional. Let $\tau(n)$ denote the nth Fourier coefficient of the unique cusp form of weight 6 whose $e^{2\pi iz}$ coefficient is 1. Already in 1916 Ramanujan had conjectured that $\sum_{n=1}^{\infty} \frac{\tau(n)}{n^s}$ has an Euler product factorization of the form $\prod_p \frac{1}{1 - \tau(p)p^{-s} + p^{11}p^{-2s}}$, and in 1917 Mordell published a proof of this conjecture together with a proof of the fact that the Dirichlet series $D(s) = \sum_{n=1}^{\infty} \frac{\tau(n)}{n^s}$ satisfies the functional equation $\frac{\Gamma(s)D(s)}{(2\pi)^s} = \frac{\Gamma(12-s)D(12-s)}{(2\pi)^{12-s}}$.

Hecke's theory constituted a far-reaching generalization of these isolated results. Ramanujan also conjectured that $|\tau(p)| \leq 2p^{11/2}$, or equivalently that the polynomial $1 - \tau(p)p^{-s} + p^{11}p^{-2s}$ cannot have unequal real roots. However, neither he nor Mordell was able to prove this. Hecke's theory suggests a natural generalization which was formulated by Petersson (1902—) in 1939. The Ramanujan-Petersson conjecture states that $|\lambda_\ell^k(p)| \leq 2p^{(2k-1)/2}$. It was finally proved by Deligne (1944—) in 1974 as a corollary of a much more general theorem.

Hecke's discovery of a basis with multiplicative Fourier coefficients for spaces of modular forms was based on the consideration of certain linear operators T_n, which had been used by Hurwitz in the classical theory of the 1880s and 1890s. However, Hecke found new properties of these operators and used them in a new way, and as a result they are now called Hecke operators. Let F be a modular form of weight k and level 1, and for each n let $T_n(F)(z) = n^{2k-1}\sum_{a,b} F\left(\frac{az+b}{d}\right)d^{-2k}$ where $d = n/a$, a varies over all divisors of n and for each a, $b = 0, 1, 2, \cdots, d-1$. Then $T_n(F)$ is also a modular form of weight k and level 1, and $F \to T_n(F)$ is a linear operator. Hecke's basic observation was that these operators not only commute with one another but vary with n in a manner strictly analogous to the way in which the Fourier coefficients of Eisenstein series vary. Specifically, $T_nT_m = \sum_d d^{2k-1}(T_{nm/d^2})$ where d varies over all positive divisors of the highest com-

mon factor of n and m. This implies that $T_nT_m = T_{nm}$ whenever n and m are relatively prime, and that $I + \Sigma z^jT_{p^j} = \frac{1}{(I - T_pz + Ip^{2j-1}z^2)}$ where I is the identity operator. Whenever the T_n are all diagonalizable, it follows from their commutivity that they are *simultaneously* diagonalizable. One thus obtains the desired basis in the space of modular forms by choosing any (suitably normalized) basis which diagonalizes all of the T_n. Hecke was able to prove the diagonalizability only for certain values of k. In the 1939 paper mentioned above, however, Petersson was able to prove diagonalizability in general. He did so by introducing a natural inner product in the space of modular cusp forms of a given dimension with respect to which all the T_n are self adjoint. This inner product turned out to be useful in the general theory of automorphic forms and is now known as the Petersson inner product.

While Hecke and Petersson were developing the theory described above, C. L. Siegel (1896—) was applying analytical considerations to a different kind of extension of the theory of n-ary quadratic forms. Siegel generalized the problem of finding the number of representations of an integer by a quadratic form to that of finding the number of representations of one quadratic form by another. Instead of being content with asymptotic results as were Hardy and Littlewood, however, or attempting to get at the lower order terms via a generalization of Hecke's theory (which was just then being developed), Siegel replaced the study of individual forms by the study of certain averages. (As explained in section 6, the same device was used by Gauss to get exact results in the theory of binary quadratic forms.) Siegel's main result is a formula for the average solution number which specializes to the singular series of Hardy and Littlewood (see section 20) when one of the forms is a form in one variable. He showed how this result could be interpreted as a product over the primes (including ∞) of p-adic or real solution densities, and also that it was equivalent to an identity between two differently defined "generalized modular forms."

Siegel's generalized modular forms are analytic functions of several complex variables which are related to the n-dimensional symplectic group in the same way that the classical modular forms are related to the group $SL(2, R)$. Let V^{2n} be a $2n$-dimensional vector space equipped with a non-degenerate alternating bilinear form $[\ ,\ ]$. The n-dimensional symplectic group $Sp(n)$ is the group of all non-singular linear transformations T of V^{2n} into V^{2n} such that $[T(\varphi), T(\psi)] = [\varphi, \psi]$ for all φ and ψ in V^{2n}. If K is a maximal compact subgroup of $Sp(n)$, then $Sp(n)/K$ has the structure of a complex manifold which can be identified with a space of complex matrices. When $n = 1$, $Sp(n)$ is isomorphic to $SL(2, R)$, and $Sp(n)/K$ becomes the upper half-plane.

Siegel's results for the special case of definite forms appeared in a long paper published in 1935. Their extensions to the indefinite case and to forms with coefficients in an algebraic number field were published in 1936 and 1937 respectively. Of course, it was now very much in order to extend the classical theory of modular forms to include Siegel's new modular forms on $Sp(n)/K$, and the foundations for such a theory were laid down by Siegel himself in two papers published in 1939 and 1943 respectively. Many mathematicians became interested and the new theory soon experienced a considerable development.

One of the more active workers in the development of Siegel's theory as well as other generalizations of modular form theory to several complex variables was H. Maass (1911—). Returning to the one variable case, Maass in 1949 published a long paper extending Hecke's theory in an unexpected direction. While Hecke's theory made it possible to give an abstract characterization of the zeta function of an imaginary quadratic extension of the rational field Q, it failed for real quadratic extensions because of the different form the functional equation takes in that case. Maass showed that it was possible to develop a modification of Hecke's theory which applied to the zeta functions of real quadratic fields if one replaced the analytic functions of the classical theory of modular forms by group invariant functions which are not analytic in the sense of complex analysis. Instead, they are eigenfunctions of the second order differential operator $y^2\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)$. This operator is (to within a multiplicative constant) the only second order differential operator that commutes with the action of $SL(2, R)$. In Maass's theory (which parallel's Hecke's in some respects but is quite different in others), the role of the weight of a modular form is played by the eigenvalue corresponding to an eigenfunction of $y^2\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)$.

That the unitary representation theory of the group $SL(2, R)$ might be closely connected with the classical theory of modular forms (and hence to number theory) is immediately suggested if one compares the identity $f\left(\frac{az+b}{cz+d}\right)=(cz+d)^{2k}f(z)$ occurring in the definition of a modular form of weight k with one form of the definition of the discrete series of irreducible unitary representations of $SL(2, R)$ given in section 21. The member V^{2k} of the discrete series had as Hilbert space a space of analytic functions on the upper half-plane and was defined by the formula $V^{2k}_{\left(\begin{smallmatrix}a&b\\c&d\end{smallmatrix}\right)}f(z)=f\left(\frac{az+b}{cz+d}\right)(cz+d)^{-2k}$. Ignoring growth conditions for the moment, one sees at once that being a modular form of weight k and level N is equivalent to being a member of the subspace of $\mathcal{H}(V^{2k})$ on which V^{2k} reduces to the iden-

tity when restricted to Γ_N. If the obvious analogue of the Frobenius reciprocity theorem (see section 15) were true in the present context, the dimension of this subspace would coincide with the multiplicity of occurrence of V^{2k} in the decomposition of the representation $U^{I^{\Gamma_N}}$ induced by the one-dimensional identity representation I_{Γ_N} of Γ_N. While neither of these two heuristic arguments can be made correct, the result they suggest is largely true. For $k = 2, 3, 4, \cdots$, the dimension of the space of all *cusp* forms of weight k and level N is precisely equal to the multiplicity with which V^{2k} occurs as a discrete direct summand of the induced representation $U^{I^{\Gamma_N}}$. The first hint that such a connection between modular forms and the discrete series might exist occurs in a paper published by Gelfand and Fomin in 1952. These authors were concerned not with modular forms or questions in number theory, but with using the theory of unitary group representations to prove the ergodicity of certain flows on *compact* homogeneous spaces of the form G/Γ where $G = SL(2, R)$ and Γ is a discrete subgroup. Because of the compactness of G/Γ, the groups Γ_N are excluded, but one has obvious analogues of modular forms in which Γ replaces Γ_N. These are the automorphic forms of Poincaré, and when G/Γ is compact all automorphic forms are cusp forms. For use in their study of ergodicity Gelfand and Fomin proved that U^{I^Γ} contains V^{2k} a number of times equal to the dimension of the space of Γ automorphic forms of weight k. The connection of the Gelfand-Fomin observation with modular forms and number theory seems not to have been noticed until after Selberg (1917—) published an extremely interesting and influential paper in 1956.

While Selberg made no mention of either the Maass non-analytic automorphic forms or of the unitary representations of semi-simple Lie groups, his paper can be most easily understood as a contribution to a generalization of the Maass theory in which a group representational interpretation of Maass's automorphic forms plays a key role. Let Γ be a discrete subgroup of $G = SL(2, R)$ such that G/Γ is compact, and consider the unitary representation U^{I^Γ} of G induced by the identity representation of Γ. It follows easily from the compactness of G/Γ that U^{I^Γ} decomposes as a discrete direct sum of irreducible unitary representations of G, each occurring with finite multiplicity. The cited theorem of Gelfand and Fomin states that the multiplicities with which half the members of the discrete series occur are equal to the dimensions of certain spaces of classical automorphic forms. What about the other irreducible unitary representations of G, especially the principal series? Maass's introduction of non-analytic automorphic forms received an extra vindication when it was pointed out by Gelfand and Pjateskii-Shapiro (1929—) in 1959 that for each principal series member L there is a real number λ with the following property: The multiplicity with which L occurs in the decomposition of U^{I^Γ} is equal to the dimension of the space of all Maass automorphic forms for the group Γ

and the eigenvalue λ. Moreover, λ may be computed from the character $\begin{pmatrix} a & 0 \\ \mu & 1/a \end{pmatrix} \to |a|^{i\sigma}$ which induces the principal series member L by means of the formula $\lambda = 1/4 - \sigma^2$. To understand this correspondence between eigenvalues and principal series members on a conceptual level, identify H^+, the upper-half plane, with the coset space G/K and consider the induced representation $U^{\eta\kappa}$. (Here K is of course the maximal compact subgroup of all $\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$.) One shows that $U^{\eta\kappa}$ has a commutative commuting algebra and hence is uniquely a direct integral of irreducibles, each of which occurs once. Since the operator $y^2\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)$ commutes with all U^{η}_{κ}, and since there are no multiplicities, the decomposition of $U^{\eta\kappa}$ decomposes $y^2\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)$ as a direct integral of constant operators. The irreducibles that occur in the decomposition of $U^{\eta\kappa}$ are (modulo sets of measure zero) precisely the members of the principal series, and we have accordingly an assignment of an eigenvalue of $y^2\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)$ to each principal series member. It is this correspondence between eigenvalues of $y^2\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)$ and principal series members which makes it possible for Selberg to avoid talking about irreducible unitary representations of $SL(2, R)$. (An analogous correspondence involving a family of invariant differential operators permits Selberg to do likewise in his generalization to homogeneous spaces other than $SL(2, R)/K$.)

Selberg's paper centers about the existence and consequences of a formula (now known as the "Selberg trace formula") which he asserts can be considered as a generalization of the classical Poisson summation formula. In its most elementary form the latter asserts that $\sum_{n=-\infty}^{\infty} f(n) = \sum_{m=-\infty}^{\infty} \hat{f}(2\pi m)$ where $\hat{f}(y) = \int_{-\infty}^{\infty} f(x)\, e^{ixy} dx$ and f is a mildly restricted complex-valued function on the real line. Noting that $e^{iny} = 1$ for all integers n if and only if $y = 2\pi m$ for some integer m, one can rewrite this formula as $\sum_{\gamma\in\Gamma} f(\gamma) = \sum_{\chi\in\Gamma^{\perp}} \hat{f}(\chi)$, where Γ is the subgroup of the additive group R of the real line consisting of the integers, $\Gamma^{\perp}$ is the subgroup of the character group $\hat{R}$ consisting of all χ with $\chi(\gamma) = 1$ for all γ in Γ, and $\hat{f}(\chi) = \int_{-\infty}^{\infty} f(x)\chi(x)dx$. Once it is so written, there is an obvious generalization in which R is replaced by an arbitrary separable locally-compact commutative group G, and Γ is any discrete closed subgroup such that G/Γ is compact. (More generally still, Γ can be an arbitrary closed subgroup, and then the formula becomes $\int_{\Gamma} f(\gamma)d(\gamma) = \int_{\Gamma^{\perp}} f(\chi)d\chi$ for suitably chosen Haar measures in Γ and $\Gamma^{\perp}$.) In order to be

led naturally to a further generalization in which G can be non-commutative, one has only to look at the formula in the commutative case from the point of view of induced representations and their characters. Consider U^{I_Γ}, the representation of G induced by the trivial representation I_Γ of Γ. When Γ is discrete and G/Γ is compact, $\Gamma^\perp$ is also discrete, and it is easy to check that U^{I_Γ} is just the direct sum of the one-dimensional representations defined by the members of $\Gamma^\perp$. While Trace $(U^{I_\Gamma}_x)$ does not exist, one can define the character of U^{I_Γ} as a linear functional by the device already discussed in section 20. One forms $U^{I_\Gamma}_f$ and defines the character to be $f \to$ Trace $(U^{I_\Gamma}_f)$. Now an easy computation shows that Trace $(U^{I_\Gamma}_f) = \sum_{\gamma \in \Gamma} f(\gamma)$. Moreover, for each $\chi \in \Gamma^\perp$, the linear functional defining the character of the one-dimensional representation defined by χ is just $\int \chi(x) f(x) d\mu(x) = \hat{f}(\chi)$. In other words, the Poisson summation formula simply asserts the equality of the character of U^{I_Γ} (as a linear functional) to the sum of the characters of its irreducible constituents.

Now let G be any separable locally-compact group and let Γ be any closed discrete subgroup of G such that G/Γ is compact. Then $U^{I_\Gamma} = \sum n_j L^j$ where the L^j are distinct irreducible unitary representations of G. In this case one can hope to find a reasonably large family of functions f for which the two sides of the equation Trace $U^{I_\Gamma}_f = \sum n_j$ Trace L^j_f make sense and are equal. To the extent that one can do this, one will have a formula which is evidently a non-commutative generalization of the Poisson summation formula. For compact G/Γ, this formula is in essence Selberg's trace formula. However, Selberg introduced certain restrictions on G and f that made it possible to avoid various difficulties associated with not knowing all the L^j which might occur and with the (possible) non-existence of Trace (L^j_f). For Selberg, G was always a Lie group with a distinguished compact subgroup K such that G/K is a Riemannian manifold and U^{I_K} has a commutative commuting algebra. Moreover, Selberg only applied his trace formula to functions f which were constant on the $K : K$ double cosets. Under these circumstances, only those irreducible unitary representations L which contain the identity when restricted to K make a contribution to the right side of the trace formula. Moreover, each such L contains the identity of K just one time. Let φ be a unit vector such that $L_k(\varphi) = \varphi$ for all k in K. Then the function $x \to (L_x(\varphi) \cdot \varphi)$ is independent of the choice of φ and is called the *spherical function* associated with L. For an f which is constant on the $K : K$ double cosets, Trace L_f is equal to $\int f(x)(L_x(\varphi) \cdot \varphi)$ whenever the former exists, and the latter may be substituted for the former in general. For the Maass case ($G = SL(2, R)$, $K =$ all $\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$), the irreducible unitary representations which contain the identity when restricted to K are just the trivial representation and the members of the principal series which

are trivial on the center. For Selberg, the problem of finding which principal series members occur in the decomposition of $U^{\Gamma\Gamma}$ appeared as the problem of finding the eigenvalues of the operator $y^2\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)$ in the space of Γ orbits in the upper half-plane.

Given the importance of the classical Poisson summation formula in number theory (Dirichlet's method of evaluating Gauss sums and deducing quadratic reciprocity, the proof of the Jacobi inversion formula, functional equations for zeta and L functions, etc.), one can hope that a noncommutative generalization will have a host of interesting new number-theoretical consequences. Indeed, Selberg's paper was first found interesting not because it helped establish a connection between number theory and unitary group representations, but because of the immediate and prospective consequences of the trace formula for classical questions in number theory and the theory of modular forms. Among other things, Selberg found new relations among class numbers for binary quadratic forms, a new way of determining the dimensions of spaces of automorphic and modular forms, and perhaps most interesting of all, a way of computing the traces of the Hecke operators T_n which are so important in Hecke's theory of Dirichlet series with Euler products.

Actually, from the outset Selberg dealt with the generalizations of the trace formula described above, which one obtains by replacing the identity representation I_Γ with an arbitrary finite-dimensional unitary representation M of Γ. Moreover, in order to obtain his results on traces of Hecke operators, he generalized in another direction by replacing Trace U_f^M on the left-hand side by Trace AU_f^M where A is a rather special member of the commuting algebra $R(U^M, U^M)$. When $M = I_\Gamma$, there is a possible A associated with each $\Gamma : \Gamma$ double coset containing only finitely many right and left Γ cosets. It would take us too far afield to give further details here.

For applications to number theory, Selberg had to modify his theory to allow for the possible non-compactness of G/Γ. Already when $G = SL(2, R)$, the most interesting Γ's for number theoretical purposes are the so-called principal congruence subgroups Γ_N, and for these G/Γ_N is never compact. Accordingly, the induced unitary representation $U^{\Gamma\Gamma}$ (more generally U^M) is in part a discrete direct sum and in part a direct integral. In the special case in which $G = SL(2, R)$ and Γ is a subgroup of Γ_1 of finite index so that there are only finitely many "cusps," it turns out that $U^{\Gamma\Gamma}$ is the direct sum of two parts. One summand is a discrete direct sum of irreducibles, as in the case in which G/Γ is compact. The other part has one contribution from each cusp, each of which is a simple known direct integral of members of the principal series. Selberg showed how to take these contributions into account in his non-compact trace formula by making use of Maass's analogue of Eisenstein series. In defining his Eisenstein series for

non-analytic automorphic forms, Maass had found that these series converged only for inappropriate values of the eigenvalue parameter. But he could obtain what he needed by analytic continuation. This "analytic continuation of Eisenstein series" was important in Selberg's considerations as well. When one tries to deal with higher-dimensional groups G and spaces G/K for noncompact G/Γ, one can no longer compactify by adding a finite number of points. One has to add higher-dimensional spaces and the whole theory becomes much more complicated. Selberg gave important indications as to how to proceed but left many open problems.

It may appear at first sight that Selberg's trace formula can only give information about the non-analytic automorphic forms of Maass — at least in the special cases considered by Selberg. But let $G = SL(2, R) \times K^1$ where K^1 is the subgroup of $SL(2, R)$ consisting of all $\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$. Let K be the subgroup of $K^1 \times K^1 \subseteq SL(2, R) \times K^1$ consisting of all k, k. Then if $\chi_n\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} = e^{in\theta}$, the irreducible unitary representation $L \times \chi_n$ contains the identity when restricted to K if and only if L restricted to K^1 contains $\overline{\chi_n}$. Thus every discrete series member W occurs in the decomposition of U^{η_K} as $W \times \chi_n$ for some χ_n, and only occurs once with any particular $\overline{\chi_n}$. This choice of G and K satisfies Selberg's axioms and he obtained his results on classical automorphic forms by applying his theory to this case.

Selberg's work was translated into the language of the theory of unitary group representations and integrated with the earlier work of Gelfand, Fomin, and Maass in seminar reports by Godement and in the 1959 paper of Gelfand and Pjateskii-Shapiro cited above. In describing the picture which emerges, it is convenient to make use of the concept of an intertwining operator. Given unitary representations V and W of the same group G, an *intertwining operator* for V and W is by definition a bounded linear operator T from $\mathscr{H}(V)$ to $\mathscr{H}(W)$ such that $TV_x = W_xT$ for all x in G. Evidently the set $I(V, W)$ of all intertwining operators for V and W is a vector space. Its dimension is defined to be the *intertwining number* $i(V, W)$ of V and W. Of course, $I(V, V)$ coincides with $R(V, V)$, the commuting algebra of V. If $T \in I(V, W)$, let N_T denote the zero space of T and let $\overline{R_T}$ denote the closure of the range of T. It is obvious that N_T and $\overline{R_T}$ are closed invariant subspaces and easy to show that the subrepresentation of V defined by $N_T{}^{\perp}$ is equivalent to the subrepresentation of W defined by $\overline{R_T}$. Indeed, the most general intertwining operator for V and W is obtained by composing an equivalence between subrepresentations with members of $R(V, V)$ and $R(W, W)$ respectively. When V is irreducible, $i(V, V)$ is one-dimensional and it follows easily that for any W, $i(V, W)$ is the multiplicity with which V occurs as a discrete irreducible constituent of W. More generally, one can deduce information about the decomposition of unitary representations W

of unknown structure from information about the members of $I(V, W)$ where V has known structure.

In the special case of automorphic and modular forms defined in the upper half-plane, the principal facts connecting their theory with the theory of unitary group representations revolve around the structure of the induced representations U^{Γ} where Γ is a discrete subgroup of $G = SL(2, R)$. When G/Γ is compact or Γ has finite index in $SL(2, Z)$, then the multiplicity with which a member of the principal series of irreducible unitary representations of G is contained discretely in U^{Γ} is equal to the dimension of the space of all Maass cusp forms for the group Γ and a fixed eigenvalue λ. Similarly, the multiplicity with respect to which the discrete series member V^{2k} for $k = \pm 2, \pm 3, \cdots$ is contained discretely in U^{Γ} is equal to the dimension of the space of all ordinary cusp forms of weight k for the group Γ. Equivalently, one can say (in either case) that the dimension of a space of cusp forms is equal to the dimension of a space of intertwining operators. Moreover, it turns out that this statement can be strengthened to the statement that there is a canonical isomorphism of one space on another. Using this canonical isomorphism to identify cusp forms with intertwining operators, one can give group-theoretical definitions of the Petersson inner product and the Hecke operators. Let L be an irreducible unitary representation of $G = SL(2, R)$, and let T_1 and T_2 be members of $I(U^{\Gamma}, L)$. Then $T_1 T_2^*$ is a self-intertwining operator for L, and since L is irreducible this self-intertwining operator is a complex multiple $B(T_1, T_2)$ of the identity. The function $T_1, T_2 \to B(T_1, T_2)$ is the Petersson inner product. If A is any member of $I(U^{\Gamma}, U^{\Gamma})$, then for each T in $I(U^{\Gamma}, L)$, AT is in $I(U^{\Gamma}, L)$. Thus there is a natural ring homomorphism of $I(U^{\Gamma}, U^{\Gamma})$ into the space of linear operators in $I(U^{\Gamma}, L)$. As will be explained below, each $\Gamma : \Gamma$ double coset with suitable finiteness properties defines a member of $I(U^{\Gamma}, U^{\Gamma})$. The corresponding linear operators in $I(U^{\Gamma}, L)$ are the Hecke operators.

When G/Γ is not compact, there may be non-discrete components in U^{Γ}. However, at least when Γ is a subgroup of finite index of $SL(2, Z)$, these are all direct integrals with respect to Lebesgue measure of members of the principal series. They may be discovered by considering the intertwining operators of U^{N} with U^{Γ} where N is the nilpotent (actually commutative) subgroup of $SL(2, R)$ consisting of all $\left(\begin{smallmatrix} 1 & 0 \\ a & 1 \end{smallmatrix}\right)$. The structure of U^{N} is easily determined from two general theorems in the theory of induced representations. Let T denote the intermediate subgroup consisting of all $\left(\begin{smallmatrix} \lambda & 0 \\ a & 1/\lambda \end{smallmatrix}\right)$ and let W be the representation of T induced by I_N. By the theorem on inducing in stages, U^{N} is equivalent to U^{W}. On the other hand, since N is normal in T, W is the regular representation of T/N lifted to T. Since T/N is commutative, the regular representation of T/N is the direct integral of

all characters with respect to Haar measure. That U^{I_N} is a corresponding direct integral of members of the principal series (each repeated twice) follows at once. To find intertwining operators between U^{I_N} and U^{I_Γ}, one exploits the fact that for finite groups one has a complete overview of *all* intertwining operators for two induced representations. When both inducing representations are the identity, these intertwining operators correspond one-to-one to the complex-valued functions on G which are constant on the double cosets for the two subgroups involved. Thus there is a basis for the space of intertwining operators consisting of those whose corresponding function is identically one on some double coset and zero on all others. An explicit formula can be written down for each of these "double coset intertwining operators" which makes *formal* sense even when the group is infinite. Thus one can seek intertwining operators for U^{I_Γ} and U^{I_N} by investigating the convergence properties of the formal double coset intertwining operators associated with the $N : \Gamma$ double cosets. The double coset intertwining operator for $\Gamma \times N$ involves an integration over the coset space $N/x^{-1}\Gamma x \cap N$, and the double cosets divide sharply into two categories according to whether or not this coset space has a finite invariant measure. Those for which a finite invariant measure exists are of course those for which the formal double coset intertwining operators exist as actual operators, and are also just those for which $x^{-1}\Gamma x \cap N \neq \{e\}$. At least one such intertwining operator will exist precisely when Γ contains elements conjugate to members of $N - \{e\}$, i.e., when Γ contains "parabolic" elements. In considering the $N : \Gamma$ double cosets, it is possible to lump together those which are in the same $T : \Gamma$ double coset. Since T normalizes N, the different double coset intertwining operators for a given $T : \Gamma$ double coset are obtainable from one another in a trivial way. Now the homogeneous space G/T is identifiable with the one point compactification of the real line, i.e., with the *boundary* of the upper half plane H^+. Correspondingly, since $T : \Gamma$ double cosets may be identified with "boundary points" of the space of Γ orbits in H^+, those $T : \Gamma$ double cosets whose $N : \Gamma$ double cosets lead to well-defined intertwining operators, as indicated above, may be identified in this way with the "cusps" of the space of Γ orbits. Thus one has one intertwining operator for each cusp, and (at least when Γ is a subgroup of finite index of $SL(2, Z)$) these collectively account for the entire continuous part of U^{I_Γ}.

These considerations show how intimately the theory of automorphic forms is related to the decomposition theory of unitary representations of the form U^{I_Γ}. The Selberg trace formula may be regarded as the tool that makes it possible to exploit this relationship to obtain information about automorphic forms from knowledge of the irreducible unitary representations of $SL(2, R)$.

In the spirit of Hecke's 1928 paper, described in section 19, it is useful (and possible) to extend the theory just described by replacing I_Γ by a more general finite-dimensional irreducible unitary representation of Γ. In particular, when Γ is a subgroup of finite index of $\Gamma_1 = SL(2, Z)$, then U^{I_Γ} is a finite direct sum of induced representations of the form U^M where M is an irreducible constituent of the (finite-dimensional) representation of $\Gamma_1 = SL(2, Z)$ induced by I_Γ. Moreover, the work of Siegel and others on modular forms in several complex variables makes it desirable to develop a corresponding theory in which $SL(2, R)$ is replaced by $Sp(n)$. Finally, once things have been properly formulated in conceptual terms, there is no reason why one should not go on from $Sp(n)$ to general semi-simple Lie groups. Indeed, one might even hope for interesting new number theoretical applications of such an extended theory. This is an immense program on which much progress has been made, but which cannot be described in further detail here. Instead I shall confine myself to a few remarks. In a short note published in 1959, Harish-Chandra formulated a definition of automorphic forms which applied to quite general systems consisting of a semi-simple Lie group G and a discrete subgroup Γ. He also indicated a proof of the finite dimensionality of the space of all generalized automorphic forms of given "type." As a corollary he was able to show under rather general conditions that U^{I_Γ} contains each discretely contained component with finite multiplicity. In the same year, Gelfand and Graev published a long paper describing what they called the "horospherical method" or "the method of integral geometry" for decomposing unitary representations of semi-simple Lie groups of the form U^{I_K} where K is a maximal compact subgroup of a semi-simple Lie group G. It is not difficult to see that this method is essentially that of examining double coset intertwining operators as indicated above for U^{I_K} and U^{I_N} where N is a maximal nilpotent subgroup of G. Soon thereafter, Gelfand and Graev applied their method to U^{I_Γ} where Γ is a discrete subgroup of a semi-simple Lie group G. In 1962 they announced the following result: Let $\mathcal{M}$ be the orthogonal complement of the linear span of all double coset intertwining operators in $I(U^{I_N}, U^{I_\Gamma})$ for all maximal nilpotent subgroups N of G. Then U^{I_Γ} restricted to $\mathcal{M}$ is a discrete direct sum of irreducibles.

The connection between unitary representations and automorphic forms as described so far is unsatisfactory in one important respect. Although it "explains" the Hecke operators in group-representational terms, it does little to throw light on the Euler product decompositions which exist for the eigenfunctions of these Hecke operators. This gap was filled in in the 1960s with the aid of an extension of the idèle-adèle notion to non-commutative groups, inaugurated by Ono (1928—) and Tamagawa (1925—) in the late 1950s simultaneously with the translation of Selberg's ideas into the language of unitary group representations. For each prime p let G_p be the

group $SL(2, Q_p)$ of all 2×2 matrices with determinant one and coefficients in the field Q_p of all p-adic numbers (see section 20), and let $G_\infty = SL(2, R)$. Let K_p be the subgroup of G_p consisting of all members of $SL(2, Q_p)$ whose coefficients are in the ring of all p-adic integers. Equivalently, K_p may be defined as the closure in $SL(2, Q_p)$ of the subgroup $SL(2, Z)$. Then K_p is a compact open subgroup of $SL(2, Q_p)$, and one can consider the subgroup of the direct product $\prod_p SL(2, Q_p)$ consisting of all $\{x_p\}$ with $x_p \in K_p$ for all but finitely many K_p. We denote this subgroup by $\prod_p{}' SL(2, Q_p)$. It will be a separable locally-compact topological group if the subgroup $K = \prod_p K_p$ has the usual product topology and if $\prod_p{}' SL(2, Q_p)$ has the unique topology such that $K = \prod_p K_p$ is an open subgroup. The adèle group G_A for $SL(2, Q)$ is then defined to be the product group $\prod_p{}' SL(2, Q_p) \times SL(2, R)$. Let θ_p be the canonical imbedding of $SL(2, Q)$ as a dense subgroup of $SL(2, Q_p)$ and let θ_∞ be the canonical dense imbedding of $SL(2, Q)$ in $SL(2, R)$. Then $\gamma \to \{\theta_p(\gamma)\}, \theta_\infty(\gamma)$ is an injective homomorphism of $SL(2, Q)$ in the adèle group G_A, and the range of this homomorphism can be shown to be closed. Thus $SL(2, Q)$ appears in a natural way as a discrete subgroup of G_A. It is called the group of *principal adèles*. More generally, one can define such locally compact subinfinite product groups for every so-called "algebraic" subgroup of $GL(n, C)$ which is "defined over" an algebraic extension k of the rationals. This means that the group consists of all invertible $n \times n$ matrices which satisfy a certain finite set of polynomial equations with coefficients in k.

The theory of such *algebraic groups* was begun by Maurer (1859-1927) in 1894 and then apparently forgotten for a half-century. Chevalley and Tuan (1914—) returned to the subject in 1945, and in 1951 Chevalley published an extensive account as volume II of his treatise on Lie groups. Chevalley's treatment made heavy use of Lie algebra techniques, but in 1956 A. Borel (1923—) transformed the subject with a long paper showing how (with the help of certain ideas of Kolchin [1916—]) to analyze the structure of algebraic groups by means of global arguments involving algebraic geometry.

When Ono introduced non-commutative adèle groups in 1957, he was influenced not only by the revival of the theory of algebraic groups and Borel's paper in particular, but also by some ideas advanced by Eichler (1912—) a few years earlier concerning a possible unification of algebraic number theory with the theory of n-ary quadratic forms (see the introduction to a paper of Ono published in 1959). Ono found that certain finiteness theorems in algebraic number theory (such as the finiteness of the number of ideal classes) can be translated into statements about the idèle group of the field. These statements not only make sense for the adèle groups of alge-

braic groups, but when specialized to the adèle groups of the orthogonal groups they reduce to known finiteness theorems for quadratic forms. Ono suggested that one try to find unified proofs for the theorems about quadratic forms and algebraic number fields, respectively, by finding proofs of the general statements for the adèle groups of sufficiently comprehensive classes of algebraic groups. In his 1957 paper he defined adèle groups in general, but proved finiteness theorems only for commutative algebraic groups. In a second paper published in 1959 (and mentioned above) he was able to take care of certain solvable groups as well. More generally (as suggested by Ono) one could now attempt to encompass much of classical number theory in a much more general theory centering around the properties of the adèle groups of algebraic groups. This program turned out to be attractive and fruitful and was soon in a rapid state of development. One of its earliest successes was the discovery by Tamagawa and M. Kneser (1928—) that some of the main results of Siegel on quadratic forms (discussed earlier in this section) are essentially equivalent to the assertion that G_A/G_Q has measure 2 with respect to a canonically defined Haar measure in G_A. Here G_A is the adèle group associated with the orthogonal group of a non-degenerate rational quadratic form in at least three variables, and G_Q is the subgroup of principal adèles. This raised the question of generalizing Siegel's results by computing this measure (now called the Tamagawa number) for the adèle groups attached to other semi-simple algebraic groups. A detailed discussion, together with several Tamagawa number computations, appears in some 1961 lecture notes of Weil. Other aspects of Ono's program were worked out by Borel (with the assistance of some joint work with Harish-Chandra). For further details, the reader may consult the published version of Borel's 1962 address to the International Mathematical Congress in Stockholm. In the Proceedings of this same Congress, the reader will also find the texts of addresses by Gelfand and Selberg respectively. These describe the state of development of the theory of automorphic forms in general semi-simple Lie groups as seen in 1962 from two quite different perspectives.

Two years after the Stockholm Congress, the ideas described in Borel's address were combined with those described in the addresses of Gelfand and Selberg by considering extensions of automorphic forms from semi-simple Lie groups to the corresponding adèle groups and by replacing the study of unitary representations induced from discrete subgroups of semi-simple Lie groups to the study of unitary representations of adèle groups induced by the identity representation of the subgroup of principal adèles. The basic early publications include a short note published in 1964 by Gelfand, Graev, and Pjateskii-Shapiro and detailed papers published in 1964 by Weil and in 1965 by C. Moore (1937—). Moore on the one hand, and Gelfand, Graev, and Pjateskii-Shapiro on the other independently laid the foundations for

studying the unitary representation theory of infinite product groups of the adèle type but applied their results in different ways. Moore was concerned with nilpotent Lie groups and the structure of U^{Γ} where Γ is a discrete subgroup with a compact quotient. The other three authors were concerned with automorphic forms (although they did not say so in 1964) and their principal example was the adèle group associated with $SL(2, R)$. They made their intentions clear in 1966 when they published a book (vol. 6 of Gelfand's series *Generalized Functions*) entitled (in English translation) *Theory of Representations and Automorphic Forms.*

On the surface, Weil's paper seems almost unrelated to the considerations in the book of Gelfand et al., and it has yet to be integrated with those aspects of the unitary representation theory of adèle groups that tie in with the Hecke theory, which are described below. On the other hand, it relates unitary representations and adèle groups to number-theoretical problems in a most interesting way. Unfortunately, the results in it do not lend themselves to brief description; the reader is referred instead to my rather lengthy review of the paper in volume 29 of *Mathematical Reviews.* Let it suffice to state here that a key role in it is played by a certain natural projective representation of a generalization of the symplectic group, that the existence of this representation is implied by the generalization to locally-compact commutative groups of the Stone-von Neumann uniqueness theorem (see section 20), and that one of the main results may be looked upon as another generalization of the Poisson summation formula. The paper was written to prepare the way for a second paper, published in 1965, which contains a group-representational proof of Siegel's main results on quadratic forms imbedded in the generalization which specifies the Tamagawa number of most semi-simple algebraic groups. Roughly speaking, one may describe Weil's proof of Siegel's theorems as the result of using the connection between group representations and automorphic forms to replace automorphic forms by group representations in Siegel's proof.

To return to the adèle group G_A associated with $SL(2, R)$, let G_Q denote the group of principal adèles and consider the representation $U^{I_{G_Q}}$ of G_A induced by the one-dimensional identity representation of G_Q. A good way to understand the relevance of the study of $U^{I_{G_Q}}$ to the theory of modular forms in the upper half-plane is to consider the restriction of $U^{I_{G_Q}}$ to the factor group $e \times SL(2, R)$, where e is the identity of $\prod_p{}' SL(2, Q_p)$. If one first restricts $U^{I_{G_Q}}$ to the intermediate subgroup $\prod_p K_p \times SL(2, R)$, where K_p is the compact open closure in $SL(2, Q_p)$ of $SL(2, Z)$, and uses certain general theorems about induced representations (cf. pages 305-308 of [*15*] for further details), one finds that this restriction is a discrete direct sum $\Sigma \dim(L) U^L$. In this sum, L varies over a certain set of finite-dimensional ir-

reducible unitary representations of $\Gamma = SL(2, Z)$. The representations L which occur are precisely those which are continuous in the $SL(2, Q_p)$ topology for all p and include all those which reduce to the identity on the principal congruence subgroups Γ_N. Thus in analyzing $U^{I_{G_Q}}$, one is simultaneously analyzing a great many induced unitary representations of $SL(2, R)$ including all those of the form $U^{I_{\Gamma_N}}$.

To see the connection with Hecke's theory of Euler products, one has to bring together the following three components:

(a) The connection described above between modular forms and the decomposition of the representations U^L;

(b) the easily established fact that any decomposition of $U^{I_{G_Q}}$ as a direct sum or direct integral restricts down to a decomposition of each U^L and indeed to a decomposition of each factor component of each U^L; and

(c) the fact established by Gelfand et al. and Moore that each irreducible unitary representation of G_A defines and is defined by a sequence $\{M^p\}$, M^∞ where M^p is an irreducible unitary representation of $SL(2, Q_p)$ and M^∞ is an irreducible unitary representation of $SL(2, R)$.

The details are too complicated to give here, but the decompositions of the factor components of the U^L brought about by (b) are by (a) reflected in corresponding decompositions of spaces of modular forms with values in $\mathscr{H}(L)$. These decompositions are those defined by the Hecke operators, and the Euler product decomposition of an irreducible component is a reflection of the factorization of the corresponding irreducible representation of G_A as $\Pi M^p \times M^\infty$. In particular, the coefficients in the Euler factors are determined by the particular M^p's that occur.

Once the relationship between Euler products and the adèle group representation $U^{I_{G_Q}}$ is understood, one sees how to approach the question of extending the Hecke-Euler product theory to modular forms associated with more general semi-simple Lie groups. A strong motivation for undertaking such a program was provided in 1967 by independent work of Weil and Langlands (1936—). Weil showed how a relatively mild generalization of Hecke's theory might allow one to identify the zeta functions of elliptic curves with the Dirichlet series of modular forms, and Langlands indicated a path toward identifying the Artin L functions (see section 19) with the Dirichlet series arising in a generalization of Hecke theory to the general linear group of degree n. In Langland's case, n is the dimension of the irreducible representation of the Galois group parameterizing the Artin L functions. Langlands began his career by making extensive contributions to the development of the Selberg-Gelfand program, and for the past decade he has been the leader in developing the new program suggested by the

above considerations. For further details about the very complex theory which has emerged, the reader is referred to Borel's 1974-1975 Bourbaki seminar report [*1*].

Naturally, a program like Langland's cannot progress very far without knowledge of the irreducible unitary representations of the p-adic analogues of the semi-simple Lie groups. Mautner began the study of the unitary representations of these p-adic groups in 1958 with an attempt to settle the type I-ness question and a determination of some irreducible unitary representations. He was soon joined by Bruhat (1929—). The road was blocked for a while by insufficient understanding of the structure of the groups, but there is now a rather large literature on both structure and representations which I cannot even sketch here. A summary account will be found on pages 316-327 of [*15*]. As with physics, a few decades earlier, the needs of number theory have greatly stimulated the development of the theory of infinite-dimensional group representations.

23. SUMMARY AND CONCLUSION

With the preceding sketch of the nature of modern applications of the theory of unitary group representations to probability, physics, and number theory, we come to the end of our story. My central theme has been the power and scope of what I have called "the method of harmonic analysis." In addition there have been several unannounced subthemes. One of these is that physics is not so mysterious as many mathematicians seem to consider it. It is rather that physicists have different values and a different viewpoint, and this leads them to explain things in a manner uncongenial to mathematicians. If one works at it, it is possible to translate practically all of physics into well-defined mathematics. Moreover, when one does so, one finds a beautifully coherent scheme, which can be rather briefly summarized.

Another subtheme is that a rather large part of modern mathematics has developed in a natural way out of attempts to understand the solutions of quite simple equations. On the one hand, much of modern algebra and number theory has arisen out of attempts to understand the solutions in integers of equations of the form $Ax^2 + Bxy + Cy^2 = D$ where A, B, C, and D are known integers (and certain straightforward generalizations). On the other, much of modern analysis originated in attempts to deal with the partial differential equations that were encountered in physics as Newton's ideas were applied to continuous matter and as electricity, magnetism, and light began to be understood more quantitatively. The fact that the method of harmonic analysis is a key tool in dealing with both types of equations helps make it possible to see much more unity in mathematics and in mathematics and physics together than usually meets the eye.

In order to develop these themes, I have presented some of the main ideas and concepts of physics and number theory in more or less chronological sequence with emphasis on the impact of harmonic analysis. Before 1800 (except for Laplace's use of generating functions in probability) there was no systematic harmonic analysis, and both number theory and mathematical physics remained in a relatively primitive state. I have tried to emphasize the considerable progress made possible in both subjects by the introduction of Fourier analysis and its (unrecognized) analogue for finite commutative groups. At the end of the nineteenth century Frobenius invented group representations—under the indirect inspiration of the needs of number theory. Thirty years later Hermann Weyl recognized the essentially group-theoretical nature of Fourier analysis and observed that the theory of Frobenius was just the finite special case of an extension of Fourier analysis from commutative to non-commutative groups. The ensuing decades saw this new and enlarged concept of harmonic analysis produce advances in physics and in number theory comparable with those made over a century earlier by the original commutative version. These advances are still being made—especially in number theory. Indeed, it is possible to hope that startling progress will be made in classical problems once the intricate interaction between unitary group representation, automorphic forms, and number theory is better understood than it is at present.

The preceding account may seem to neglect such powerful tools as the theory of functions of a complex variable. As I have shown in the text, however, this latter theory can be regarded as an integral part of harmonic analysis, as can certain other techniques which are superficially rather different.

Although probability theory has been mentioned both as one of the fields in which harmonic analysis originated and as one of the fields to which the modern theory of group representations may be applied, the connection has developed in rather a different manner than has been the case for number theory and physics. Probability *theory* did not advance much during the nineteenth century (although many new applications were found). It rejoined the mainstream of modern mathematics when it was integrated with measure theory in the 1930s, and the main applications of harmonic analysis to it have been via the developments in the classical commutative theory made possible by the introduction of measure theory. While ergodic theory and the ergodic theorem are not usually thought of as being a part of harmonic analysis, I have taken pains in the text to show that in fact they are. It is this relatively new and undeveloped branch of harmonic analysis which has the most far-reaching connections with probability theory at present.

NOTES

1. Shortly after the typescript of this paper had been sent to the editors, I received a preprint of an article by K. I. Gross entitled "On the Evolution of Noncommutative Harmonic Analysis," scheduled to appear in the *American Mathematical Monthly*. The central theme of Gross's article is the same as that of this one. His execution is different in being more elementary, less than one fifth as long, and much less detailed. The reader may find it helpful to read Gross's treatment as an introduction to this one.

Sections 14 through 22 (about three-quarters of the paper) were written while I was a member of the Institute for Advanced Study in Princeton. I wish to express my gratitude to the Institute as well as to the following individuals who read all or part of the typescript and made helpful comments and corrections: Allan Adler, Armand Borel, Philip Green, Harish-Chandra, Howard Jacobowitz, Ian MacDonald, B. Simon, Robert Stanton, M. Taylor, David Vogan, and A. Weil.

2. To call the prime number theorem a conjecture of Riemann's is perhaps too loose a statement. Other mathematicians such as Gauss had suggested the truth of such a result rather earlier. Riemann's contribution was to suggest a method of proof.

3. Rayleigh himself suggested an *ad hoc* modification of his law which made it more reasonable at low temperatures and short wave lengths.

4. Since writing these paragraphs about Planck's contribution, I have examined the original papers and decided that the disclaimer "This of course is not how Planck proceeded" is an inadequate indication of the amount of "poetic" license I have taken in trying to make clear the essential point in Planck's discovery. This point is of course a lot clearer now than it was then. In actual fact Planck published two papers, a few months apart. In the first he found his now famous radiation law by making a (physically unmotivated) mathematical adjustment in Wien's derivation of Wien's law. He was not satisfied with his reasoning and published the law only because its predictions agreed with those of Wien's law at one end of the spectrum and with recent experimental results at the other. In the second paper he showed that the law could be derived in a more satisfying way by making his famous discreteness assumption. He does not talk about passing to the limit. It should also be mentioned that Planck did not refer to Rayleigh's law as such, but only to the experimental results confirming its validity at high temperatures. For a full account of the complex history of the old quantum theory, the reader is referred to Hermann [*12*] and Kuhn [*13*].

5. With this recognition of the group theoretical character of harmonic analysis, the further development of the subject proceeded along two semi-independent paths. The followers of one path concerned themselves with obtaining ever deeper and more refined results about harmonic analysis on the line and the circle (and later with extensions to n dimensions). The followers of the other path concerned themselves with extending the easier theorems to more general groups and to the new applications which this made possible. This paper does not pretend to be a complete history of harmonic analysis, but is concerned rather with harmonic analysis as a method for exploiting symmetry. Accordingly, I shall concern myself almost exclusively with the work of the followers of the second path, and with all due respect will say nothing about the work of such important mathematicians as Zygmund, Beurling, and Salem. For similar reasons, the reader will find no discussion of the generalization of harmonic analysis associated with the Gelfand map in the theory of Banach algebras.

6. I am informed by Professor Weil that his conversation with von Neumann took place before he had seen Koopman's paper and that he independently had the idea of introducing the representation V.

7. For details see pages 657-658 of [*14*].

8. For details see page 270 of [*15*].

BIBLIOGRAPHY

The references given below do not begin to exhaust the sources I have drawn upon in writing this historical survey. In addition to consulting many original papers, I have read articles in biographical dictionaries such as [6], obituary notices, survey articles, etc. Since much of this was done unsystematically and over a period of years, compiling a complete list of sources would be an impossibly difficult undertaking. I have chosen instead to give only a list of books and articles specifically referred to in the text or used especially extensively. I have not listed any of the original memoirs in which the various discoveries mentioned in the text were reported. There are far too many. The interested reader should have no difficulty tracking down those that interest him, using the date given in the text and such aids as collected works or abstracting and reviewing journals such as mathematical reviews. Items *14* and *15* contain two halves of a lengthy bibliography for sections 21 and 22.

[1] BOREL, A., *Formes Automorphes et Series de Dirichlet*]d'apres R.P. Langlands], Séminaire Bourbaki 1974/75, #466.

[2] BELL, E. T., *Men of Mathematics* (New York: Simon and Schuster, 1937).

[3] ———, *The Development of Mathematics,* 2nd ed. (New York and London: McGraw Hill, 1942).

[4] DAVENPORT, H., *Multiplicative Number Theory* (Chicago: Markham Publishing Co., 1967).

[5] FEIT, WALTER, "Theory of Finite Groups in the Twentieth Century," to appear in *Graduate Studies, Texas Tech University.*

[6] GILLISPIE, C. C., ed., *Dictionary of Scientific Biography.*

[7] GRATTAN-GUINNESS, I., *Joseph Fourier, 1768-1830* (Cambridge, Mass.: M.I.T. Press, 1972).

[8] HAWKINS, THOMAS, *Lebesgue's Theory of Integration: Its Origins and Development* (Madison: University of Wisconsin Press, 1970).

[9] ———, "The Origins of the Theory of Group Characters," *Archive for History of Exact Sciences* **7** (1970-71): 142-170.

[10] ———, "Hypercomplex Numbers, Lie Groups and the Creation of Group Representation Theory," *Archive for History of Exact Sciences* **8** (1971-72): 243-287.

[11] ———, "New Light on Frobenius' Creation of the Theory of Group Characters," *Archive for History of Exact Sciences* **12** (1974): 217-243.

[12] HERMANN, ARMIN, *The Genesis of Quantum Theory (1899-1913)* (Cambridge, Mass., and London: M.I.T. Press, 1971).

[13] KUHN, T. S., *The Black-Body Problem and the Quantum Discontinuity, 1894-1912* (New York and Oxford: Oxford Univ. Press, 1978).

[14] MACKEY, G. W., "Infinite Dimensional Group Representations," *Bull. Amer. Math. Soc.* **69** (1963): 628-686.

[15] ———, *The Theory of Unitary Group Representations* (Chicago and London: University of Chicago Press, 1976).

[16] WHITTAKER, E., *A History of the Theories of Aether and Electricity,* Vol. I, *The Classical Theories* (New York: Philosophical Library, 1951).

[17] ———, *A History of the Theories of Aether and Electricity, Vol. II, The Modern Theories (1900-1926)* (New York: Philosophical Library, 1954).

HARVARD UNIVERSITY