Representation Theory*

Its Rise and Its Role in Number Theory

Introduction

By representation theory we understand the representation of a group by linear transformations of a vector space. Initially, the group is finite, as in the researches of Dedekind and Frobenius, two of the founders of the subject, or a compact Lie group, as in the theory of invariants and the researches of Hurwitz and Schur, and the vector space finite-dimensional, so that the group is being represented as a group of matrices. Under the combined influences of relativity theory and quantum mechanics, in the context of relativistic field theories in the nineteen-thirties, the Lie group ceased to be compact and the space to be finite-dimensional, and the study of *infinite-dimensional representations* of Lie groups began. It never became, so far as I can judge, of any importance to physics, but it was continued by mathematicians, with strictly mathematical, even somewhat narrow, goals, and led, largely in the hands of Harish-Chandra, to a profound and unexpectedly elegant theory that, we now appreciate, suggests solutions to classical problems in domains, especially the theory of numbers, far removed from the concerns of Dirac and Wigner, in whose papers the notion first appears.

Physicists continue to offer mathematicians new notions, even within representation theory, the Virasoro algebra and its representations being one of the most recent, that may be as fecund. Predictions are out of place, but it is well to remind ourselves that the representation theory of noncompact Lie groups revealed its force and its true lines only after an enormous effort, over two decades and by one of the very best mathematical minds of our time, to establish rigorously and in general the elements of what appeared to be a somewhat peripheral subject. It is not that mathematicians, like cobblers, should stick to their lasts; but that humble spot may nevertheless be where the challenges and the rewards lie.

J. Willard Gibbs, too, offered much to the mathematician. An observation about the convergence of Fourier series is classical and now known as the Gibbs phenomenon. Of far greater importance, in the last quarter-century the ideas of statistical mechanics have been put in a form that is readily

* Appeared in Proceedings of the Gibbs Symposium, AMS (1990).

accessible to mathematicians, who are, as a community, very slowly becoming aware of the wealth of difficult problems it poses. His connection with representation theory is more tenuous. In the period between his great researches on thermodynamics and statistical mechanics, he occupied himself with linear algebra, even in a somewhat polemical fashion. In particular, he introduced notation that was more than familiar to mathematics students of my generation, but that perhaps survives today only in electromagnetic theory and hydrodynamics.

In these two subjects, the dot product $\boldsymbol{\alpha}\cdot\boldsymbol{\beta}$ and the cross product $\boldsymbol{\alpha}\times\boldsymbol{\beta}$ introduced by Gibbs are ubiquitous and extremely convenient. The usual orthogonal group comes to us provided with a natural representation because it is a group of 3×3 matrices, and all objects occurring in elementary physical theories must transform simply under it.The two products of Gibbs appear because the tensor product of this natural representation with itself is reducible and contains as subrepresentations both the trivial representation, which assigns 1 to every group element, and, if attention is confined to proper rotations, the natural representation itself. Tensor products play an even more important role, implicitly or explicitly, in spectroscopy, and it is their explicit use for the addition of angular momenta, especially by Wigner, that drew representation theory so prominently to the attention of physicists.

The use of compact groups and their Lie algebras became more and more common, especially in the classification of elementary particles, often with great success, as a terminology peppered with arcane terms from occidental literature and oriental philosophy attests, but the investigation of the representations of groups like the Lorentz group, in which time introduces a noncompact element, was left by and large to mathematicians. To understand where it led them, we first review some problems from number theory.

Number Theory

Two of the most immediate and most elementary aspects of number theory that are yet at the same time the most charged with possibilities are *diophantine equations* and *prime numbers*. Diophantine equations are equations with integral coefficients to which integral solutions are sought. A simple example is the equation appearing in the pythagorean theorem,

$$x^2 + y^2 = z^2, \tag{a}$$

with the classic solutions,

$$3^2 + 4^2 = 5^2, \qquad 5^2 + 12^2 = 13^2,$$

that were in my youth still a common tool of carpenters and surveyors. It is also possible to study the solutions of equations in fractions and to allow the coefficients of the equations to be fractions, and that often amounts to the same thing. If $a = x/z$ and $b = y/z$ then the equation (a) becomes

$$a^2 + b^2 = 1.$$

Prime numbers are, of course, those like $2, 3, 5$ that appear in the factorization of other numbers, $6 = 2 \cdot 3$ or $20 = 2 \cdot 2 \cdot 5$, but that do not themselves factor. They do not appear in our lives in the homely way that simple solutions of diophantine equations once did, but their use in cryptography, in for example public-key cryptosystems, has perhaps made them more a subject of common parlance than they once were.

The two topics may be combined in the subject of *congruences*. Consider the equation $x^2 + 1 = 0$. If p is any prime number we may introduce the congruence,

$$x^2 + 1 \equiv 0 (\operatorname{mod} p).$$

A solution of this is an integer x such that $x^2 + 1$ is divisible by p. Thus,

$$\begin{aligned}
x &= 1, \qquad 1^2 + 1 = 2 \equiv 0\,(\operatorname{mod} 2),\\
x &= 2, \qquad 2^2 + 1 = 5 \equiv 0\,(\operatorname{mod} 5),\\
x &= 3, \qquad 3^2 + 1 = 10 \equiv 0\,(\operatorname{mod} 2, \operatorname{mod} 5),\\
x &= 4, \qquad 4^2 + 1 = 17 \equiv 0\,(\operatorname{mod} 17),\\
x &= 5, \qquad 5^2 + 1 = 26 \equiv 0\,(\operatorname{mod} 2, \operatorname{mod} 13),\\
x &= 6, \qquad 6^2 + 1 = 37 \equiv 0(\operatorname{mod} 37).
\end{aligned}$$

The list can be continued and the regularity that is already incipient continues with it. The primes for which the congruence can be solved are 2 and $5, 13, 17, \ldots$ all of which leave the remainder 1 upon division by 4, whereas primes like $7, 19, 23, \ldots$ that leave the remainder 3 upon division by 4 and that have not yet appeared never do so; the congruence is not solvable for them.

This regularity, at first blush so simple, is the germ of one of the major branches of the higher number theory and was the central theme of a development that began with Euler and Legendre in the eighteenth century, and continued down to our own time, with contributions by Gauss, Kummer, Hilbert, Takagi, and Artin. Current efforts to extend it will be the theme of this essay.

We shall be concerned with more complex congruences, involving, for example, two unknowns and our concern will be strictly mathematical, but I mention in passing that these more abstruse topics may also impinge on our daily life, even disageeably, since the sophisticated theory of congruences modulo a prime is exploited in coding theory and thus in the transmission of information (and of misinformation, not to speak of unsolicited sounds and images that are simply unpleasant). It is sometimes also necessary, and even important, to consider congruences modulo integers that are not prime, not only for strictly theoretical purposes but also for ends that may be regarded as more practical. The difference between congruences modulo primes and modulo composite numbers is great enough that it may be used to very good effect in the testing of numbers for primality, and so can be exploited in cryptography.

Zeta-functions

The most familiar of the zeta-functions, and the one that has given its name to the others, is that to which the name of Riemann is attached,

$$\begin{aligned}\zeta(s) &= \sum_{n=1}^{\infty} \frac{1}{n^s} \\ &= \Big(1 - \frac{1}{2^s}\Big)^{-1}\Big(1 - \frac{1}{3^s}\Big)^{-1}\Big(1 - \frac{1}{5^s}\Big)^{-1}\dots \\ &= \prod_p \frac{1}{1 - \frac{1}{p^s}}\end{aligned}$$

The equality, due to Euler, is obtained by applying the expansion

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \dots$$

to

$$\frac{1}{1 - \frac{1}{p^s}}$$

to obtain

$$1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \dots,$$

and then recalling that every positive integer can be written in one and only one way as a product of prime powers.

The principal use of the Riemann zeta-function is for the study of the distribution of the primes amongst the integers. Observing that the sum

$$1 + \frac{1}{2^s} + \frac{1}{3^s} + \dots$$

behaves like the integral

$$\int_1^\infty \frac{1}{x^s} ds,$$

we conclude correctly that the series defining the zeta-function converges for $s > 1$. To use it to any effect in the study of the distribution of primes it is, however, necessary to define and calculate it for other values of s.

There are manifold ways to do this. The Γ-function, defined by the integral

$$\Gamma(s) = \int_0^\infty e^{-t} t^{s-1} dt$$

when $s > 0$, is readily shown upon an integration by parts to satisfy the relation $\Gamma(s) = (s-1)\Gamma(s-1)$. It can therefore be defined and its value calculated for any s.

To deal with the zeta-function, introduce two further functions defined by infinite series:

$$\phi(t) = \sum_{n=1}^\infty e^{-\pi n^2 t};$$

$$\theta(t) = 1 + 2\phi(t) = \sum_{-\infty}^\infty e^{-\pi n^2 t}.$$

Calculating the integral,

(b) $$\int_0^\infty \phi(t) t^{\frac{s}{2}-1} dt,$$

term by term, we obtain

$$\pi^{-\frac{s}{2}} \Gamma\left(\frac{s}{2}\right) \zeta(s).$$

As a consequence, to calculate the zeta-function we need only calculate the integral (b) for all values of s. The integrand behaves badly at $t = 0$ if $s \le 1$, so that as it stands the integral still does not serve our purpose.

We next exploit the elements of Fourier analysis, using a device to which Poisson's name is attached. We observe that $\theta(t)$ is the value at $x = 0$ of the periodic function

$$\sum_{n=-\infty}^\infty e^{-\pi(n+x)^2 t}.$$

It is easy enough to calculate the Fourier expansion of this function. Doing so, and using the expansion to calculate its value at $x = 0$, we discover that

(c) $$\theta(t) = \frac{1}{\sqrt{t}} \theta\left(\frac{1}{t}\right).$$

With this relation in hand we return to the integral (b), replace it by two integrals, one from 0 to 1, and one from 1 to ∞. The second is defined for any value of s. In the first, we replace $\phi(t)$ by

$$\frac{1}{2}\left(t^{-\frac{1}{2}}-1\right)+t^{-\frac{1}{2}}\phi\left(\frac{1}{t}\right).$$

The first two terms yield integrals that can be integrated explicitly, the result being

$$\frac{1}{s-1}-\frac{1}{s},$$

which gives the anticipated pole at $s=1$. In the last term, we substitute $1/t$ for t obtaining an integral from 0 to ∞ that is readily seen to converge for any s.

Since the function $\phi(t)$ decreases so rapidly as $t \to \infty$, the two integrals that we have not calculated explicitly are easily evaluated to any degree of accuracy. This is the main lesson to be drawn from this technical digression. We have not yet linked diophantine equations to zeta-functions, but when we do, we shall see that as a result of a circle of conjectures and theorems the use of zeta-functions offers a way of deciding whether a given diophantine equation has a solution that can be thousands of times more effective than even sophisticated searches for it. For this, however, it is necessary to be able to calculate their values at certain points with precision, although it need not be very great. If our experience in this century is any guide, this can be done only by showing that the zeta-functions attached to diophantine equations are equal to others, defined ultimately in terms of representations of noncompact groups, that are amenable to the same analysis as the Riemann zeta-function.

A classic example, in which the use of infinite-dimensional representations is unnecessary, is provided by the equation $x^2+1=0$. We attach to it the function,

$$L(s)=\frac{1}{1+\frac{1}{3^s}}\cdot\frac{1}{1-\frac{1}{5^s}}\cdot\frac{1}{1+\frac{1}{7^s}}\cdots,$$

the general factor being

$$\frac{1}{1\mp\frac{1}{p^s}},$$

according as the congruence

$$x^2+1\equiv 0 \pmod p$$

does or does not have a solution. This is a series defined by a diophantine equation. The notation $L(s)$ is due to Dirichlet and one often speaks of L-function rather than zeta-function, following conventions that are unimportant here.

On the other hand, by the regularity whose importance we have already been at pains to stress, the general factor could also be defined by the alternative that p leave the remainder $+1$ or the remainder -1 upon division by 4. A Dirichlet character χ is a multiplicative function on the integers, thus a function satisfying $\chi(ab) = \chi(a)\chi(b)$, and such that $\chi(a)$ depends only on the remainder after division of a by some positive integer n. For simplicity, one customarily takes $\chi(a) = 0$ if a and n have a divisor in common. For example a Dirichlet character modulo 4 is given by:

$$\chi(1) = 1; \chi(2) = 0; \chi(3) = -1; \chi(4) = 0.$$

The general factor of the product defining $L(s)$ is then

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

Expanding the product defining $L(s)$ just as we expanded the product for the zeta-function we see that

$$\begin{aligned} L(s) &= 1 - \frac{1}{3^s} + \frac{1}{5^s} - \frac{1}{7^s} + \frac{1}{9^s} - + \dots \\ &= \sum_{n=0}^{\infty} \frac{\chi(n)}{n^s}. \end{aligned}$$

Since the numerator of this series is a periodic function of n we can once again use elementary Fourier analysis to define and calculate $L(s)$ for any s.

Because it is so important for the susequent discussion, I repeat that the function $L(s)$ is originally defined by a product of factors determined according to a diophantine alternative, simple though it be. At first glance there is no reason to think that the function has a meaning outside the region, $s > 1$, where the product obviously converges, but thanks to the regularity that we have observed, the product can be converted to a series that is defined by a periodic function and that can be put in a form that has a meaning for any s.

We can not expect the regularity always to be so simple; for then it would not have been so elusive. Two basic classes of examples serve as an introduction to the general problem; both illustrate its difficulty. The first, equations with icosahedral Galois groups, has a greater historical appeal since they are the simplest equations completely inaccessible to the classical theory of equations with abelian Galois groups; the solution that appears to be correct for them was suggested directly by ideas originating in representation theory. The second, equations in two variables defining elliptic curves, illustrates, however, far more cogently to a nonspecialist the value of L-functions for diophantine equations and is one of the earliest and most striking uses of the computer as an aide to pure mathematics; here the solution that appears to be correct was suggested earlier and on different grounds. Both solutions turn out to be part of the same larger pattern, and although the evidence for them is extensive and overwhelming, neither has been established in any generality. We begin with equations in a single variable.

Galois groups

In our discussion of the congruence $x^2 + 1 \equiv 0 \pmod p$, we emphasized the search for solutions, but this is tantamount to the search for factorizations,

$$x^2 + 1 \equiv x^2 + 2x + 1 = (x + 1)^2 \pmod 2,$$
$$x^2 + 1 = x^2 - 4 + 5 \equiv (x - 2)(x + 2) \pmod 5.$$

The polynomial $x^2 + 1$ cannot be factored modulo 7 or 11 or any prime that leaves the remainder 3 upon division by 4, but factors into two linear factors modulo any prime leaving the remainder 1.

Another example, the reasons for whose choice will be explained later, is

$$x^5 + 10x^3 - 10x^2 + 35x - 18. \tag{d}$$

It is irreducible modulo p for $p = 7, 13, 19, 29, 43, 47, 59, \ldots$ and factors into linear factors modulo p for $p = 2063, 2213, 2953, 3631, \ldots$. These lists can be continued indefinitely, but it is doubtful that even the most perspicacious and experienced mathematician would detect any regularity. It is none the less there.

To explain why we have chosen this equation and not another as an example, we first consider the general equation in one variable,

$$x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \ldots + a_0 = 0,$$

in which all coefficients are rational numbers. This equation will have roots $\theta_1, \ldots, \theta_n$ and there will be various relations between these roots with coefficients that are also rational,

$$F(\theta_1, \ldots, \theta_n) = 0.$$

For example, the roots of $x^3 - 1 = 0$ are $\theta_1 = 1, \theta_2 = (-1 + \sqrt{-3})/2, \theta_3 = (-1 - \sqrt{-3})/2$ and two of the many valid relations are:

$$\theta_1 = 1; \qquad \theta_2\theta_3 = \theta_1.$$

We can associate to the equation the group of all permutations of its roots that preserve all valid relations. In the example the sole possibility in addition to the trivial permutation is the permutation that fixes θ_1 and interchanges θ_2 and θ_3. This group is extremely important, and is known as the $Galois$ $group$ of the equation. We denote it by G.

For simplicity (it is easy to achieve) suppose that the equation has no multiple roots and that the coefficients are integers, and introduce the discriminant,

$$\Delta = \prod_{i \neq j} (\theta_i - \theta_j).$$

It is an integer and it is fundamental to the theory of diophantine equations that, following Dedekind and Frobenius, we can attach to any prime p that does not divide Δ an element F_p in G that determines among other things how the equation factors modulo p. More precisely, it is the conjugacy class of F_p within the group G that is determined, and that suggests, as is indeed the case, that to define F_p a little theory is necessary.

It also suggests the use not of F_p itself but of the trace of the matrix $\rho(F_p)$, where ρ is some (finite-dimensional) representation of the group G, for the numbers $trace(\rho(F_p))$ taken for all ρ determine the class of F_p. We can also, fixing ρ, consider the matrices $\rho(F_p)$ or, better, their conjugacy classes.

Rather than asking how the factorization of the original equation $f(x) = 0$ varies with p and whether it manifests any regularity, we can ask whether, for a given ρ, the conjugacy classes $\rho(F_p)$ do. The simplest possibility is that ρ is one-dimensional, and this is the classical theory; $\rho(F_p)$ is then a number that is determined by the remainder left upon division of p by a certain integer n that depends on the equation and on ρ.

The next possibility is that ρ is a representation by 2×2 matrices that we may suppose unitary. Recalling the close relation between the group of unitary matrices in two variables and the group of proper rotations in three variables, the second being a homomorphic image of the first, we classify finite subgroups of the unitary group by their image in the group of proper rotations. Taking the finite subgroup to be $\rho(G)$ and excluding the possibility that ρ is reducible, we obtain dihedral, tetrahedral, octahedral, and icosahedral representations. Dihedral representations can be treated by the classical theory; tetrahedral and octahedral representations require modern ideas but have been completely dealt with [L2]. Icosahedral equations, however, remain intractable in general, and even for specific examples verifying numerically the validity of the regularity suggested is a task that requires great skill and ingenuity and is by no means assured of success.

The first step is to find examples of equations with icosahedral Galois groups, and that requires a computer search of equations of degree 5. Two possibilities that present themselves are:

$$x^5 + 10x^3 - 10x^2 + 35x - 18 = 0\,;$$
$$x^5 + 6x^3 - 12x^2 + 5x - 4 = 0\,.$$

It is the first that has been studied closely and not the second, although it has smaller coefficients, so that one might believe calculations with it would be more efficient. A peculiar aspect, however, of the theory of diophantine equations is that a great deal of theory is required in order to recognize which are simpler. Neither the size of the coefficients nor the form of the equation is a guide.

For equations in one variable there are two criteria: the Galois group and the conductor. We have already chosen the Galois group as simple as possible if the equation is to offer difficulties. The conductor is a positive integer related to the discriminant but more complicated. To calculate it requires a painstaking examination of the properties of the equation modulo powers, often quite high, of the primes that divide the discriminant, but with enough time and effort that can always be done. The conductor of the first of our two equations is 800 and that of the second 4256. This difference in size entails a great reduction in the number of calculations, and not being an expert I am not even sure whether the second equation is accessible to numerical investigations.

A good-sized monograph [Bu] was found necessary to explain the methods used to establish the proposed regularity for the first. This regularity, which is expressed in terms of modular forms, has yet to be described; we have first to establish a conviction that it is necessary, for it is of a transcendental nature, and simply asserts the equality of two quite differently defined sequences. The criterion that the statement of regularity has to fulfill in order to be accepted as significant is analytic. For equations in a single variable, its importance is obscure except in a theoretical context like that of Emil Artin's papers of the nineteen-twenties, in which the representation theory of finite groups was fused with the notion of an L-series. For elliptic curves the importance will be clearer .

It is curious that, although one of the two currents that merged in Artin's papers, representation theory, owed its very existence to Richard Dedekind and F. G. Frobenius [Ha], who also contributed essential ideas to the study of L-series and the theory of equations, it is not clear to what extent the rise of representation theory was a response to problems posed by L-series. The correspondence between

Dedekind and Frobenius is extant, and the letters of Dedekind have been published [D], but it would be imprudent for a reader inexperienced in the ways of historical research to read too much into them.

If G is the icosahedral group attached to the equation (d) and ρ its two-dimensional representation, then for all primes p that do not divide the conductor we can form the 2×2 matrix $\rho(F_p)$, which has two eigenvalues, α_p and β_p. The L-function of Artin is then

$$L(s) = \prod_p \frac{1}{1 - \frac{\alpha_p}{p^s}} \cdot \frac{1}{1 - \frac{\beta_p}{p^s}},$$

the primes dividing the conductor being omitted from the product. This product converges for $s > 1$ and one of the simplest cases of the problem posed by Artin in 1923 is to show that as an analytic function of a complex variable it can be defined for all s. For equation (d), this was first achieved in the monograph [Bu] in 1970. It was a critical test of the general ideas formulated in [L1].

Elliptic Curves

As examples of elliptic curves we take the equation in two variables,

$$y^2 = x^3 + Dx, \tag{e}$$

borrowing from the account in a lecture of G. Harder before the Rheinische-Westfäl- ische Akademie der Wissenschaften [H]. Take $D = (6577)^2$. The equation has the solution $x = 0, y = 0$. Does it have further solutions with x and y both rational? One way to attempt to answer this question is to conduct a computer search.

An intelligent search entails an appeal to the theory of the equation, in the hope of replacing x and y by numbers with possibly smaller numerators and denominators. The first step is to set

$$x = \frac{y_1^2}{4x_1^2}; \qquad y = \frac{y_1(x_1^2 + 4D)}{8x_1^2}.$$

Then we put

$$x_1 = lt^2; \qquad y_1 = lst,$$

and finally

$$t = \frac{U}{2V}; \qquad s = lW(2V)^2; \qquad l = 6577.$$

If U, V and W are integers satisfying the equation

$$U^4 - 64V^4 = 6577W^2,$$

then x and y will be rational solutions of the original equation. Further transformations of a similar nature but with more complex equations are made, and finally, at some point, a more or less blind search for integers U, V, W satisfying the new equation begins. That takes time,apparently some twenty minutes on an IBM 370 if no false starts are made, and ultimately a triple is found.

$$U = 10\,500\,084\,257\,375\,984\,596\,799$$

$$V = 1\,980\,407\,963\,453\,953\,023\,564$$

$$W = 1\,303\,262\,616\,226\,128\,053\,329\,966\,805\,106\,514\,807\,822\,601$$

Of course, there is no need except as a mathematical amusement to find solutions to the equation (e), but the problem of finding some method more efficient than promiscuous searching to decide whether it has a solution has an immediate appeal that needs no further justification. That such a method may exist is, nevertheless, an astonishing discovery due to Bryan Birch and Peter Swinnerton-Dyer [BSD], one of the earliest uses of the computer in pure mathematics and perhaps still that with the deepest implications.

Their investigations were partly inspired by the study of congruence zeta-functions, which also originated with Dedekind and Artin but whose evolution was strongly determined by the geometrical and topological ideas that infused the mathematics of the middle decades of this century. More than anything else, it is these functions that have shaped current number theory.

As a simple example we take $D = -1$ in equation (e) and then for each prime number p, except 2 and 3, we count the number of solutions modulo p. For example, if $p = 5$ we simply list the possibilities for x and y and substitute them in the equation to see if it is satisfied modulo p.

$y\backslash x$	0	1	2	3	4
0	+	+	−	−	+
1	−	−	+	−	−
2	−	−	−	+	−
3	−	−	−	+	−
4	−	−	+	−	−

The plus sign indicates that the congruence is satisfied and the minus sign that it is not. There are seven plus signs so that the number of solutions of the congruence is $N_5 = 7$. For any prime p we can count N_p in the same way and define α_p and β_p by the conditions:

$$N_p = p + \alpha_p + \beta_p; \quad \beta_p = \bar{\alpha}_p; \quad \alpha_p\beta_p = p.$$

These conditions will define α_p and β_p only if $|N_p - p| \leq 2\sqrt{p}$. This is so, and can be proved by considerations that began with Gauss, but the relation is also a special case of a powerful general theorem whose statement and proof are one of the great triumphs of the application to number theory of the topological ideas appearing during this century.

Given the two sequences α_p and β_p, for any D but now in particular for $D = (6577)^2$, we can introduce, in what has become almost a reflex action, the function

$$L(s) = \prod_p \frac{1}{1 - \frac{\alpha_p}{p^s}} \cdot \frac{1}{1 - \frac{\beta_p}{p^s}}, \tag{f}$$

omitting $p = 2$ and $p = 3$ from the product. Because the equation (e) has a symmetry $x \to -x,\ y \to iy$, it satisfies an analogue of the first criterion for simplicity, which was that the Galois group be commutative. I forego attempting an exact formulation; the upshot is that there is a regularity that allows this function, like the zeta-function of Riemann and the L-functions defined by Dirichlet characters, to be calculated in a fraction of a second for any s and not just in the region $s > \frac{3}{2}$ where the product (f) converges.

Its value at $s = 1$ is of special interest. According to a special case of the conjecture of Birch and Swinnerton-Dyer, but not one that has been established, the equation (e) must have a solution if

$$L(1) = 0.$$

This is readily decidable since $L(1)$ should be an integer times

$$\frac{1}{4} \int_0^\infty \frac{(\sqrt{x^3 + Dx})}{dx}.$$

Judgements as to the value of a given mathematical theorem are diverse and various criteria may be applied, but one that mathematicians of all casts accept as decisive is that in concrete classical problems it replaces a laborious calculation with uncertain results with one that is rapid, at least with modern tools, and certain. The conjecture of Birch and Swinnerton-Dyer achieves this, for it predicts that the existence of nontrivial solutions to certain equations is decided by the values of zeta-functions, and Yutaka Taniyama had already proposed [Sh] a method for calculating these values. Taniyama's proposal was not known, however, to Birch and Swinnerton-Dyer who tested their conjecture on elliptic equations that like (e) are sufficiently simple, in the appropriate sense, to be susceptible to a more classical treatment.

Not long after the appearance of the conjectures of Birch and Swinnerton-Dyer the proposal of Taniyama was taken up by Andre Weil, who made it more precise, and it has since been subject to a thorough numerical testing, which curiously enough is much more readily performed for elliptic curves than for icosahedral equations. Once again the apparently simpler equations are the more difficult of access. The regularity in the numbers N_p implied by Taniyama's proposal is ultimately of the same nature as that for the traces $\alpha_p + \beta_p$ of the elements F_p in the Galois group of an icosahedral equation, but it has a more immediate geometric meaning.

A typical form for the equation of an elliptic curve is:

$$y^2 = x^3 + ax^2 + bx + c. \tag{g}$$

Then $\int \frac{dx}{y}$ is an elliptic integral and that is of course the source of the terminology.

For these equations too, there is a second criterion for simplicity, the conductor that can be calculated by examining the associated congruence modulo high powers of primes. The conductor of the equation,

$$y^2 = x^3 - x^2 + \frac{1}{4},$$

is for example 11 and this is the smallest conductor that can be achieved. It is surprisingly small.Replacing the variable y by $y + \frac{1}{2}$, we may also write the equation as

$$y^2 + y = x^3 - x^2. \tag{h}$$

The equation is also surprisingly simple. We change it in two steps. We first set, following Vélu [V],

$$X = x + \frac{1}{x^2} + \frac{2}{x-1} + \frac{1}{(x-1)^2};$$
$$Y = y - (2y+1)\left(\frac{1}{x^3} + \frac{1}{(x-1)^3} + \frac{1}{(x-1)^2}\right)$$

If x and y satisfy (g) then the new quantities satisfy the equation,

$$Y^2 + Y = X^3 - X^2 - 10X - 20, \tag{i}$$

which, although slightly more complicated, is still elliptic. There is clearly a close relation between solutions of the two equations. If we then set,

$$\sigma = \frac{2Y+1}{11}, \qquad \tau = -\frac{X-5}{11},$$

we obtain two quantities satisfying the equation,

$$\text{(j)} \qquad \sigma^2 = 1 - 20\tau + 56\tau^2 - 44\tau^3,$$

which ceases altogether to recommend itself by its simplicity. Its advantage is that it appeared in an altogether different context long ago, in Klein's lectures on modular forms ([KF], p.440).

As a consequence, although this was understood only much later after the researches of M. Eichler and G. Shimura in the nineteen-fifties, the L-series of the curve (h), (i), or (j) — they are all equal — is readily calculated, and provides the simplest example of the method proposed by Taniyama.

Modular Curves

The set of all complex solutions to the equation (j) forms a surface, because for each complex value of τ there are in general two possible values of σ. This surface can be obtained in another way from the upper half-plane of complex numbers $z = x + iy$, $y > 0$. If

$$\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is a real matrix of determinant 1 and z lies in the upper half-plane, then $\gamma z = (az + b)/(cz + d)$ again lies in the upper half-plane. Consider the group Γ of integral matrices γ that have determinant 1 and for which c is divisible by 11.

In the lectures of Klein, as well as in the later treatise of Fricke([F], p. 406, [Li], Section 4) two meromorphic functions (thus complex-analytic except for poles), $\sigma(z), \tau(z)$, on the upper half-plan are constructed such that $\sigma = \sigma(z), \tau = \tau(z)$ is a point on the curve (j) for any z, provided that we admit ∞, ∞ as such a point. Furthermore, all points of the curve except $\tau = 0, \sigma = \pm 1$ are obtained in this way, and two values z and z' yield the same point if and only if there is a γ in the group Γ such that $z' = \gamma z$.

If we replace z by the variable $q = exp(2\pi iz)$, then $\frac{1}{\sigma}\frac{d\tau}{dq}$ has an expansion,

$$\frac{1}{q}\sum_{n=1}^{\infty} A_n q^n,$$

and the function,

$$L(s) = \sum_{n=1}^{\infty} \frac{A_n}{n^s},$$

is, according to Shimura, equal to the L-function attached to the elliptic curve (j) and to (h) and (i), so that its value is easily calculated for any value of s

Our notions of simplicity are once again overturned. If one avoids equations that are in any sense abelian then some equations in two variables, those of elliptic curves, appear to be simpler than those in one variable. They exist with much smaller conductor, and the connection with modular functions, like σ and τ, is much more immediate.

This is not peculiar to the example (j). If

$$f(z) = \frac{1}{\sigma}\frac{d\tau}{dz},$$

then, for any γ in the group Γ

$$f(\gamma z) = (cz + d)^2 f(z). \tag{k}$$

Such a function is called a modular form of weight 2 with respect to the group Γ. If the exponent 2 is replaced by 1 the equation (k) defines modular forms of weight 1. Forms of weight 2 are simply derivatives of invariant functions and therefore much easier to treat than forms of weight 1. Icosahedral equations are so much more difficult because their regularity is at best expressed by forms of weight 1.

Taniyama's suggestion, as we now understand it, is that for any elliptic curve (g) with rational coefficients and therefore with a conductor that will be a positive integer N, we can find two meromorhic functions $x(z), y(z)$ that parametrize the curve and that are invariant under the group Γ_N of matrices γ that are integral of determinant 1 and satisfy the conditions,

$$c \equiv 0 (\operatorname{mod} N); \quad a \equiv 1 (\operatorname{mod} N).$$

It cannot, however, be required in general that two points z and z' yield the same point on the curve only if $z' = \gamma z$ with some γ in Γ_N. This suggestion can not yet be proved but it can be easily tested, because for a given integer N there are efficient methods for finding all elliptic curves that admit parametrizations by functions invariant under Γ_N [M]. Of course, even if Taniyama's suggestion can be established, the conjecture of Birch and Swinnerton-Dyer will remain.

We have introduced L-functions as adjuncts to the study of diophantine equations, but they also appear in the study of automorphic forms, of which the modular forms of weight 1 and 2 in this section are a particular manifestation. It is in this second guise that the values of the functions can be calculated and the regularity on which we have insisted asserts that an L-function attached to a diophantine problem is always, in spite of the utterly different way it is defined, equal to one attached to an automorphic form. It is possible to pass to the general theory of automorphic forms from the theory of modular functions, but for the purposes of this essay, in which the emphasis is on the notion of automorphic L-function, it is best, in order to present it quickly and persuasively, to take our stance within representation theory.

Representations of groups over the field of real numbers

We begin with a review of the pertinent facts from the representation theory of Lie groups, taking as our first example the group $\mathrm{GL}(n, \mathbb{R})$, a group whose underlying manifold is neither compact(closed) nor connected, so that we cannot expect its irreducible representations, and it is these that are of primary interest, usually to be finite-dimensional.

There are at least two ways to analyze or to construct representations of a Lie group: by passing to the Lie algebra and searching with purely algebraic means for realizations of it; or to start from a natural action of the group on a manifold, then to pass to the associated action on functions and to decompose it irreducibly. The standard example is the orthogonal group in three variables, which acts on the sphere of radius 1 and thus on the functions on this sphere. This action commutes with the angular term in the Laplacian,

$$Df = \frac{1}{\sin\,\theta}\left(\frac{1}{\sin\,\theta} f_{\phi\phi} + (f_\theta \sin\,\theta)_\theta\right),$$

the subscripts denoting partial differentiation. Then the action of the orthogonal group on the space of spherical harmonics of order n, defined by

$$Df + n(n+1)f = 0,$$

yields, for $n = 0, 1, 2, \ldots$, the irreducible representation of degree $2n + 1$ of the orthogonal group.

One manifold that is always available is the group manifold itself, on which the group acts by left or right multiplication, and as appears most clearly in the paper of Peter-Weyl [PW], if the group is compact all representations can be obtained in the space of functions on the group itself. For example, the representation π being given on the space V, we may choose a fixed nonzero linear form u on V and sending the vector v to the function

(1) $$\langle \pi(g)v, u\rangle$$

realize V together with π on a space of functions. Since the group is compact these functions are in fact square-integrable, so that, as in [PW], standard elementary techniques from analysis can be applied. The functions (l) are known as matrix coefficients.

For groups such as $\mathrm{GL}(n,\mathbb{R})$ that are not compact similar techniques can be used, but they cease to be elementary, the system being treated being best compared to a Schrödinger equation for which both asymptotically independent and bound states appear.

Since every element of $\mathrm{GL}(n,\mathbb{R})$ is a product k_1ak_2 where k_1 and k_2 are orthogonal matrices and a is a diagonal matrix with positive eigenvalues,

$$\begin{pmatrix} \alpha_1 & & \\ & \ddots & \\ & & \alpha_n \end{pmatrix}$$

the simplest representations should have n freely assignable parameters, and are analogous to a system of n interacting but asymptotically independent particles to which arbitrary momenta can be assigned. In addition the presence of the factors k_1 and k_2 may entail the appearance of discrete quantum numbers. Moreover as a result of the symmetries implied by the possibility of permuting the α_i at the cost of modifying k_1 and k_2 we are dealing with questions of reflection more than of scattering, and as a consequence permutations of the n parameters do not change the representation.

These simple representations have simple constructions. In the group $G = \mathrm{GL}(n,\mathbb{R})$ let B be the group of superdiagonal matrices,

$$\text{(m)} \qquad b = \begin{pmatrix} \alpha_1 & * & * & \cdots & * \\ & \alpha_2 & * & & * \\ & & \alpha_3 & & * \\ & & & \ddots & \vdots \\ & & & & \alpha_n \end{pmatrix}.$$

Given n real parameters $a_1, \dots, a_n$ and n numbers $m_k = \pm 1$ (the supplementary discrete quantum numbers) we introduce the characters

$$\chi_k : \alpha \to sgn(\alpha)^{m_k} |\alpha|^{ia_k}$$

of $\mathbb{R}^\times$, as well as the character χ of B that sends the matrix b to

$$\prod_{k=1}^{n} \chi_k(\alpha_k)|\alpha_k|^{\delta-k}.$$

The number $\delta = (\mathrm{n}+1)/2$ and the supplementary exponent $\delta - k$ appear because the group manifold of $G = \mathrm{GL}(n, \mathbb{R})$ is curved. We associate to χ the representation of G on the space of functions f on G that satisfy

$$f(bg) = \chi(b) f(g) \tag{n}$$

for all b in B and all g in G. It is irreducible. Moreover the representations associated to two sequences of parameters χ_k are equivalent if and only if one sequence is a permutation of the other.

If we replace the exponents ia_k by arbitrary complex numbers, we may still study the representation on the space of functions satisfying (n). It will in general be neither unitary nor irreducible. None the less, there is a well determined process for choosing a specific irreducible factor of the representation that is then taken as the representation associated to the sequence of characters, $\chi_1, \ldots, \chi_n$. In other words, we allow complex momenta; the theory of numbers demands that we do so. Once again, the representations attached to two sequences are equivalent only if the sequences are permutations of each other.

In his basic paper [Ba], Bargmann first constructs representations algebraically by examining commutation relations, but his most important discovery was, at least if we tailor its formulation to our present purposes, that there are irreducible representations of $\mathrm{GL}(2, \mathbb{R})$ whose matrix coefficients are square-integrable over the submanifold $\mathrm{SL}(2, \mathbb{R})$. In other words analogues of two-particle bound states appear; they are known as discrete-series representations of $\mathrm{GL}(2, \mathbb{R})$. They are associated to one continuous parameter (or, if the analogy is pursued,one momentum) and one discrete parameter (or quantum number).

As long as we remain with the general linear group over the real field there are no analogues of n-particle bound states for $n > 2$. The reason is that the real numbers have a single algebraic extension, the field of complex numbers, and it is of degree two. Observe that in the context of the analogy, a one-particle state is simply a (continuous) homomorphism of $\mathrm{GL}(1, \mathbb{R})$, the multiplicative group of real numbers, into $\mathbb{C}^\times$, thus, an irreducible representation of $\mathrm{GL}(1, \mathbb{R})$, which is necessarily, the group being commutative, of dimension 1. To have a convenient terminology, it is usual to call a (continuous) homomorphism of a group into $\mathbb{C}^\times$ a character, even if its absolute value is not 1.

For the group $\mathrm{GL}(n, \mathbb{C})$ of complex matrices, the representation theory is simpler. An irreducible representation of $\mathrm{GL}(1, \mathbb{C})$ is simply a homomorphism from $\mathrm{GL}(1, \mathbb{C}) = \mathbb{C}^\times$ to $\mathbb{C}^\times$, thus a character of $\mathrm{GL}(1, \mathbb{C})$. There are no discrete series for $n > 1$ and therefore no n-particle bound states, because, as the fundamental theorem of algebra implies, the field of complex numbers has no algebraic extension.

For either field, the general irreducible representation of $\mathrm{GL}(n)$ is the analogue of r interacting but asymptotically independent bound states with $n_1, ..., n_r$ particles, the sum of the n_k being n, and the n_k being subject to the constraints appropriate to the field. To construct the representation, we introduce the group P of matrices of the form (m) except that n is to be replaced by r, each α_k is a matrix with n_k rows and columns, and the asterisks represent block matrices of the appropriate size. The space on which the represention act is once again given by functions satisfying (n), where, however, χ is no longer a character, but the tensor product of representations χ_k of $\mathrm{GL}(n_k)$, so that the function f takes values in an infinite-dimensional space. If the momenta are all real, then this representation is irreducible; otherwise it is necessary to pass again to a specific factor. As before, the order of the χ_k is irrelevant.

We conclude that, proofs that involve an elaborate analysis of the pertinent scattering problems aside, the classification of irreducible representations of the general linear group has an extreme formal simplicity. For the complex field, to specify a representation of $\mathrm{GL}(n)$ we simply specify n homomorphisms from $\mathrm{GL}(1, \mathbb{C}) = \mathbb{C}^\times$ to $\mathbb{C}^\times$. Hence the conclusion, simple to state but fraught with consequence, that the irreducible representations of $\mathrm{GL}(n, \mathbb{C})$ are parametrized by the (completely reducible) n-dimensional representations of the multiplicative group $\mathbb{C}^\times$ of nonzero complex numbers. Observe that we are comparing collections of two very different objects: on one hand, *irreducible* and in general *infinite-dimensional* representations of the group $\mathrm{GL}(n, \mathbb{C})$; on the other, *finite-dimensional* indeed n*-dimensional* but in general *reducible* representations of $\mathbb{C}^\times$

A similar statement for the real field requires a group that is not commutative but whose irreducible representations are of degree at most two, in order to accomodate the existence of two-particle bound states. The appropriate group is labeled $W_\mathbb{R}$ or $W_{\mathbb{C}/\mathbb{R}}$ and is obtained by adjoining to the group $\mathbb{C}^\times$ an element w such that $w^2 = -1$ and $wz = \bar{z}w$, for z in $\mathbb{C}^\times$. Thus it is a kind of dihedral group.

Bound states correspond to irreducible representations and their momentum to the determinant of the representation. Since the homomorphism

$$z \to z\bar{z}, \qquad w \to -1$$

yields $\mathbb{R}^\times$ as the maximal abelian quotient of $W_{\mathbb{C}/\mathbb{R}}$, the determinant is indeed a character of $\mathbb{R}^\times$. Typical representations that yield two-particle bound states are:

$$z \to \begin{pmatrix} z^m & 0 \\ 0 & \bar{z}^m \end{pmatrix}; \qquad w \to \begin{pmatrix} 0 & 1 \\ (-1)^m & 0 \end{pmatrix}.$$

The integer m must, however, be different from zero for the representation to be irreducible.

The group $W_{\mathbb{R}}$ is referred to as the Weil group of the real field $\mathbb{R}$. The Weil group of the complex numbers is just the multiplicative group $\mathbb{C}^{\times}$. For both fields, the irreducible representations of $\mathrm{GL}(n)$ are parametrized by n-dimensional representations of the Weil group.The Weil group can be attached to many of the fields appearing in number theory. Its simplicity for the real and the complex field is misleading, for it incorporates many of the deepest facts about the theory of equations known at present. In the following section we shall comment on its definition for p-adic fields, but we first add some observations on its use for the field of real numbers.

As a device for inspiring some confidence that L-functions are important mathematical objects, we have focused on the problem of calculating their values. For this purpose, the groups $\mathrm{GL}(n)$ suffice. On the other hand, some central principles in the subject were suggested by Harish-Chandra's practice of treating all reductive groups uniformly, and, despite the extra baggage they entail, the theory gains both in flexibility and power when reductive groups are admitted.

In the present context a few remarks suffice, simply to suggest the possibilities. Take the field to be $\mathbb{R}$. A typical reductive group is the special orthogonal group G in $2n+1$ variables, of which there are several forms depending on the index of the quadratic form, which can, for example, be of Euclidean or Minkowski type, so that, even for a fixed n, we are dealing with groups whose representation theory is at first glance quite distinct.

In order to parametrize the representations of G with the help of the Weil group, we attach to G its L-group ${}^{L}G$, whose general definition is given in terms of the theory of equations and of the root systems of Killing and Cartan, and which, in this specific case turns out to be the symplectic group in $2n$ variables over the *complex* number field, thus, the group of complex matrices leaving invariant an alternating form,

$$x_1y_2 - x_2y_1 + \ldots + x_{2n-1}y_{2n} - x_{2n}y_{2n-1}.$$

The principle that the irreducible representations of the real Lie group G are classified by the (continuous) homomorphisms of $W_{\mathbb{R}}$ into ${}^{L}G$ is valid, but there are some cautionary remarks. Observe first of all that a reducible n-dimensional representation of $W_{\mathbb{R}}$ is a homomorphism of $W_{\mathbb{R}}$ into $\mathrm{GL}(n,\mathbb{C})$ such that the image of $W_{\mathbb{R}}$ commutes with a subgroup of $\mathrm{GL}(n,\mathbb{C})$ that is isomorphic to $\mathbb{C}^{\times}$ and not in the center of $\mathrm{GL}(n,\mathbb{C})$. Hence the notion of irreducibility has an obvious analogue for homomorphisms of $W_{\mathbb{R}}$ into ${}^{L}G$, and homomorphisms that are irreducible in this sense correspond to representations whose matrix coefficients are square-integrable over the group manifold.

Of course, if the quadratic form chosen is Euclidean then G is compact and all its representations have this property, so that any homomorphism of $W_{\mathbb{R}}$ into ${}^LG$ that is not irreducible in the sense indicated will have to be excluded from the list of parameters. A similar but less restrictive condition must also be imposed for the groups associated to forms with other indices, but with standard notions from the theory of reductive groups it can be simply and elegantly formulated, and functions in general.

It also turns out that the classification provided by $W_{\mathbb{R}}$ and ${}^LG$ is coarse; some homomorphisms correspond to several irreducible representations of G; but this has not marred the theory. On the contrary, the analysis of the families of finitely many representations associated to given homomorphisms has revealed unexpected facets of harmonic analysis and representaton theory.

The most fecond principle suggested by the introduction of the L-group is that of $functoriality$. A homomorphism from a group H to a group G does not provide any way to transfer irreducible representations from one of these groups to the other, unless of course G is abelian, nor are homomorphisms between H and G usually related to homomorphisms between ${}^LH$ and ${}^LG$. On the other hand a homomorphism from ${}^LH$ to ${}^LG$ does, by composition,

$$W_{\mathbb{R}} \to^L H \to^L G,$$

yield a map from the set of parameters for the irreducible representations of H to those for the irreducible representations of G, and thus implicitly a way to pass from irreducible representations of H to irreducible representations of G. It has not yet been possible to define this passage directly, and it promises to be very difficult to do so. The possibility of this passage is known as the principle of functoriality in the L-group. The analogous principle for p-adic fields and for the field of rational numbers would be extremely powerful, strong enough, for example, to imply all forms of the Ramanujan conjecture and to treat the L-functions of icosahedral representations, and is certainly expected to be valid; but at present it is far beyond our reach.

Representations of groups over p-adic fields

We have already insisted on the importance of the conductor and therefore on the importance of studying equations not only modulo a prime number but modulo any integer. Because of the Chinese Remainder Theorem, it is sufficient to treat congruences modulo a prime power.

We have seen, for example, that

$$x^2 - 5 = (x-1)^2 + 2x - 6 \equiv (x-1)^2 \pmod 2.$$

For the modulus 4 we have,

$$x^2 - 5 = (x-1)(x+1) - 4 \equiv (x-1)(x+1) \pmod 4.$$

On the other hand, there is no integer m such that 8 divides $m^2 - 5$ and hence no possibility of factoring $x^2 - 5$ modulo 8.

The polynomial $x^2 - 17$ behaves in quite a different fashion.

$$x^2 - 17 = (x-1)(x+1) - 16 \equiv (x-1)(x+1) \pmod{16}.$$

We can continue,

$$x^2 - 17 = (x-9)(x+9) + 64 \equiv (x-9)(x+9) \pmod{64},$$

$$x^2 - 17 = (x-41)(x+41) + 1664 \equiv (x-41)(x+41) \pmod{128},$$

because$1664 = 13 \cdot 128$, and then

$$x^2 - 17 = (x-105)(x+105) + 11008 \equiv (x-105)(x+105) \pmod{256},$$

because $11008 = 43 \cdot 256$. In general, if

$$x^2 - 17 = x^2 - m^2 + a2^k,$$

with two integers m and a, and $k \geq 3$, then m is odd and

$$x^2 - 17 = x^2 - (m - a2^{k-1})^2 + b2^{k+1},$$

with some other integer b. Thus we can proceed indefinitely, constructing a sequence $m_1, m_2, \ldots$ such that

$$x^2 - 17 \equiv (x - m_k)(x + m_k) \pmod{2^k}.$$

The sequence m_k does not converge in the usual sense, because the difference between m_k and m_{k+1} is equal to an integer times 2^{k-1}. On the other hand, the difference between m_k and any m_l, $l \geq k$, is also divisible by 2^{k-1}, so that if our concern is with divisibility properties rather than with metric properties then m_k and m_l are close when k and l are large. To simplify theoretical discussions, we treat such sequences as real objects, calling them $\mathit{2\text{-}adic\ numbers}$.

More generally, for any prime number p a sequence $a_k = m_k/n_k$ of rational numbers such that the numerator of the differences,

$$a_k - a_l = m_k/n_k - m_l/n_l, \qquad\qquad l \geq k,$$

is divisible by higher and higher powers of p as k approaches infinity is called a p-adic number, the p-adic number that this sequence of rational numbers approaces. Of course, constant sequences $a, a, a, \ldots$ are admitted, so that every rational number is a p-adic number. Indeed in almost every respect p-adic numbers can be treated like real numbers. In particular, they can be added,subtracted, multiplied, and divided. The only difference is metric.

For the purpose of studying congruences modulo powers of a given prime p, we agree that a rational number is small if its denominator is prime to p and its numerator is divisible by a high power of p. Thus, if $p = 5$ then, in this sense, $1,000,000$ is small and $1,000,000,000$ even smaller. More precisely, we take the size of a number a whose numerator and denominator are both prime to p to be 1 and of ap^k, where k is any integer, positive, negative, or zero, to be p^{-k}.

$$|ap^k|_p = p^{-k}.$$

For example,

$$|\frac{3}{2}|_5 = 1, \quad |\frac{1}{25}2|_5 = \frac{1}{25}.$$

The p-adic distance between two rational numbers, a, b is then just $|a - b|_p$. This distance extends by passing to limits to a distance defined between any two p-adic numbers, as does the p-adic absolute value $|a|_p$, so that even the metric properties of p-adic numbers are analogous to those of real numbers, although in practice quite dissimilar.

We have stressed in our discussion of zeta functions and diophantine equations that the study of rational solutions of equations usually requires a preliminary examination, often lengthy and painstaking, of the congruential properties of the equations. The introduction of the field of p-adic numbers allows us to interpret this as the study of the equations in these fields. Thus the study of rational solutions of equations is generally preceded by a study of their p-adic solutions for every prime p.

This requires, among other things, some understanding over p-adic fields of equations of the form,

$$x^n + a_{n-1}x^{n-1} + \ldots + a_0 = 0.$$

This is the theory of equations over p-adic fields, and is considerably more difficult than the theory over the real field, because for any p-adic field there are irreducible equations of arbitrary degree, and not, as for the real field, of degree at most two. Thus the field of p-adic numbers has algebraic extensions of arbitrarily high degree and the Galois groups of these extensions are quite complicated.

As our previous discussion suggests, the extensions with abelian Galois groups are much more accessible than the others, and can be classified. Even so, this requires an elaborate theory that we cannot rehearse here. The basic results can be encapsulated in the existence of the Weil group, a way of expressing them that is concise and suggestive, especially for our present purposes, but hardly reveals their import to the uninitiated.

If we denote the p-adic field by F, then a Galois extension K of F is of course a field obtained by adjoining to F all the roots of some equation. Rather than using the Weil group W_F, we use the groups $W_{K/F}$, from which W_F is ultimately constructed. The group $W_{K/F}$ has as subgroup the multiplicative group $K^\times$ of the field K. It is such that if we divide it by this subgroup to obtain a quotient group then this quotient is isomorphic to the Galois group of the extension K. All the art is in piecing these two constituents together. Two properties worth mentioning are: (i) there is always a homomorphism from $W_{K/F}$ onto $F^\times$; (ii) if K contains L then there is a homomorphism from $W_{K/F}$ to $W_{L/F}$.

The Weil group is the correct tool for the classification of the representations of the group $\mathrm{GL}(n, F)$ of $n \times n$ matrices with entries from F. This is a topological group, although not a Lie group. Nevertheless we can study its (continuous) irreducible representations. There are now discrete-series representations for any value of n, thus n-particle bound states, because there are irreducible representations of the Weil group of any degree. Over a p-adic field the classification of the (continuous) irreducible, and of course generally infinite-dimensional representations of $\mathrm{GL}(n, F)$ by (continuous) finite-dimensional and, in general, reducible representations of the Weil group is not yet complete, although much is known [He].

To have a statement exactly like that for the real field, one uses not the Weil group itself but a thickened form, its product with the group $\mathrm{SL}(2, \mathbb{C})$ or, recalling the unitary trick of Hermann Weyl, the group $SU(2)$. In contrast to the Weil group over the real field, in which the subgroup $\mathbb{C}^\times$ can transmit geometric information, the Weil group over a p-adic field contains only information from the theory of equations, and could be defined entirely in terms of Galois groups. The classification by the thickened group implies that representation theory, like the theory of automorphic forms, to which we are using it as a steppingstone, is adequate to the description of geometric phenomena that first appear for the curves and surfaces defined by equations in more than one variable. It is, however, rather the simplest representations of the group $\mathrm{GL}(n, F)$ that we need to understand before passing to automorphic forms.

The construction used over the real and the complex field may be used here to construct the representation associated to r discrete-series representations π_k of $\mathrm{GL}(n_k, F)$ with

$$n_1 + \ldots + n_r = n.$$

This is then the counterpart of r interacting but asymptotically independent n_k-particle bound states.

The representations that are of particular importance, the counterparts of n asymptotically independent particles all of whose supplementary quantum numbers are 0 and which are therefore characterized by their momenta alone, are usually referred to as unramified representations. More precisely, a 1-particle state corresponds to an irreducible representation of $\mathrm{GL}(1, F) = F^\times$ and is thus simply a homomorphism from $F^\times$ to $\mathbb{C}^\times$. The simplest such homomorphisms are those of the form,

$$x \to |x|_p^{ia}.$$

They are characterized by the complex number a alone, which can be regarded as their momentum, and putting n of them together, we construct an unramified representation on the space of functions satisfying the relation (n).

Thus an unramified representation π is defined by n numbers $a_1, \ldots, a_n$ or, rather, by the numbers $p^{ia_1}, \ldots, p^{ia_n}$, because $|x|_p$ is always a power of p. Since it is only the set of these numbers that matters and not their order, we can prescribe them simply by giving a matrix with them as eigenvalues, for example,

$$A(\pi) = \begin{pmatrix} p^{ia_1} & & & \\ & p^{ia_2} & & \\ & & \ddots & \\ & & & p^{ia_n} \end{pmatrix}.$$

Indeed, it suffices to give the conjugacy class of $A(\pi)$.

Observing that we can attach to the matrix $A(\pi)$ the function

$$L(s, \pi) = \mathrm{Det}(I - A(\pi)/p^s)^{-1},$$

we are ready to pass to automorphic forms.

Automorphic forms

If we are given for every prime p except those lying in some finite set S an unramified representation π_p of the general linear group $\mathrm{GL}(n, \mathbb{Q}_p)$ of matrices with coefficients from the p-adic field $\mathbb{Q}_p$, then we can form the infinite product

$$L(s) = \prod_{p \notin S} L(s, \pi_p). \tag{o}$$

It converges if s is a sufficiently large real number, say $s \geq a + 1$ provided that the eigenvalues of each $A(\pi_p)$ are less than p^a in absolute value. There is, however, no reason to believe that it will even define a function having a meaning for any other values of s, and much less that the function could be calculated.

In order that this procedure lead to a function defined outside the region of convergence of the infinite product, there must be some coherence between the matrices $A(\pi_p)$, and the appropriate way to assure this is to demand that the collection of representations π_p be associated to an automorphic form.

An efficient, but abstract, way to approach the subject of automorphic forms is by the introduction of *adeles*, rather ungainly objects that nevertheless, once familiar, spare much unnecessary thought and many useless calculations. We shall use them first as a conceptual aid, adding later some remarks to show how one works with them in practice. The fields of p-adic numbers were introduced as a compact way to study congruences modulo ever higher powers of prime numbers. If we want to consider congruences modulo arbitrary integers m and at the same time monitor the size, in the usual sense, of the numbers involved, then we use the adeles. An adele is a sequence $\{a_\infty, a_2, a_3, a_5, \ldots, a_p, \ldots\}$ in which a_∞ is a real number, and, for each p, a_p is a p-adic number. The only constraint is that for all but a finite number of p the number a_p be integral, in the sense that $a = bp^k$, with $k \geq 0$ and $|b|_p = 1$. Thus if b were rational then both its numerator and its denominator would be prime to p. Observe, for it is very important to us, that if we take any rational number a then the sequence $\{a, a, a, a, a, \ldots\}$ is an adele because the denominator of a is divisible by only a finite number of primes. It is possible to add, substract, and multiply adeles, but not necessarily to divide one by another. In particular, in order that the adele $1/a$, $\{a = a_\infty, a_2, \ldots\}$, be defined it is necessary that $|a_p|_p = 1$ for all but finitely many p

We may consider matrices whose entries are adeles and form their determinants. If the determinant d of an adelic matrix has an inverse $1/d$ then, using minors in the customary way, we can form the inverse of the matrix itself. Thus the collection of all adelic matrices with invertible determinant forms a group $\mathrm{GL}(n, \mathbb{A})$, or more briefly G. Taking the entries one coordinate at a time we can construct from an element g of G a sequence of elements $g_\infty, g_2, g_3, \dots$ with g_∞ in the group $\mathrm{GL}(n, \mathbb{R})$ and with g_p in the group $\mathrm{GL}(n, \mathbb{Q}_p)$. As a consequence, from an irreducible representation π of the group G we can construct irreducible representations of the group $\mathrm{GL}(n, \mathbb{R})$ and of the groups $\mathrm{GL}(n, \mathbb{Q}_p)$, and this is the first clue to the construction of coherent sequences of π_p.

Since every rational number is an adele, every matrix γ in $\Gamma = \mathrm{GL}(n, \mathbb{Q})$ is also a matrix in G so that Γ is a subgroup of G. An irreducible representation is said to be automorphic if it can be realized on a space of functions f on the group G satisfying

$$f(\gamma g) = f(g),$$

for all $g \in G$ and all $\gamma \in \Gamma$. The operator $\pi(h)$, $h \in G$, sends f to the function $f'(g) = f(gh)$. The automorphic forms themselves are these functions f. An automorphic representation then yields a sequence π_p that is unramified for almost all p and sufficiently coherent that the function

$$L(s) = L(s, \pi) \tag{p}$$

defined by (o) is meaningful for all except a finite number of values of s. Moreover, from f we can calculate its values as precisely as we like.

The functions (p) are called automorphic L-functions and appear to be exactly what we need for diophantine purposes, although it is extremely difficult to show that the L-function associated to a problem in diophantine equations is equal to an automorphic L-function, and our conjectures have far outstripped our ability to prove them. None the less, although the theorems necessary to deal with elliptic curves are not yet available in any real generality, there is a great deal of very solid evidence available for other problems [K1], so that there are no serious doubts that we are on the right track.

Convenient though they are, the use of adeles does at first leave most operational questions unanswered, so that, for example, it is by no means obvious how we are going to use elementary Fourier analysis to calculate the values of our L-functions, or that the functions (p) are in any way related to the functions (f).

To see that they are, take $n = 1$. The group G is then just the group of invertible ideles, usually denoted $\mathbb{A}^\times$. Since it is abelian, all irreducible representations are of dimension one, simply characters of the group. What kind of coherence is entailed by the demand that the representation be automorphic?

In the present context, where the group is abelian, this is the condition that the character equal 1 on the subgroup $\mathbb{Q}^\times$ of $\mathbb{A}^\times$.

Introduce the subgroup U of elements $a_\infty, a_2, a_3, \dots$ such that $|a_p|_p = 1$ for all prime numbers p and a_∞ is positive. Every element of the group $\mathbb{A}^\times$ can be written in a unique way as the product of an element in $\mathbb{Q}^\times$ and an element of U. Indeed given any adele with an inverse, so that all its coordinates are nonzero, we first multiply it by ± 1 to obtain an adele with positive coordinate a_∞. If the other coordinates are a_p and if

$$|a_p|_p = p^{-m_p},$$

then m_p is zero for all but finitely many p, so that we can form the positive rational number

$$b = \prod p^{m_p}.$$

Since $|a_p/b|_p = 1$ for all p, the original adele is equal to $\pm b$ times an element of U. Upon a little reflection one sees that this factorization is unique.

Thus an automorphic representation becomes even simpler, just a character χ of U. The condition that it be continuous, on which we have not previously insisted, turns it into a completely elementary object. It implies that χ depends on only finitely many coordinates $u_\infty, u_{p_1}, \dots, u_{p_r}$ of an element u and that further in its dependence on each u_{p_i}, it depends only on the residue of u_{p_i} modulo some power p^{n_i} of p_i, the same for all u, but, of course, not for all χ.

The set S of primes to exclude from the product (p) is $\{p_1, \dots, p_r\}$. The representations π_p associated to the automorphic representation π defined by χ are themselves characters, but of the group $\mathbb{Q}_p^\times = \mathrm{GL}(1, \mathbb{Q}_p)$, and are easily constructed. The representation π sends an element $a = bu$, where $b \in \mathbb{Q}^\times$ and $u \in U$, to $\pi(a) = \chi(u)$. To obtain π_p we take an element a_p in $\mathbb{Q}_p$, extend it to an adele a with the coordinate a_p at p but all other coordinates equal to 1, and then set $\pi_p(a_p) = \pi(a)$.

The function $L(s, \pi)$ is a product of the factors

$$L(s, \pi_p) = (1 - A(\pi_p)/p^s)^{-1},$$

the matrix $A(\pi_p)$ becoming for $n = 1$ a number. It is equal to $\pi_p(p^{-1})$ because $\pi_p(p^{-1}) = p^a$ if, in general, $\pi_p(a_p) = |a_p|_p^a$. To calculate $\pi_p(p^{-1})$, an expression in which p is a p-adic number, we have to write the adele $1, 1, \ldots, 1, p^{-1}, 1, \ldots$ as the product of a rational number and an element u in U. The rational number is p^{-1} itself, but then u must be

$$p, p, \ldots, p, 1, p, \ldots.$$

Thus it has the coordinate 1 only in $\mathbb{Q}_p$. Choosing a character χ that does not depend on the coordinate u_∞ of u in $\mathbb{R}$, we conclude finally, on recalling the properties of χ, that $A(\pi_p)$ depends only on the residue of p modulo the numbers $p_1^{n_1}, \ldots, p_r^{n_r}$. Thus Dirichlet characters are automorphic representations of $\mathrm{GL}(1, \mathbb{A})$. Indeed, there is almost no distinction between the two notions.

Hence, for $n = 1$, our constructions lead us back to objects that we had already introduced in a much simpler way. This is not very persuasive evidence either for the necessity or even for the utility of the notion of adele. For $n = 2$, however, the constructions lead to quite different objects, in particular, to modular forms in the usual sense. This similarity between Dirichlet characters and modular forms that the use of adeles renders evident went unnoticed for decades, even apparently by Erich Hecke who did so much with both; so that, although the arithmetic of equations with abelian Galois group, begun in the eighteenth century, was well understood at the end of the first quarter of this century, and even in many respects earlier, another forty years elapsed before it was realized that the vehicle for expressing the regularities that, it was hoped, would appear for all equations was already at hand, and had been so for a very long time. For $n > 2$, there is seldom an advantage in deciphering the adelic language.

Take $n = 2$ and, for simplicity, consider an automorphic representation that is unramified at all primes p. We now introduce another group U, a group of 2×2 matrices u with adelic entries, and with further special properties to be prescribed. Suppose that

$$u = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Each of the four entries has coordinates that we denote by the appropriate subscripts. We suppose first that

$$\begin{pmatrix} a_\infty & b_\infty \\ c_\infty & d_\infty \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

For each p we suppose that

$$|a_p|_p \leq 1, \quad |b_p|_p \leq 1, \quad |c_p|_p \leq 1, \quad |d_p|_p \leq 1,$$

and that

$$|a_p d_p - b_p c_p|_p = 1.$$

Thus u_p and its inverse are both integral matrices in the p-adic sense.

If an automorphic representation is unramified then it can be realized on a space containing functions f satisfying

$$\text{(q)} \qquad f(\gamma g u) = f(g),$$

for all $g \in G$, $\gamma \in \Gamma$, and $u \in U$. On the other hand, elementary considerations a little more elaborate than those involved for $n = 1$ show that every element of the group $G = \mathrm{GL}(2, \mathbb{A})$ is a product $g = \gamma h u$, where h is a matrix in $\mathrm{GL}(2, \mathbb{R})$. Since $f(g) = f(h)$, we may think of f as a function on $\mathrm{GL}(2, \mathbb{R})$. It is by no means an arbitrary function. First of all, if δ is an integral matrix with integral inverse then $f(\delta h) = f(h)$; and secondly, the representation π_∞ on the space of functions obtained in this way is irreducible.

To make the connection with modular forms, we suppose that π_∞ belongs to the discrete series and that the associated momentum is 0, or, in straightforward terms, that $\pi_\infty(z) = 1$ if z is a scalar matrix. We can find an even positive integer $l \geq 2$ such that the function f of (q) satisfies in addition the relation,

$$\text{(r)} \qquad f(hk) = e^{-il\theta} f(h),$$

if

$$k = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

is an orthogonal matrix. If l is chosen minimal then, in addition, f satisfies the relation,

$$\text{(s)} \qquad \frac{d}{dt} f(g exp(tX_1)) - i \frac{d}{dt} f(g exp(tX_2)) = 0,$$

where

$$X_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}; \qquad X_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

If we then define the function $h(z)$ on the upper half-plane by

$$h(gi) = (ci + d)^l f(g), \quad gi = \frac{ai + b}{ci + d}, \quad g = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

then equation (s) becomes the Cauchy-Riemann equation guaranteeing that h is an analytic function of z and (q) and (r) imply that it is a modular form of weight l. Because the automorphic representation was taken to be unramified it is even a modular form with respect to the group Γ_1. In general, a conductor N appears.

As well as giving a little more solid or intuitive content to otherwise tenuous notions, these somewhat technical discussions for $n = 1$ and $n = 2$ make it clear that there is very little difference between functions on $Gl(n, \mathbb{A})$ invariant under $\mathrm{GL}(n, \mathbb{Q})$ and functions on $\mathrm{GL}(n, \mathbb{R})$ invariant under $\mathrm{GL}(n, \mathbb{Z})$, or, more precisely, a subgroup of it. Since $\mathrm{GL}(n, \mathbb{R})$ almost completely fills the vector space V of $n \times n$ real matrices, it is not astonishing that we can use Fourier analysis on this space to define and calculate the value of the L-functions (p) for all values of s, although the arguments and calculations are not completely patent. They are more transparent if one remains with $Gl(n, \mathbb{A})$.

Final reflections

The introduction of infinite-dimensional representations entailed an abrupt transition in the level of the discourse, from explicit examples to notions that were only described metaphorically, but at both levels we are dealing with a tissue of conjectures that cannot be attacked frontally.

The aesthetic tension between the immediate appeal of concrete facts and problems on the one hand, and, on the other, their function as the vehicle to express and reveal not so much universal laws as an entity of a different kind, of which these laws are the very mode of being, is perhaps more widely acknowledged in physics, where it has long been accepted that the notions needed to understand perceived reality may bear little resemblance to it, than in mathematics, where oddly enough, especially among number theorists, conceptual novelty has frequently been deprecated as a reluctance to face the concrete and a flight from it. Developments of the last half-century have matured us, as an examination of Gerd Falting's proof of the Mordell conjecture makes clear [Fa], but there is a further stage to reach.

It may be that we are hampered by the absence of a central unresolved difficulty and by the extremely large number of currently inaccessible conjectures, at whose extent we have hardly hinted. Some are thoroughly tested; others are in doubt, but they form a coherent whole. What we do in the face of them, whether we search for specific or general theorems, will be determined by our temperament or mood. For those who thrive on the interplay of the abstract and the concrete, the principle of functoriality for the field of rational numbers and other fields of numbers has been particularly successful in suggesting problems that are difficult, that deepen our understanding, and yet before which we are not completely helpless. In particular, there are methods, *the Arthur-Selberg trace formula*, and classes of diophantine problems, those defining *Shimura varieties*, that are intimately connected to representation theory and whose mastery probably has to precede any attack on the general form of the principle of functoriality. The papers [AC] and [K2] surmount longstanding obstacles to the application of the trace formula and to the analysis of the zeta-functions of Shimura varieties, and are rich sources of technique for anyone wishing to penetrate the area.

Despite its importance as a guiding principle, I have given no prominence to functoriality, alluding to it only briefly, preferring, for the sake of cogency, to place the sequence, *diophantine equation, values of L-functions, automorphic L-functions*, in stark relief. This entails an emphasis on a fixed $\mathrm{GL}(n)$, to the prejudice of other groups, whereas the primary force of the principle of functoriality is its insistence on the intimate connection between representations, above all automorphic representations, of different groups. I have also restricted myself to the simplest possible equations that could serve as illustrations. The L-functions attached to algebraic curves are closely related to algebraic integrals on these curves, so that their theory cannot be simpler than that of these integrals, whose concrete geometric implications for curves of higher genus are much less clear than its theoretical consequences. The same disadvantages are attached to the use of L-functions.

References

[AC] J. Arthur and L. Clozel, *Simple algebras, base change, and the advanced theory of the trace formula,* Annals of Math. Studies, no. 120, Princeton University Press (1989).

[Ba] V. Bargmann, *Irreducible unitary representations of the Lorentz group,* Ann. of Math. 48, 568-640 (1947)

[BSD] B. Birch and P. Swinnerton-Dyer, *Notes on elliptic curves, I* and *II,* Jour. f. d. reine u. ang. Math., Bd 212, 7-25 (1963), and Bd. 218, 79-108 (1965).

[Bu] J. P. Buhler, *Icosahedral Galois representations,* SLNin Math., v. 654 (1978).

[D] R. Dedekind, *Gesammelte mathematische Werke,* reprinted by Chelsea Publishing Co. (1969).

[Fa] G. Faltings *Endlichkeitssätze für abelsche Varietäten über Zahlkörper,* Inv. Math. v. 73, 561-584 (1983).

[F] R. Fricke, *Die elliptische Funktionen und ihre Anwendungen,* T. II, B. G. Teubner,1922.

[H] G. Harder, *Experimente in der Mathematik,* Rheinisch-Westfälische Akad. der Wiss., Vorträge, N.321 (1983)

[Ha] T.Hawkins, *The creation of the theory of group characters,* Rice University studies, v. 64 (1978)

[He] G. Henniart, *Les conjectures de Langlands locales pour* GL*(n),* in Journées arithmétiques de Metz, Astérisque, no. 94, 67-85 (1982).

[KF] F. Klein and R. Fricke, *Vorlesungen über die Theorie der elliptischen Modulfunktionen, Bd. II,* B. G. Teubner, Leipzig (1892)

[K1] R. Kottwitz, *On the λ-adic representations associated to some simple Shimura varieties,* to appear

[K2] R. Kottwitz, *Shimura varieties and λ-adic representations,* to appear

[L1] R. Langlands, *Problems in the theory of automorphic forms,* in SLNin Math., v. 170 (1970).

[L2] R. Langlands, *Base change for* $\mathrm{GL}(2)$, Annals of Math. Studies, no. 96 (1980).

[Li] G. Lizogat, *Courbes modulaires de genre 1,* Bull. de la Soc. Math. de France, Mém. 43, v. 103 (1975)

[M] J.-F. Mestre, *Courbes de Weil et courbes supersingulières,* Seminar on Number Theory 1984-1985, Exp. No. 23, Univ. Bordeaux I, Talence, 1985.

[PW] F. Peter and H. Weyl, *Die Vollständigkeit der primitiven Darstellungen einer geschlossenen kontinuierlichen Gruppe,* Math. Ann. Bd. 97, 737-755 (1927).

[Sh] G. Shimura, *Yutaka Taniyama and his time, very personal recollections,* Bull. London Math. Soc., v. 21, 186-196 (1989).

[V] J. Vélu, *Isogénies entre courbes elliptiques,* C. R. Acad. Sc. Paris, A, t. 273, 238-241 (1971).