The Mathematics of Fermat's Last Theorem

Welcome to one of the most fascinating areas of mathematics. There's a fair amount of work involved in understanding even approximately how the recent proof of this theorem was done, but if you like mathematics, you should find it very rewarding. Please let me know by email how you like these pages. I'll fix any errors, of course, and try to improve anything that is too unclear.

Introduction

If you have ever read about number theory you probably know that (the so-called) Fermat's Last Theorem has been one of the great unsolved problems of the field for three hundred and fifty years. You may also know that a solution of the problem was claimed very recently—in 1993. And, after a few tense months of trying to overcome a difficulty that was noticed in the original proof, experts in the field now believe that the problem really is solved.

In this report, we're going to present an overview of some of the mathematics that has either been developed over the years to try to solve the problem (directly or indirectly) or else which has been found to be relevant. The emphasis here will be on the "big picture" rather than technical details. (Of course, until you begin to see the big picture, many things may look like just technical details.)

We will see that this encompasses an astonishingly large part of the whole of "pure" mathematics. In some sense, this demonstrates just how "unified" as a science mathematics really is. And this fact, rather than any intrinsic utility of a solution to the problem itself, is why so many mathematicians have worked on it over the years and have treated it as such an important problem.

The statement of Fermat's Last Theorem (FLT for short) is about as simple as any mathematical proposition could be:

> The equation $x^n + y^n = z^n$ has no solution for non-zero integers x, y, and z if n is an integer greater than 2.

The Proof of FLT

How can something like FLT be proved? Since it is a statement about the non-existence of something, the proof has to be somewhat indirect. Of course, if one could actually find a solution for some set of numbers, that would disprove the theorem and solve the problem. But we want a proof that FLT is true. The "easiest" way to show that something *doesn't* exist is to show that the supposed existence would lead to a contradiction.

At the highest level, the proof is extremely simple to understand, since it follows from just 2 theorems:

Theorem A. *If there is a solution (x, y, z, n) to the Fermat equation, then the elliptic curve defined by the equation*

$$Y^2 = X(X - x^n)(X + y^n)$$

is semistable but not modular.

And

Theorem B. *All semistable elliptic curves with rational coefficients are modular.*

However, both of these theorems are very difficult themselves, and both have been proven only in the last 10 years. But given that both are now known, it follows that, in order to avoid a contradiction, there cannot be any solution to the Fermat equation.

Don't worry too much now about the terminology used in these theorems. The purpose of this report is to explain some of the terms and many related concepts - and in the process give a bird's eye view of a vast amount of mathematical terrain.

Theorem A is obviously rather special in that it applies only if the Fermat equation has a solution. (And since we now know this isn't the case, the theorem has no further use.) It was first conjectured around 1982 by Gerhard Frey, and finally proved in 1986 by Ken Ribet, with help along the way from Jean-Pierre Serre.

Theorem B is even harder still, and it is the theorem of which Andrew Wiles first claimed a proof in 1993, thus proving FLT as well. Although problems were found in Wiles' original proof, he managed to nail it down a year later, with help from Richard Taylor.

Actually, Theorem B was conjectured earlier (in a special form) by Yutaka Taniyama around 1955, and increasingly more general forms since then by Goro Shimura and Andre Weil. It is a special case of what is now known as the Taniyama-Shimura Conjecture (which dispenses with the technical semistable requirement). And the latter conjecture is a special case of much more general conjectures that are part of what is known as the Langlands Program, after Robert Langlands.

Theorem B certainly seems, to one unfamiliar with the territory, to be quite technical and abstruse. However, on closer examination, it can be seen to be both surprising and beautiful. The reason is that it concerns two apparently quite different sorts of mathematical objects - elliptic curves and modular forms. Each of these is relatively simple and has been studied intensively for over 100 years. Along the way some very surprising parallels have been observed in the theory of each (which we will discuss). And the theorem states that the parallels are in fact the result of a fundamental underlying connection between the two.

Wiles and Taylor proved Theorem B only with the semistable restriction given here, but many experts believe that much more general versions may be true. This is a very popular area of active research at present, and a number of the experts are hard at work trying to prove generalizations.

The Mathematics of FLT

We're now going to give a whirlwind tour of number theory and related mathematical fields that are relevant to FLT and the concepts that have turned out to be fundamental to its proof. There will be many terms tossed out rather casually, and unless you have done graduate work in mathematics many will probably be unfamiliar at first. Don't let that dismay you, though - we intend to provide explanations and hypertext links to begin fleshing out some of the concepts and interrelationships.

Of course, any reasonably complete understanding is attainable only by dedicated study of graduate level texts and (eventually) research papers. But, we think, it is possible to learn your way around the ideas enough to orient yourself and to see how things fit together. How far you want to go beyond that is up to you.

FLT is a statement in number theory. The earliest attempts to prove it, by founders of number theory such as Euler, Dirichlet, and Legendre, usually involved only "elementary" techniques - that is, arguments which (though often very clever and creative) can be understood by anyone who knows what is now high school algebra.

Matters suddenly took a more profound turn when Kummer realized that necessary assumptions about unique factorization of numbers into primes that hold for ordinary integers fail for the generalized integers of an algebraic number field. (An algebraic number field is a finite "extension" of the ordinary rational numbers

to include the solutions of specific polynomial equations.) To solve this problem, Kummer invented a new kind of "ideal" numbers where unique factorization still occurs. Several decades of refinement of Kummer's ideals led directly to such ideas of modern algebra as rings, and then to modern algebraic number theory as we know it.

Despite the great power and importance of Kummer's ideal theory, and the subtlety and sophistication of subsequent developments such as class field theory, attempts to prove FLT by purely algebraic methods have always fallen short.

But something else rather surprising happened. Bernard Riemann was one of the greatest mathematicians of the 19th century, perhaps best known for putting integral calculus on a rigorous footing (with the Riemann integral). But he did a lot more that's quite relevant to number theory and FLT as well.

In the 1850s Riemann investigated the properties of a certain complex function called the zeta function, which had been of interest much earlier to people like Euler and Dirichlet. The zeta function is perhaps the simplest of a class of functions defined by a series expansion named after Dirichlet. The analytic behavior of this function, in particular the location of its zeros and poles, turned out to have a profound connection with the distribution of prime numbers. Knowledge of the zeta function eventually allowed Hadamard to prove the "prime number theorem", which gives an asymptotic formula for the number of primes there are less than any given bound. A stronger and still unproven conjecture about the zeta function, the Riemann Hypothesis (which says that the only zeros of the zeta function in the strip $0 \leq Re(z) \leq 1$ lie on the line $Re(z) = 1/2$), implies much more precise information about the distribution of primes.

Over the years, other mathematicians have invented and investigated generalizations of the zeta functions and Dirichlet series which turn out to be as intimately involved with generalizations of the ordinary rational numbers as the zeta function is with the rational numbers themselves. For instance, there are zeta functions of finite algebraic extensions of the rationals, and similar functions called L-functions that express facts about the Galois group of the extension field. There are also zeta and L-functions of elliptic curves and of finite fields. There are even p-adic analogues of zeta and L functions, defined over p-adic fields.

Various analogues of zeta and L-functions are used heavily in number theory and related areas. In particular, it is possible to formulate an equivalent of the Taniyama-Shimura conjecture as the assertion that for every elliptic curve there is a modular form which has the same associated L-function. This represents a very tantalizing and deep relationship of algebraic and analytic mathematical objects.

Riemann, in a relatively brief career, fertilized a large number of mathematical fields. As if what we've already mentioned weren't enough for anyone, he also made absolutely fundamental contributions to complex analysis by his invention of the concept of Riemann surfaces. A Riemann surface is a generalizaton of the complex plane and a natural domain of definition of analytic functions. Riemann surfaces make it possible to define and study in a natural way a very interesting class of functions called elliptic functions, which were investigated by Weierstrass. These turn out to be very closely related to elliptic curves (i.e., the sort of curve involved in Theorems A and B). By looking at functions defined on a different Riemann surface from that of elliptic functions one can construct another type of functions known as modular functions. Theorem B and more general forms of the Taniyama-Shimura Conjecture can be viewed in yet another way to affirm that there is a very significant relationship between modular functions and elliptic curves. But even well before that, modular functions have been investigated for their many properties that imply quite elegant number theoretic results.

Incidentally, Riemann was *also* responsible for Riemannian geometry, i.e. the study of curves and surfaces by techniques of differential calculus. In fact, as his invention of Riemann surfaces suggests, Riemann contributed as much to geometry as to analysis. Indeed, he did a great deal to unify the two fields. Such concepts as tensor calculus and differential manifolds are a direct result of his work - and they became the essential tools of Einstein's general relativity theory.

That, then, is a very brief overview of the mathematical cast of characters which play leading roles in

the eventual resolution of Fermat's theorem. There are various directions you can take from here. Each direction will often draw on concepts and facts that lie in one or more of the other directions, so you will have to be willing to wait until you've explored them all to get the best understanding of what's going on. With that willingness to accept ideas which are only explained elsewhere, you can choose almost any path for the next step:

Next...

A note on prerequisites

There is a lot of heavy-duty math in the following pages. That's the whole point. There's no use in pretending you will get much out of the discussion unless you've had at least a couple of college-level math courses. If you've had the courses but perhaps forgotten a little, that's OK. There are reminders of the basic definitions and a glossary. An introduction to abstract algebra (groups, rings, fields) is almost essential. A course in linear algebra would be nice too. Introductory calculus will come in handy sometimes. A course in complex analysis would be a real plus, but if you haven't had it, you can get by if you take a lot of basic results on faith.

Elliptic curves and elliptic functions

Elliptic curves are relatively simple objects that helped inspire the field of algebraic geometry because of some very special properties.

Elliptic curves and modular functions

A modular function is something like an elliptic function. Both are special cases of automorphic functions, which means they are invariant under certain group operations on their domains of definition. This introduces considerations of group theory and symmetry into the study of complex functions and Riemann surfaces. There turn out to be many parallels between the theory of elliptic curves and that of modular functions, which have deep consequences for both theories.

Zeta and L-functions

These Dirichlet series and their generalizations tie together number theoretic and analytic information in deep and mysterious ways.

Galois representations

Another kind of mathematical construct which can be made for both elliptic curves and modular forms. We look at Galois groups and their representations as matrices over various rings, including the p-adic numbers.

The Proof of Fermat's Last Theorem

This is a sketch of the results of Ribet (proving Theorem A) and Wiles (proving Theorem B). Together they prove FLT. You can read this first if you just want the highest-level overview of the proof.

Elliptic Curves and Elliptic Functions.

Contents:

What is an elliptic curve?

An elliptic curve is *not* an ellipse! The reason for the name is a little more indirect. It has to do, as we shall explain shortly, with "elliptic integrals", which arise in computing the arc length of an ellipse. But this happenstance of nomenclature isn't too significant, since an elliptic curve has different, and much more interesting, properties as compared to an ellipse.

Instead, an elliptic curve is simply the locus of points in the $x-y$ plane that satisfy an algebraic equation of the form $y^2 = Ax^3 + Bx^2 + Cx + D$ (with some additional minor technical conditions). This is deliberately vague as to what sort of values x and y represent. In the most elementary case, they are real numbers, in which case the elliptic curve is easily graphed in the usual Cartesian plane. But the theory is much richer when x and y may be any complex numbers (in $\mathbb{C}$). And for arithmetic purposes, x and y may lie in some other field, such as the rational numbers $\mathbb{Q}$ or a finite field $\mathbb{F}_p$.

So an elliptic curve is an object that is easily definable with simple high school algebra. Its amazing fruitfulness as an object of investigation may well depend on this simplicity, which makes possible the study of a number of much more sophisticated mathematical objects that can be defined in terms of elliptic curves.

It is very natural to work with curves in the complex numbers, since $\mathbb{C}$ is the *algebraic closure* of the real numbers. That is, it is the smallest algebraically closed field that contains the roots of all possible polynominals with coefficients in $\mathbb{R}$. Being algebraically closed means that $\mathbb{C}$ contains the roots of all polynomials with coefficients in $\mathbb{C}$ itself. It's natural to work with a curve in an algebraically closed field, since then the curve is as "full" as possible.

The case of elliptic curves in the complex numbers is especially interesting, not only because of the algebraic completeness of $\mathbb{C}$, but also because of the rich analytic theory that exists for complex functions. In particular, the equation of an elliptic curve defines y as an "algebraic function" of x. For every algebraic function, it is possible to construct a specific surface such that the function is "single-valued" on the surface as a domain of definition. It turns out that an elliptic curve, defined as a locus of points, is also the Riemann surface associated with the algebraic function defined by the equation.

So an elliptic curve is a Riemann surface. In fact, it is of a special type: a compact Riemann surface of genus 1. And not only that, but the converse is also true: every compact Riemann surface of genus 1 is an elliptic curve. In other words, elliptic curves over the complex numbers represent exactly the "simplest" sorts of compact Riemann surfaces with non-zero genus. Topologically, the genus counts the number of "holes" in a surface. A surface with one hole is a torus.

This topological equivalence of an elliptic curve with a torus is actually given by an explicit mapping involving the Weierstrass $\wp$-function and its first derivative. This mapping is, in effect, a parameterization of the elliptic curve by points in a "fundamental parallelogram" in the complex plane.

The $\wp$-function was originally studied for its analytic properties, specifically the fact that it is doubly periodic. That is, it is periodic with respect to two distinct complex numbers ω_1 and ω_2 (where one isn't a real number multiple of the other), in contrast to an exponential or trigonometric function, which has only one fundamental period.

The periodicity of the $\wp$-function means that it assumes exactly the same values on corresponding points of opposite sides of the parallelogram which is defined by the origin and the two periods ω_1 and ω_2. Therefore, the $\wp$-function can still be well-defined even if the two pairs of opposite sides of the parallelogram are identified. But when this identification is made, the parallelogram becomes, topologically, a torus. (Imagine taking a rectangular piece of paper and taping together first one pair of opposite sides. You'd have a cylinder. If the paper were flexible enough so you could tape together the two ends of the cylinder, you'd get a torus.)

The topological space that results from identifying opposite sides of a period parallelogram is called a complex torus. The fundamental periods ω_1 and ω_2 that define the parallellogram generate a *lattice* in $\mathbb{C}$ consisting of all sums of integral multiples of ω_1 and ω_2. If $\mathcal{L}$ denotes the lattice, then $\mathcal{L} = Z_{\omega_1} \oplus Z_{\omega_2}$. The complex torus can then be described as $\mathbb{C}/\mathcal{L}$. What this all means, therefore, is that a (complex) torus is the "natural" domain of definition of the $\wp$-function, or any doubly periodic complex function.

Classically, such doubly periodic functions were called elliptic functions, since they occurred in the elliptic integrals which represent the arc length of an ellipse. Elliptic curves got their name in this indirect way.

One of the properties of the $\wp$-function is that it satisfies the equation

$$\wp'^2 = 4\wp^3 - g_2\wp - g_3$$

where $\wp'$ is the first derivative and g_2, g_3 are constants. Thus, for any z, setting $y = \wp'(z)$ and $x = \wp(z)$ shows we have an elliptic curve, and the correspondence $z \mapsto (\wp(z), \wp'(z))$ is the explicit map from the fundamental parallelogram to the elliptic curve, which itself is embedded in $\mathbb{C} \times \mathbb{C}$ (or, more accurately, in $P\mathbb{C} \times P\mathbb{C}$, where $P\mathbb{C}$ is the projective plane, i.e. the Riemann sphere, i.e. $\mathbb{C}$ with one "point at infinity" added.)

The group structure of an elliptic curve

Now the plot thickens. It is remarkable enough, if you think about it, that there is such a tidy mapping between a complex torus and an elliptic curve. Especially so, because there are many other noteworthy analytic properties of elliptic functions that were discovered in the 19th century by Weiestrass and others and which we haven't even mentioned. Many of these properties turn out to have simple interpretations in terms of the geometry of Riemann surfaces, which is quite a deep subject in its own right.

But if that were not enough, it happens that elliptic curves have purely algebraic properties which are quite remarkable too. Most importantly, one can easily define an operation on the points of an elliptic curve that turns the whole curve into an abelian group.

Though the definition of the group law is easy, it isn't especially obvious. The simplest way to see it is to go back to looking at the elliptic curve with a given defining equation over the real projective plane, i.e. the ordinary real $x - y$ plane with a point at infinity added. Since the defining equation is a cubic in x, any straight line not parallel to the y-axis (i. e. a line where x isn't constant) will intersect the curve in either 1 point or 3 points. (Since under a rotation this becomes a question about the roots of a cubic polynomial, and there are always either 1 or 3 real roots.)

The definition of the group operation then becomes "simple". If a and b are two distinct points (i.e. $x - y$ pairs) on the curve then they define a straight line which intersects the curve at two points, hence at a third. Suppose the third point on the line is (x, y). Then the result of the group operation, which we denote as $a + b$, is defined to be $(x, -y)$. The identity element of the group will be the point at infinity, designated as O, so $a + O = O + a = a$. (This works out precisely because we defined the group operation not as the third point on the line, but instead as the reflection across the x-axis of the point.)

To define $a + a$, where we have only one point of the curve involved, we use a line which is tangent at the point. (The definition of elliptic curve excludes certain pathological cases of curves which don't have tangents everywhere.) With this definition, then, it is easy, though a little tedious, to verify that points on an elliptic curve do form an abelian group under the + operation, with O as the identity element.

All of the work to define the group structure is purely algebraic, so it can be done over any field, not just the reals. In particular, it can be done over the complex numbers too. In that case, the elliptic curve is a compact Riemann surface - and the group operation makes it a complex Lie group. Such objects have been studied in a very general setting - it has been a very busy area of mathematics for over a hundred years.

The mathematical field of algebraic geometry deals with "curves" in any number of dimensions, called algebraic varieties. When these varieties also have a group operation (that is regular as a mapping of varieties), it is called an algebraic group. It turns out that there are just two kinds of algebraic groups, with very different properties. One kind is a type of algebraic variety with the technical property of being "complete", called an abelian variety, since the group operation (it turns out) must be commutative. The other kind is a linear algebraic group, which is (isomorphic to) an algebraic subgroup of a general linear group — i.e. a group of matrices. Further, the only algebraic group that is of both types is the trivial group. An elliptic curve belongs to the abelian variety type of algebraic group.

We're mentioning all this to emphasize that an elliptic curve is just a special case of a much more general class of objects that have been studied quite extensively in the general setting. But a lot of the motivation for this study comes from the remarkable properties of elliptic curves. One might hope that by finding analogous properties of the more general objects it will become possible, eventually, to prove a vast number of interesting and/or useful results, of which Fermat's Last Theorem is just one example.

In case there is any question about the mention of possible "useful" results, it should be noted that the study of elliptic curves has also led to very concrete results about factoring large numbers, which in turn has an awful lot to do with the contemporary science of cryptography.

Arithmetic on elliptic curves

We are interested in "arithmetical" questions, since the ultimate purpose here is to study diophantine equations, i.e. polynomial equations having integral coefficients, and their solutions which are integral. The Fermat equation is the prime example. In general, an elliptic curve has the form $y^2 = Ax^3 + Bx^2 + Cx + D$, but for considering arithmetical questions, it is natural to restrict our attention to the case where A, B, C, D are all rational. This assumption will usually be in effect when we are considering properties of elliptic curves involving arithmetical questions (as opposed to their more general analytic properties). If all coefficients are rational, the elliptic curve is said to be *defined over* $\mathbb{Q}$. The all-important Taniyama-Shimura conjecture concerns only elliptic curves defined over $\mathbb{Q}$.

The fact that any elliptic curve (not necessarily defined over $\mathbb{Q}$) has an abelian group structure means that we can learn a lot about it by studying various of its subgroups. For considering arithmetical (i.e. number theoretic) questions, we restrict our attention to curves defined over $\mathbb{Q}$. In that case, there are several interesting subgroups we can consider.

The first is the group of all points on the curve E which have an order that divides m for some particular

integer m. That is, m "times" such a point is the identity element. Such points are called "m-division points", and the subgroup they make up is denoted $E[m]$. The reason for the name is that any point in $E[m]$ generates a cyclic subroup of E (and $E[m]$) whose order divides m. If the order is actually m, then the points in the cyclic group generated by the point divide E into m segments.

It isn't necessarily the case that the coordinates of a point in $E[m]$ have integral or rational coordinates. However, the coordinates will be algebraic numbers (i.e., roots of an algebraic equation with coefficients in $\mathbb{Q}$). It's relatively easy to show that as an abstract group $E[m]$ is just the direct sum of two cyclic groups of order m, i.e. $\mathbb{Z}/m\mathbb{Z} \oplus \mathbb{Z}/m\mathbb{Z}$, so its order is m^2. We shall see later that its real interest lies in the fact that we can construct representations of other groups of transformations that act on $E[m]$. Such representations will consist of 2×2 matrices with integral entries, i.e. elements of $GL_2(\mathbb{Z})$.

Another interesting subgroup of E is the set of all points whose coordinates are rational. Such points are said to be *rational points*. If the curve is defined over $\mathbb{Q}$, then it is a simple fact that the set of all rational points (if there are any) is a subgroup.

Many years ago (1921), Louis Mordell proved the theorem named after him, that the group of all rational points on an elliptic curve (over $\mathbb{Q}$) is finitely generated. (There is also a conjecture due to Mordell, that the set of rational points on a algebraic curve of genus > 1 is actually finite. Note that an elliptic curve has genus 1. This conjecture was proved by Gerd Faltings in 1983.)

One of the pricipal facts of elementary group theory is that any finitely generated abelian group is the direct sum of a finite group and a finite number of infinite cyclic groups (isomorphic to the integers $\mathbb{Z}$). If we denote the group of rational points of E by $E(\mathbb{Q})$, then $E(\mathbb{Q}) = \mathbb{Z}^r \oplus E(\mathbb{Q})_t$. The number of copies of $\mathbb{Z}$ is called the rank of $E(\mathbb{Q})$, and it's a very important invariant of E.

The subgroup $E(\mathbb{Q})_t$ of torsion points of $E(\mathbb{Q})$, i.e. points of finite order, is also very interesting. A recent (and difficult) theorem by Barry Mazur says that $E(\mathbb{Q})$ must be one of only 15 possible cases, and in fact the order of an element of $E(\mathbb{Q})_t$ must be ≤ 12 (and not 11). There are still many open questions about $E(\mathbb{Q})$, such as how large the rank r can be and whether there is an effective algorithm for computing r. Then there is the famous (among mathematicians) conjecture of Birch and Swinnerton-Dyer which says that r is actually the order of the zero at $s = 1$ of $L(E, s)$, the "L-function" of E, about which function we shall say quite a bit more later. In particular, if this is true, E has infinitely many rational points if and only if $L(E, 1) = 0$. So fans of elliptic curves would like to know a whole lot more about $L(E, s)$.

The definition of $L(E, s)$ will be made based on details about a series of other groups connected with E. These arise by considering E as an elliptic curve over the finite fields $\mathbb{F}_p$. This is the same as taking the original equation and reducing the coefficients mod p. If the equation of E has rational but non-integral coefficients, we would need to assume none of their denominators are divisible by p, so we might as well assume all coefficients to be integral to begin with (since if the denomimators are prime to p they have inverses mod p). Further, the definition of an elliptic curve requires that there are no repeated roots of the polynomial in x, and this may fail to be true when reducing mod p for some primes. Such primes are said to have "bad reduction". There will be only a finite number of these for any particular curve (they will divide the discriminant), but they have to be dealt with specially.

For any prime p where E has good reduction, we can consider the elliptic curve $E(\mathbb{F}_p)$ over $\mathbb{F}_p$ Since $\mathbb{F}_p$ is finite, there are only a finite number of points on $E(\mathbb{F}_p)$, so it is a finite group. The order of this group, $\#(E(\mathbb{F}_p))$, turns out to be a very important number.

There is a general approach in number theory of trying to deal with "global" problems, such investigating the structure of $E(\mathbb{Q})$, by looking at a closely related "local" problem mod p for all primes p. This is why we are interested in $E(\mathbb{F}_p)$. In particular, if $E(\mathbb{F}_p)$ is "large" for most p, we would expect $E(\mathbb{Q})$ to be large too.

We will see that the numbers $\#(E(\mathbb{F}_p))$ are studied by relating them to coefficients of the Dirichlet

series of $L(E, s)$, the L-function of E.

Further basic concepts and results

So far, we have thought of an elliptic curve as defined by its equation. However, a specific equation is not unique for determining a locus of points. Simple substitutions such as $x' = 2x$ obviously lead to different equations with the same locus of points that satisfy the equation. Substitutions like this amount to a change of coordinate system.

It turns out that if an elliptic curve with an equation of the general form $y^2 = f(x)$, where $f(x)$ is a cubic polynomial, then by a change of coordinates, we can find a new equation for the same curve in the form

$$y^2 = x^2 + ax + b$$

and this is called the *Weierstrass normal form*.

In order for this equation to define an elliptic curve, it must have no repeated roots. Elementary algebra shows this happens if and only if the discriminant of $x^3 + ax + b$, which is $4a^3 + 27b^2$, is not zero. It is customary to define a slightly different form of this:

$$\Delta = -16(27b^2 + 4a^3)$$

We say that two elliptic curves are isomorphic if they have defining equations which are the same under some change of coordinate system. Since we can always change coordinates to put the equation in the normal form, we only need to work with that form. However, that form still isn't quite unique — there are different equations in normal form that define isomorphic elliptic curves. In other words, there are coordinate transformations that change the coefficients but preserve the normal form. Such transformations thus lead to isomorphic curves which have different discriminants.

However, it turns out that the quantity

$$j = (12^3)4a^3/(4a^3 + 27b^2) = -1728(4a)^3/\Delta$$

is invariant no matter what normal form of the equation is used. This is called the j-invariant of the elliptic curve. Two elliptic curves are isomorphic if and only if they have the same j-invariant. (The reason for the constant coefficient 1728 is that j, being dependent on the lattice periods ω_1 and ω_2, has an explicit formula in terms of them out of which 1728 falls out naturally.)

Although the discriminant of a defining polynomial isn't an invariant of an elliptic curve, it is close. It happens that there is a related quantity called the *minimal discriminant* that is invariant. If we consider all equations in normal form for the same elliptic curve, we can choose the one whose discriminant has the fewest distinct prime factors. That discriminant is the minimal discriminant.

The most important fact about the minimal discriminant is that the primes which divide it are precisely the ones at which the curve has bad reduction. In other words, except for those primes, the reduced curve is an elliptic curve over $\mathbb{F}_p$.

There is still another invariant of an elliptic curve E, called its *conductor*, and often denoted simply by N. The exact definition is rather technical, but basically the conductor is, like the minimal discriminant, a product of primes at which the curve has bad reduction. Recall that E has bad reduction when it has a singularity modulo p. The type of singularity determines the power of p that occurs in the conductor. If the singularity is a "node", corresponding to a double root of the polynomial, the curve is said to have "multiplicative reduction" and p occurs to the first power in the conductor. If the singularity is a "cusp",

corresponding to a triple root, E is said to have "additive reduction", and p occurs in the conductor with a power of 2 or more.

If the conductor of E is N, then it will turn out that N is the "level" of certain functions called modular forms (not yet defined) with which, according to the Taniyama-Shimura conjecture, E is intimately connected.

If N is square-free, then all cases of bad reduction are of the multiplicative type. An elliptic curve of this sort is called *semistable*. It is for elliptic curves of this sort that Wiles proved the Taniyama-Shimura conjecture.

We might add a few more words about the j-invariant. It is a complex number that characterizes elliptic curves up to isomorphism: two curves are isomorphic if and only if they have the same j-invariant. Not only that, but for any non-zero complex value, there actually exists an elliptic curve with a j-invariant equal to that value. So there is a 1–1 correspondence between (isomorphism classes of) elliptic curves and $\mathbb{C}^*$.

Now, we have already seen that an elliptic curve as a complex torus is essentially determined by the period lattice of the $\wp$-function that parameterizes the curve. More precisely, two tori are isomorphic if and only if their corresponding lattices are "similar", that is, if and only if one is obtained from the other by a "homothety", i.e. multiplication by a non-zero complex number.

But there is another way to characterize similar lattices. Suppose we have two lattices. Each has a $\mathbb{Z}$-basis of the form $\{\omega_1, \omega_2\}$. Applying a homothety, we can just consider the period ratios and assume the two bases are $\{1, \tau\}$, $\{1, \tau'\}$, with both τ and τ' in the upper half plane $\mathbb{H} = \{ z \mid Im(z) > 0 \}$. These define the same lattice if and only if they are related by a transformation in $SL_2(\mathbb{Z})$. This latter is essentially what is known as the modular group Γ. So there is a 1–1 correspondence of similar lattices and elements of $\mathbb{H}/\Gamma$.

In summary, there are 1–1 correspondences between each of the following

* isomorphism classes of elliptic curves
* isomorphism classes of complex tori
* similarity clases of lattices
* elements of $\mathbb{H}/\Gamma$
* points of $\mathbb{C}^*$

Returning to the j-invariant, it is the 1–1 map between isomorphism classes of elliptic curves and $\mathbb{C}^*$. But by the above it can also be viewed as a 1–1 map $j : \mathbb{H}/\Gamma \to \mathbb{C}$. j is therefore an example of what is called a *modular function*. We'll see a lot more of modular functions and the modular group. These facts, which have been known for a long time, are the first hints of the deep relationship between elliptic curves and modular functions.

Riemann Surfaces.

Contents

Introduction

In simplest terms, Riemann surfaces were invented because the square root operation is not a single-valued function. That is, the equation $y^2 = x$ does not define a single valued function (the dependent variable y) of the independent variable x. For almost all values of x (a complex number), there are exactly two possible values of y.

A multi-valued function isn't really a function at all in the mathematical sense, and can't be dealt with directly except in the most cumbersome ways. (Especially for anything more complicated than a square root.) Yet multi-valued functions arise quite often, as inverse functions for instance. The square root is the inverse of $x \mapsto x^2$. Another example is the logarithm, which is the inverse of $x \mapsto e^x$.

The resolution of this problem is obvious and ingenious at the same time. Since the problem is that there can't be two different values y corresponding to the same x, we simply increase the number of x's so that there are enough for each possible y. We need to do this consistently, so that there is a concept of when two points are near to each other, i. e. a topology. Intuitively, since x is a point in the complex numbers $\mathbb{C}$, what we do is work with a suitable numbers of copies of $\mathbb{C}$.

How many copies of $\mathbb{C}$ we need depends on the function in question. For $y^2 = x$ we need two copies. For any given x, we pick one of two possible values of y to be $\sqrt{x}$ and let that be the value of the function at x on one of the copies of $\mathbb{C}$. We let $-\sqrt{x}$ *be* the value of the function at x on the other copy of $\mathbb{C}$. Since we are dealing with a continuous function, it is well-defined what the function value should be in an entire neighborhood of x on each copy of $\mathbb{C}$.

This gives us a perfectly well-defined single-valued function on the new topological space consisting of two copies of $\mathbb{C}$, where each copy has the same topology as $\mathbb{C}$ itself. This, really, is all there is to Riemann surfaces.

There is one technical difficulty, which is what to do when there is some collapsing of multi-valuedness. With $y^2 = x$, the point $x = 0$ has only the corresponding value $y = 0$. Since there is only one value here, we don't have any clue how to define which of the two possible values to pick for the function at points in the neighborhood of 0.

The problem is that in any neighborhood of 0 it is not clear how to perform the operation of "analytic continuation". A point like this is called a *point of ramification*, and it has to be dealt with specially when defining the topology on the Riemann surface.

Elliptic Curves and Modular Functions.

External references for this section: [Cox], [Hus], [Maz], [Ser], [Sil])

Contents:

The modular group

Recall that both main steps (Theorem A and Theorem B) of the proof of Fermat's Last Theorem refer to a particular property of an elliptic curve, that of being "modular". But we said hardly anything about what that means. It turns out that there are several different ways of defining that property—and each of them has very interesting consequences.

The term "modular" comes from "modular group". The modular group is a group Γ, consisting of certain "fractional linear transformations" of the complex plane. A fractional linear transformation is merely the simplest sort of rational function of the form

$$g(z) = \frac{az + b}{cz + d}$$

where the coefficients a, b, c, d are integers and $ad - bc = 1$.

The group operation of such functions is composition. Simple substitution shows that elements of the modular group behave under composition just like the multiplication of 2-by-2 matrices, if one uses the matrix

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

to correspond to the transformation with the coefficients a, b, c, d. Although there are many matrices that could yield the same rational function, the condition $ad - bc = 1$ (i.e. the determinant of the matrix) makes the correspondence almost unique.

The group of 2-by-2 matrices with integral coefficients and determinant 1 is called $SL_2(\mathbb{Z})$. It's pretty easy to see that the map from $SL_2(\mathbb{Z})$ to the modular group is a surjective group homomorphism with kernel just $\{ I, -I \}$, so that the modular group is isomorphic to $SL_2(\mathbb{Z})/\{ I, -I \}$. (Sometimes $SL_2(\mathbb{Z})$ is taken as the modular group so it isn't necessary to be fussy about speaking of equivalence classes of matrices modulo $\{ I, -I \}$.)

The important thing about the modular group is that it acts as a group of transformations on the upper half of the complex plane, $\mathbb{H} = \left\{ z \;\middle|\; Im(z) > 0 \right\}$. That is, if T is in Γ and z is in $\mathbb{H}$, $T(z)$ is also in $\mathbb{H}$.

Like the full complex plane, $\mathbb{H}$ can be treated as a Riemann surface. One way to think of Γ is as a group of "symmetries" on the geometric object $\mathbb{H}$.

In an analogous way, translations can be viewed as symmetry operations on the full complex plane. Given any two complex numbers z_1 and z_2, that aren't multiples of each other by a real number, one can construct the "free abelian group" on two generators, which are translation by z_1 and z_2. This group has a "fundamental domain" with the property that any point in the whole plane is a transformation of a point in the fundamental domain by an element of the group. A fundamental domain for the group generated by two translations is simply a parallelogram with vertices 0, z_1, z_2, z_1+z_2.

The fundamental domain of this group of translations looks suspiciously like the period parallelogram of an elliptic function, and in fact, for any given pair of non-collinear points z_1, z_2, an elliptic function can be constructed that has the given numbers as primitive periods. The period parallelogram is then the fundamental domain of a transformation group with the property that for any T in the group, $f(Tz) = f(z)$ for any z in $\mathbb{C}$, where f is any elliptic function with the given primitive periods. In other words, f is invariant with respect to the action of the symmetry group.

Modular functions

Returning to the modular group Γ of symmetries of $\mathbb{H}$, we can define a *modular function* as any meromorphic function on $\mathbb{H}$ which is invariant under the action of Γ. In other words,

$$f((az+b)/(cz+d)) = f(z)$$

for any complex numbers a, b, c, d.

This makes modular functions closely analogous to elliptic functions, where we have just chosen a slightly different domain of definition ($\mathbb{H}$ instead of $\mathbb{C}$) and symmetry group Γ. Modular and elliptic functions are both special cases of the concept of an automorphic function, which is a meromorphic function of 1 or more complex variables defined on a particular complex manifold and invariant under a particular group of analytic transformations (symmetries) of the manifold.

In practice, we want to consider a slightly more general class of functions that are not strictly invariant under transformations in Γ. We say that a function $f(z)$ on $\mathbb{H}$ is *modular of weight k* if

$$f((az+b)/(cz+d)) = (cz+d)^k f(z)$$

for all transformations in Γ, some integer $k \geq 0$, and $z \in \mathbb{H}$. If the weight is 0, the function is modular in the strict sense that it is automorphic with respect to Γ.

Note that k must be even if f isn't identically 0, since we can take $a = d = -1$ and $b = c = 0$ to require $f(z) = (-1)^k f(z)$.

Furthermore, if $f(z)$ is modular, it is periodic of period 1, $f(z+1) = f(z)$, because the transformation $T(z) = z + 1$ is in Γ. Therefore, $f(z)$ has a Fourier expansion:

$$f(z) = \sum_{n=-m}^{\infty} c_n e^{2\pi i n z}$$

(We make the additional requirement that the lower limit of the sum is some finite number $-m$.) This is a Laurent series in $q = e^{2\pi i z}$.

If $f(z)$ is analytic for all $z \in \mathbb{H}$ then we say that it is a *modular form*, which may have non-zero weight, and is an important special case. This condition means that $f(z)$ has no "poles" (singularities) in $\mathbb{H}$. In

particular, $f(z)$ is analytic at ∞, so there are no terms in the Fourier series with negative indices, and we define $f(\infty) = c_0$. An even more special case is if $f(\infty) = c_0 = 0$, and then we say that $f(z)$ is a *cusp form*.

The condition of analyticity on a modular form is very restrictive. It turns out that the only modular forms of weight 0 are constants, although there are certainly non-trivial modular functions of weight 0 (but they have singularities). In fact, a modular form that isn't trivial must have a weight that is even and ≥ 4. A cusp form must have a weight ≥ 12.

Symmetry: groups and Riemann surfaces

We stress the terminology of "symmetry", because it is a very apt term for certain transformations of a geometric object. In elementary geometry, a symmetry is some operation on the object which leaves it "unchanged". In other words, it has to do with the concept of "sameness" in spite of difference. Given any geometric object, two distinct points z_1 and z_2 on the object can be *considered* the "same", or "equivalent", if there is some element T of a transformation group, i.e. a symmetry operation, such that $T(z_1) = z_2$. The fact that the set of transformations form a group means that this relationship is reflexive (there's an identity element), symmetric (there's a group inverse), and transitive (because of the group operation). So the relationship has the defining characteristics of what is called an *equivalence relation*.

Any time there is an equivalence relation on a set, the set can be partitioned into disjoint subsets of equivalence classes. A single equivalence class is sometimes called an "orbit", since it consists of all images of a given point under some element of the group. For instance, the set of all rotations of the plane about the origin is a group, and the orbit of any particular point in the plane is a circle whose radius is the distance of the point from the origin.

If the set which is acted upon has a topology, then the set of orbits also has a topology which is called the "quotient topology", and the resulting topological space is called the "quotient space". Since a Riemann surface has a topology, a group of analytic transformations acting upon it defines a quotient space, which is also a Riemann surface. In this way, whenever we have a class of functions on a Riemann surface that are invariant under the operation of a symmetry group, we can regard the functions as actually defined on the quotient space.

In particular, for the modular group Γ, one can consider the quotient space $\mathbb{H}/\Gamma$ of the upper half plane. This consists of the space of orbits of points lying in the fundamental domain of Γ. The fundamental domain D of Γ consists of the set of points in the strip $\left\{ z \;\middle|\; |Re(z)| \leq 1/2, |z| \geq 1 \right\}$ lying above the circle $|z| = 1$. Then every point in the upper half plane is $T(z)$ for some T in Γ and z in D.

Every element of $\mathbb{H}/\Gamma$ is an orbit, and there is one and only one point in the fundamental domain that lies in the orbit. This means there is a 1–1 correspondence of points in $\mathbb{H}/\Gamma$ (orbits) and points in D. In fact, $\mathbb{H}/\Gamma$ and D are topologically equivalent, and essentially the same as Riemann surfaces. So the automorphic functions on $\mathbb{H}$ with respect to Γ, i.e. the modular functions, are essentially the meromorphic functions on D considered as a Riemann surface by its isomorphism with $\mathbb{H}/\Gamma$.

We have stressed this idea of symmetry because of the way it relates the analytic and geometric properties of an object like a Riemann surface to the algebraic properties of a group. From long experience with symmetries of simple plane geometric figures we have a lot of intuitive "knowledge" of how to think in terms of symmetries. We know, for instance, that most geometric objects have only certain specific symmetries that go a long way to actually defining the object. For example, The possession of a finite cyclic group of order 5 as a symmetry group, but no larger group, pretty much characterizes a regular pentagon among all convex polygons. Symmetry is one of the fundamental concepts of mathematics.

In the case of a Riemann surface viewed as a geometric object, there are other constructs that say a lot about the object, and in particular, the space of meromorphic functions defined on the surface. If the

surface happens to be a quotient space with respect to a symmetry group on another surface, then the space of all its meromorphic functions corresponds to a very special class of functions on the "larger" surface: the automorphic functions.

Thus the elliptic functions are essentially the automorphic functions on the extended complex plane corresponding to the group of translations by two non-collinear values. If we look at a smaller group, consisting just of translations by one quantity w, we get a larger space of automorphic functions that also includes all rational functions of the exponentials $e^{2\pi ix/\omega}$. (Since elliptic functions actually have two distinct periods, they are also in this space.)

The modular group Γ is a rather less intuitive group of symmetries of the upper half plane than the group of translations of the plane, but it plays essentially the same role. It is an algebraic object that encodes geometric information about the half plane $\mathbb{H}$. Note that while $\mathbb{H}$ admits a symmetry of translation by a real number, it does not admit a translation by any non-real number. However, it does admit the transformation $z \mapsto -1/z$, which is inversion in the unit circle. This latter transformation has finite order 2. It turns out that the modular group has a presentation with generators $T(z) = z + 1$ and $S(z) = -1/z$ and relations $S^2 = (ST)^3 = 1$.

Modular functions are, then, the automorphic functions on the upper half plane under the action of the modular group. They correspond to the space of all meromorphic functions on the quotient space of the upper half plane under the action of Γ.

Subgroups of the modular group

Associating a group with a geometric object provides a very powerful way of studying the object, since the algebraic structure of the group has a close relation to geometric properties of the object. For instance, a regular hexagon has a cyclic group of order 6 as a symmetry group. But a cyclic group of order 6 is a "direct sum" of cyclic groups of orders 2 and 3, i.e. it is "generated" by elements of orders 2 and 3. Correspondingly, a regular hexagon has 2-fold and 3-fold rotational symmetry as well as 6-fold symmetry.

The modular group is infinite, so it has quite a bit of structure. Since it is defined to consist of matrices with integral entries, it is natural to consider arithmetic properties of entries of members of the group. It turns out that there are a number of interesting subgroups defined by congruence conditions.

There is, first of all, the *principal congruence subgroup of level* N, where N is a positive integer. This is denoted by $\Gamma(N)$. It is defined by the congruence conditions that $a \equiv d \equiv 1 (\mathrm{mod}\ N)$, and $c \equiv b \equiv 0 (\mathrm{mod}\ N)$. This just means that members of $\Gamma(N)$ are congruent to the identity matrix mod N. So it's not surprizing that this is a subgroup (i.e. it is closed under the group multiplication and inverse operations). If N is 1, $\Gamma(1)$ is Γ, since any element of Γ is congruent to the identity mod 1.

$\Gamma(N)$ is in fact a "normal" subgroup, since it is the kernel of the map of reduction mod N. So the "quotient group" $\Gamma/\Gamma(N)$ can be defined. Moreover, the index of $\Gamma(N)$ in Γ, which is the order of $\Gamma/\Gamma(N)$, is finite and equal to

$$N^{3/2} \prod_{p|N} (1 - 1/p^2)$$

if $N > 2$ (and 6 if $N = 2$).

Other subgroups of finite index in Γ are called *congruence subgroups* if they contain $\Gamma(N)$ for some N. If Γ' is such a subgroup, Γ' is said to have level N if N is the least integer with $\Gamma' \supseteq \Gamma(N)$. (Note that if M is a multiple of N the congruence conditions for $\Gamma(M)$ are stronger than for $\Gamma(N)$, so $\Gamma(N) \supseteq \Gamma(M)$.)

By relaxing the congruence conditions on $\Gamma(N)$ a little, we can get larger groups of the same level N. For instance, if we require only $a \equiv d \equiv 1 (\mathrm{mod}\ N)$ and $c \equiv 0 (\mathrm{mod}\ N)$, we get $\Gamma_1(N)$, and only $c \equiv 0 (\mathrm{mod}\ N)$

we get $\Gamma_0(N)$ (i.e., upper triangular matrices, mod N). Note that $\Gamma(N)_0 \supseteq \Gamma_1(N) \supseteq \Gamma(N)$.

In the theory of elliptic curves, we will often have to deal with subgroups of Γ rather than the full modular group. We will be working with functions that are automorphic only with respect to such subgroups, which is a weaker condition than full modularity, since fewer transformations are involved. In such cases, we shall continue to say things like $f(z)$ is a modular function or a modular form "with respect to the subgroup".

Modular curves

References: [Hus], [Sil]

We saw above that if Γ is the (full) modular group, then $\mathbb{H}/\Gamma$ is a Riemann surface that is isomorphic to the fundamental domain D of Γ. So it seems plausible that if Γ' is a subgroup, we should be able to consider $\mathbb{H}/\Gamma'$ as a Riemann surface.

If we have a subgroup of the modular group, we can construct a Riemann surface that is related to the subgroup in the same way that the quotient space $\mathbb{H}/\Gamma$ is related to Γ. For any subgoup $\Gamma' \subseteq \Gamma$, the fundamental domain D' of Γ' contains the fundamental domain D of Γ. (It's larger because Γ' is smaller than Γ, so there must be more points in the fundamental domain to allow any point of $\mathbb{H}$ to be a transform of a point in D' by an element of Γ'.) Since the quotient spaces $\mathbb{H}/\Gamma'$ and $\mathbb{H}/\Gamma$ are isomorphic as Riemann surfaces to the fundamental domains D' and D respectively, $\mathbb{H}/\Gamma'$ is in some sense larger than $\mathbb{H}/\Gamma$.

Just as the complex plane can be "compactified" by adding a "point at infinity" to give the "Riemann sphere", the space $\mathbb{H}/\Gamma$ can be compactified. The result is denoted $\mathbb{X}(\Gamma)$. The same can be done for $\mathbb{H}/\Gamma'$ if Γ' is any subgroup of Γ of finite index, and the result is $\mathbb{X}(\Gamma')$.

Furthermore, there is a natural many-to-1 mapping $\mathbb{H}/\Gamma' \to \mathbb{H}/\Gamma$, since every orbit in $\mathbb{H}/\Gamma'$ is contained in an orbit in $\mathbb{H}/\Gamma$. Technically this map is what's called a covering, since each point of $\mathbb{H}/\Gamma$ has an open neighborhood U whose pre-image is a disjoint union of open sets which are homeomorphic to U. Intuitively, this means that $\mathbb{H}/\Gamma'$ is (locally) like multiple copies of $\mathbb{H}/\Gamma$. The covering can be done for the compactified spaces $\mathbb{X}(\Gamma')$ and $\mathbb{X}(\Gamma)$ also.

So far, what we have seen is that for subgroups Γ' of finite index in Γ, the spaces $\mathbb{X}(\Gamma')$ are (compact) abstract Riemann surfaces, essentially the quotient spaces. But much more is actually true — for certain Γ', $\mathbb{X}(\Gamma')$ is in fact an algebraic curve, that is, a locus of points (x, y) in $\mathbb{C}^2$ where x and y are related by a polynomial equation $f(x, y) = 0$. (Technically, $\mathbb{X}(\Gamma')$ is what is termed a complete algebraic curve.) When Γ' is a congruence subgroup $\Gamma(N)$, the corresponding curve $\mathbb{X}(N)$ is called a modular curve. If Γ' is $\Gamma_0(N)$, the corresponding curve is written $\mathbb{X}_0(N)$.

Quite a lot of technical effort is required to verify all the necessary details to prove that these Riemann surfaces are actually algebraic curves. In general, one can explicitly construct a map $j : \mathbb{H}/\Gamma' \to \mathbb{X}(\Gamma')$ and this map has specific, significant properties. In particular, this can be done when Γ' is $\Gamma(N)$, $\Gamma_0(N)$, or $\Gamma_1(N)$.

For example, if Γ' is the full modular group Γ, $\mathbb{X}(\Gamma)$ is the Riemann sphere (i.e. the 1-dimensional complex projective line). And the map $j : \mathbb{H}/\Gamma \to \mathbb{X}(\Gamma)$ is given by $J(z)$, which was studied in the classical theory of modular functions and is called the "fundamental modular function". There is a simple explicit formula for $J(z)$.

If Γ' is $\Gamma_0(N)$ so that $\mathbb{X}(\Gamma')$ is $\mathbb{X}_0(N)$, then for some N it turns out that something rather surprising can happen. Namely, we are able to find an elliptic curve E over $\mathbb{Q}$ and a surjective map $f : \mathbb{X}_0(N) \to E$. This is called a "parameterization of the elliptic curve by modular functions". (We'll explain the terminology below.) N will be the "conductor" of E, which is (roughly) the product of primes where E has "bad reduction".

It is here that the importance of the modular curves lies, because when an elliptic curve is parameterized by modular functions in this sense, there is a modular form (of weight 2) which has an L-function (suitably defined, as we will do later) that is the same as the L-function of E (again suitably defined). If $f(z)$ is this modular form, it turns out that $f(z)dz$ is a differential 1-form, invariant under the action of Γ_0, which is the "pull-back" using the map $\mathbb{X}_0(N) \to E$ of the "fundamental" differential 1-form on E.

Furthermore, the L-function of the elliptic curve is especially nice in that it has an analytic continuation to the whole plane and satisfies a functional equation. There is a conjecture known as the Hasse-Weil conjecture which says this is true for the L-function of any elliptic curve over $\mathbb{Q}$. The Hasse-Weil conjecture is in turn part of a larger research program named after Langlands. We will go into much more more detail on L-functions later.

The property of an elliptic curve of being parameterized by modular functions is one way of defining a *modular* elliptic curve, and the Taniyama-Shimura conjecture asserts that every elliptic curve is modular. Before Wiles' recent results, only elliptic curves with the property known as "complex multiplication" had been shown to be parameterised by modular functions (by Shimura in 1971).

There's only one thing left to do here: to explain why we call a map $f : \mathbb{X}_0(N) \to E$ a "parametrization of E by modular functions". But this is simple. Since E is an elliptic curve, it consists of points (x, y) in $\mathbb{C}^2$ where x and y are related by a polynomial equation, specifically $y^2 = 4x^3 + Ax + B$. So we get two functions $f_1, f_2 : \mathbb{X}_0(N) \to \mathbb{C}$ such that

$$f_2(t)^2 = 4f_x(t)^3 + Af_1(t) + B$$

Now, except at a finite number of points, a function on $\mathbb{X}_0(N)$ can be "lifted" to a function on $\mathbb{H}$ which is invariant under the action of $\Gamma_0(N)$ — i.e. a function that is modular with respect to $\Gamma_0(N)$. So we have an explicit parameterization of the curve E by modular functions (for a certain subgroup of Γ of finite index).

Why does $\Gamma_0(N)$ play the leading role here instead of other congruence subroups of level N such as $\Gamma_1(N)$ and $\Gamma(N)$? It is because the fact we have a covering $\mathbb{X}_0(N) \to E$ is the "best" we can do for a particular N. There are also coverings $\mathbb{X}(N) \to \mathbb{X}_1(N) \to \mathbb{X}_0(N)$ because $\Gamma(N) \subseteq \Gamma_1(N) \subseteq \Gamma_0(N)$. So there are coverings of E by $\mathbb{X}(N)$ and $\mathbb{X}_1(N)$ also (just by composition), and therefore parameterizations of E by functions modular with respect to $\Gamma(N)$ and $\Gamma_1(N)$. But since those are subgroups of $\Gamma_0(N)$, the same functions that are modular for $\Gamma_0(N)$ are for the others as well.

L-functions and elliptic curves.

External references for this section: [Gou], [Rih]

Contents:

Dirichlet series, zeta functions, and *L*-functions

A *Dirichlet series* is an infinite series of the form

$$F(s) = \sum_{n=1}^{\infty} \frac{a_n}{n^s}$$

(It's traditional to denote the function's argument, an arbitrary complex number, by s.)

Such series can easily be shown to converge in a right half-plane of the form $Re(s) > t$ for some $t \in \mathbb{R}$, and in that region they represent analytic functions. For many such series of interest, the corresponding function can be extended to a meromorphic function on the whole plane by a process known as "analytic continuation".

Dirichlet functions are of great importance in number theory because of the following formal property. Suppose that $\{ a_p \}$ is a sequence of numbers that are defined for all integer primes p. Then we can formally define an infinite product, called an *Euler product*:

$$F(s) = \prod_{p} \frac{1}{1 - a_p p^{-s}}$$

and it is easy to show that if this product converges, then

$$F(s) = \sum_{n=1}^{\infty} \frac{a_n}{n^s}$$

where a_p are defined for non-prime n such that if $n = p_1^{n_1} \cdots p_k^{n_k}$, then $a_n = a_{p_1}^{n_1} \cdots a_{p_k}^{n_k}$. This is well-defined precisely because n has a unique representation as a product of primes.

Not all Dirichlet series are of this simple form. In fact, a Dirichlet series has this form if and only if its coefficients $\{ a_n \}$ have the property that $a_{mn} = a_m a_n$. An Euler product whose factors have the form $1/(1 - a_p p^{-s})$ is called a linear Euler product. The next simplest kind of Euler product ("quadratic") has factors which are the reciprocals of polynomials that are quadratic in p^{-s}, and we shall have examples of

these as well. The relationships among the coefficients of the corresponding Dirichlet series are rather more complicated, however. A Dirichlet series need not have an Euler product form at all, of course.

The simplest possible example occurs when all $a_n = 1$. This yields the famous Riemann zeta function:

$$\zeta(s) = \prod_p \frac{1}{1 - p^{-s}} = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

In view of its definition, it shouldn't be surprising that $\zeta(s)$ has some remarkable number-theoretic properties. Because of a number of results and conjectures stated by Riemann regarding $\zeta(s)$, it attracted a lot of attention, and many interesting properties have been discovered.

For instance,

$$\ln \zeta(s) = s \int_2^{\infty} \frac{\pi(x)}{x(x^s - 1)} dx$$

where $\pi(x)$ is the number of primes $\leq x$.

$$\frac{1}{\zeta(s)} = \sum_{n=1}^{\infty} \frac{\mu(n)}{n^s}$$

where $\mu(n)$ is the Mobius function.

$$\zeta(s)^2 = \sum_{n=1}^{\infty} \frac{\tau(n)}{n^s}$$

where $\tau(n)$ is the number of divisors of n.

$$\frac{\zeta(s)^2}{zeta(2s)} = \sum_{n=1}^{\infty} \frac{2^{\nu(n)}}{n^s}$$

where $\nu(n)$ is the number of different prime factors of n.

Some of the analytic properties of $\zeta(s)$ are just as interesting. In particular, $\zeta(s)$ has a meromorphic continuation to the whole complex plane and satisfies a functional equation:

$$\pi^{-s/2}\Gamma(s/2)\zeta(s) = \pi^{-(1-s)/2}\Gamma((1-s)/2)\zeta(1-s)$$

There are also various integral formulas for $\zeta(s)$, in particular

$$\zeta(s) = s \int_0^{\infty} \frac{\lceil x \rceil - x}{x^{s+1}} dx$$

where $\lceil x \rceil$ is the largest integer $\geq x$, and provided $0 < Re(s) < 1$. We mention this because of its similarity to a "Mellin transform", which will be important a little later.

The functional equation in turn implies other properties, such as the fact that $\zeta(s) = 0$ when $s = -2k$ for positive integers k. These are known as the "trivial zeros" of $\zeta(s)$. Also, $\zeta(s) \neq 0$ if $Re(s) > 1$ or if $Im(s) = 0$ and $0 < s < 1$. So all "non-trivial" zeros lie in the "critical strip" $0 \leq Re(s) \leq 1$ and are distributed symmetrically with respect to both the real axis and the line $Re(s) = 1/2$. The famous Riemann Hypothesis says that all of the non-trivial zeros actually lie on $Re(s) = 1/2$.

Since this is a part of a discussion of Fermat's Last Theorem, we can't fail to mention that $\zeta(s)$ is also related to the mysterious and intriguing "Bernoulli numbers" B_n, which are defined by the generating function

$$\frac{x}{e^x - 1} = \sum_{n=1}^{\infty} \frac{B_n x^n}{n!}$$

This implies various recurrence relations, and these further imply that all B_n are rational.

Specifically, it is elementary, and noted by Euler, that B_n are related to the values of $\zeta(s)$ at even positive integers:

$$\zeta(2n) = \frac{2n\pi^2|B_{2n}|}{(2n)!}$$

The B_n have many curious number theoretic properties. But the most intriguing fact appears if p is an odd prime number that does not divide the numerators of any of $B_2, \ldots, B_{p-3}$. Such a p is said to be a "regular prime". Kummer discovered that Bernoulli numbers play a part in the theory of "cyclotomic fields" (i. e. extensions of $\mathbb{Q}$ by an n-th root of unity). Specifically, if p is an odd prime, then p does not divide the "class number" of the extension $\mathbb{Q}(\xi)$ if and only if p is regular (where ξ is a primitive p-th root of unity). Further, and most importantly, Fermat's equation for exponent p has no nontrivial solution in integers if p is a regular prime. Unfortunately for early attempts to prove FLT, it hasn't yet been shown that there are an infinite number of regular primes, even though it is known there are an infinite number of irregular primes.

There are many more number theoretic applications of $\zeta(s)$, but one of the deepest is in the proof of the "prime number theorem", which gives an asymptotic formula for $\pi(x)$, the number of primes $\leq x$. This theorem, first proved by Hadamard, states that

$$\lim_{x\to\infty} \frac{\pi(x)\ln(x)}{x} = 1$$

i.e., $\pi(x)$ is asymptotically equal to $x/\ln(x)$. Since $\ln(x)$ grows much more slowly than x, this says that there really are a lot of primes. In fact, around a large number x the average gap between primes is roughly 2.3 times the number of digits in the decimal expansion of x.

The proof of this result depends on a fact which can be deduced from the functional equation that there are actually no zeros of $\zeta(s)$ on the line $Re(s) = 1$. It turns out that if the much stronger fact expressed in the Riemann hypothesis is true, then much better estimates of the distribution of primes can be given.

Throughout number theory and related fields, such as algebraic geometry and the theory of automorphic functions, many generalizations of the zeta function and Dirichlet series have been found useful. The L-functions of elliptic curves and modular forms we are about to discuss provide one important source of examples. In general, it has been fruitful to define L-functions in terms of certain representations of Galois groups, and we will be seeing some of this eventually.

The L-function of an elliptic curve

External references for this section: [Sil]

We are interested in the question of, in some sense, "how many" rational points there are on a given elliptic curve E over $\mathbb{Q}$. In the introduction to the discussion of elliptic curves, we mentioned that global questions can often be studied by looking at them locally, i.e. mod p for all primes p. So we should consider the question of how many points there are on the reduction of E at p, i.e. the corresponding curve over $\mathbb{F}_p$.

Since there are only p elements in $\mathbb{F}_p$, there are at most $p+1$ points on any curve over $\mathbb{F}_p$ (counting the "point at infinity"). Let A_p be the number of points of the curve actually in $\mathbb{F}_p$. Then define $a_p = p+1-A_p$, which represents, roughly, how many points are "missing". Note that A_p, and hence a_p, is defined even if E isn't an elliptic curve over $\mathbb{F}_p$ because its equation has repeated roots mod p ("bad reduction"). (But if E is an elliptic curve, A_p is the order of $E(p)$, the group of points of E in $\mathbb{F}_p$.)

For each p we define

$$L_p(E,s) = \frac{1}{1 - a_p p^{-s} + p^{1-2s}}$$

if E has good reduction at p, and

$$L_p(E,s) = \frac{1}{1 - a_p p^{-s}}$$

otherwise. Finally, the L-function of E is defined as

$$L(E,s) = \prod_p L_p(E,s)$$

It can be shown that the Euler product converges for $Re(s) > 3/2$, and furthermore that the Dirichlet series for $L(E,s)$ satisfies

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

where a_n is as defined above when n is prime.

This function is known as the Hasse-Weil L-function. As usual, we expect much more to be true about it. Specifically, the Hasse-Weil conjecture states that $L(E,s)$ has a meromorphic extension to the whole complex plane. Furthermore, there should be a functional equation like that of the Riemann zeta function. If

$$L^*(E,s) = N^{s/2}(2\pi)^{-s}\Gamma(s)L(E,s)$$

where $\Gamma(s)$ is the usual gamma function, and N is the *regulator* of E (which is, roughly, the product of the primes where there is bad reduction), then

$$L^*(E,s) = wL^*(E,2-s)$$

where $w = \pm 1$ depends on the curve E.

This conjecture could be proven directly for elliptic curves with the property known as "complex multiplication". Also, the conjecture was known if E is modular in the sense described earlier, i.e. if E has a parameterization by modular functions (which was known to be true in case E has complex multiplication). Now that the Taniyama-Shimura conjecture is known for "semistable" curves E, we know that such E have a parameterization by modular functions, so the Hasse-Weil conjecture holds for them also.

The conjecture of Birch and Swinnerton-Dyer goes even farther and says that $L(E,s)$ has a zero at $s = 1$ of order equal to the rank of E (i.e. the number of infinite cyclic group factors in the group of rational points on E).

Clearly, the Hasse-Weil L-function of E has some pretty impressive properties. And it is but a special case in the "Langlands program", which conjectures that members of a much broader class of Dirichlet series are meromorphic and have a functional equation. In a little more detail, it is possible to define a Hasse-Weil L-function for "projective varieties" over a number field. A "projective variety" is basically a higher dimensional analog of an algebraic curve, while a number field is a finite extension of $\mathbb{Q}$. A priori, such L-functions are defined only in a certain right half plane. The Langlands program involves studying these L-functions by looking at related L-functions of automorphic representations of "reductive algebraic groups". We will see the special case of this for elliptic curves when we look at Galois representations. Langlands' program is central to contemporary algebraic number theory.

The L-function of a modular form

External references for this section: [Hus]

And we aren't done with the Hasse-Weil L-functions yet. They turn out to be the same as the L-functions associated with modular forms by a completely different definition.

If $f(z)$ is a modular form, $f(z+1) = f(z)$, so there is a Fourier expansion,

$$f(z) = \sum_{n=0}^{\infty} a_n e^{2\pi i n z}$$

where $z \in \mathbb{H}$, and $a_n \in \mathbb{C}$. In other words, there is a related function f^* of $q = e^{2\pi i z}$ with $f(z) = f^*(q)$, and

$$f^*(q) = \sum_{n=0}^{\infty} a_n q^n$$

is the ordinary Laurent expansion around $q = 0$. Recall that f is said to be holomorphic at ∞ if f^* is holomorphic at 0, in which case

$$f(\infty) = f^*(0) = a_0$$

Given a Fourier expansion like that, there is a standard way to construct a related Dirichlet series, using an integral transform called the *Mellin transform*. This is defined by

$$M(f, s) = \int_0^{\infty} f(t) t^{s-1} dt$$

If we let $f_1(t) = f(it) - f(\infty)$, then we can define the L-function of f as

$$L(f, s) = \frac{(2\pi)^2 M(f_1, s)}{\Gamma(s)}$$

The classical Gamma function is defined as

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

i.e. $\Gamma(s) = M(1/e^x, s)$. From this it is easily shown by a change of variables that

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

The function $L(f, s)$ is called the L-function of the modular form f. If f has weight k, so that $f(-1/z) = z^k f(z)$, then from the foregoing, it is easy to see that the related function defined by

$$L^*(f, s) = (2\pi)^{-2} \Gamma(s) L(f, s)$$

satisfies the functional equation

$$L^*(f, k - s) = (-1)^{k/2} L^*(f, s)$$

and has poles at most at $s = 0$ and $s = 2$.

We want to relate all this to elliptic curves. There is one technical snag, however, in that as we have seen, the subgroup $\Gamma_0(N)$ rather than the full modular group Γ (where N is the conductor of E) is the largest symmetry group for which we have a useful theory as far as elliptic curves are concerned. If a function f is modular only with respect to $\Gamma_0(N)$, then we don't have the relation $f(-1/z) = z^k f(z)$ since the inversion $z \mapsto -1/z$ isn't in $\Gamma_0(N)$, so we can't derive the desired functional equation.

Fortunately, it can still be shown that if f is a modular form of weight k with respect to $\Gamma_0(N)$, if $L(f, s)$ is defined as before, and if we define

$$L^*(f, s) = N^{s/2} (2\pi)^{-s} \Gamma(s) L(f, s)$$

then $L^*(f, s)$ extends to a meromorphic function on $\mathbb{C}$ and it has the functional equation

$$L^*(f, k - s) = w(-1)^{k/2} L^*(f, s)$$

where $w = \pm 1$ depends on f. In particular, for forms of weight 2, $L^*(f, 2 - s) = -wL^*(f, s)$, which just happens to be the functional equation satisfied by the L-function of an elliptic curve.

Relating elliptic curves and modular forms through their L-functions

We can see where this is leading. We have already defined the L-function $L(E,s)$ of an elliptic curve E over $\mathbb{Q}$ having Dirichlet series

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

so we would expect that the function

$$f(z) = \sum_{n=0}^{\infty} a_n e^{2\pi i n z}$$

should be very interesting — perhaps some sort of modular function. In fact, it ought to be a modular function of weight 2 for $\Gamma_0(N)$, so that $L(f,s) = L(E,s)$ has the proper functional equation. Although we can easily write down the series, all the hard work lies in showing convergence and that the resulting function is modular with weight 2.

Actually, $f(z)$ can be defined directly from $L(E,s)$ without explicit mention of the series coefficients by means of an "inverse Mellin transform":

$$f(z) = (2\pi i)^{-1} \int_{c-i\infty}^{c+i\infty} L(E,s) z^{-s}\, ds$$

using the "Mellin inversion formula". The validity of this inversion depends on specific properties of the functions involved, but it's basically similar in nature to the inversion formula for Fourier integrals.

In fact, it is a rather difficult result established by Langlands and Deligne in 1972 that if E is an elliptic curve with a parameterization by modular functions as defined previously (i.e. there is a surjection of $\mathbb{X}_0(N)$ on E over $\mathbb{Q}$), then indeed $f(z)$ is a modular cusp form (i.e. it vanishes at 0) for $\Gamma_0(N)$ of weight 2 and $L(E,s) = L(f,s)$. Since this result demonstrates the functional equation for $L(E,s)$, it verifies the Hasse-Weil conjecture when E is a modular curve (i.e. has parameterization by modular functions).

But the Taniyama-Shimura conjecture says that all elliptic curves over $\mathbb{Q}$ are modular, so if it's true, the Hasse-Weil conjecture is also true. Wiles' proof that Taniyama-Shimura holds for the semistable case therefore definitely proves Hasse-Weil in this case as well.

The cusp form $f(z)$ can be identified even more precisely when E is a modular curve. Specifically, differential geometry says that a surface like E has a "canonical differential", i.e. a differential 1-form. Given the map, $\mathbb{X}_0(N) \to E$, this 1-form pulls back to a 1-form on $\mathbb{X}_0(N)$, which is (up to a constant multiple) $f(z)dz$. Since $f(z)$ has weight 2, the differential $f(z)dz$ is invariant under transformations of $\Gamma_0(N)$, because

$$d((az+b)/(cz+d)) = (cz+d)^{-2} dz$$

Galois representations and elliptic curves.

External references for this section: [Gou]

Contents:

Introduction and motivation

So far we have seen two essentially different ways to formulate the notion of "modularity" for an elliptic curve E over q. The first is that E is modular if it has a parameterization by modular functions, i.e. a map $\mathbb{X}_0(N) \to E$. The second is that E is modular if there is a modular form f of weight 2 for $\Gamma_0(N)$ such that there is an equality of L-functions, $L(E, s) = L(f, s)$. The Taniyama-Shimura conjecture says that both of these conditions are true for any E.

In practice, it has proven difficult to use either of these properties, either to develop a proof of the conjecture or to apply the conjecture to other problems (like Fermat's Last Theorem). The difficulty, perhaps, lies in the disparity between the essentially analytic nature of the properties and the algebraic nature of an elliptic curve and the kind of problems to which we want to apply the theory.

We seem to need some more algebraic formulation of what it means for an elliptic curve to be modular. It now appears that in fact work during the last 10 or 15 years by Wiles and others (especially Ribet, Mazur, and Serre), has provided just what we need in the form of a definition of modularity involving the theory of representations of a Galois group.

In order to approach this, we need to review several standard areas of abstract algebra.

Galois theory

Galois theory is essentially the "complete" theory of the roots of polynomial equations in one variable. That is, it presents as complete a picture as possible in the general case of the solutions of polynomial equations. The study of such equations is one of the oldest parts of mathematics, of course. The explicit formulas for solutions of a quadratic equation, and sometimes a cubic equation, are taught in secondary schools. There are also explicit formulas for quartic equations, but not for quintics and equations of higher degree. Classic geometric problems like construction with ruler and compass of regular polygons and trisection of angles can be interpreted in terms of Galois theory (and thereby classified as solvable or not).

Galois theory is hardly a new part of mathematics. It is, like most things, the work of many people, but the most important ideas and results were conceived by Evariste Galois in 1832. In modern terminology, it is formulated using the concept of an algebraic structure called a *field*. A field is a set, which may be finite or infinite, that has two distinct but closely related group structures on it. The most common examples are the rational numbers, $\mathbb{Q}$, the real numbers, $\mathbb{R}$, and the complex numbers $\mathbb{C}$. Each of these has group structures corresponding to the operations of addition and multiplication. The two operations are related in that multiplication is required to be "distributive" with respect to addition, i.e. $a(b+c) = ab+ac$. Everything else follows from the group axioms and the distributive rule.

In Galois theory, the primary object of interest is the polynomial equation in one variable,

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0$$

where the coefficients $\{ a_k \}$ are all in some specific "base" field. The goal of the theory is to say as much as possible about the roots of such equations, that is, values of x for which the equation is true. In general, the roots of the equation will not be members of the same base field as the coefficients. One may think of the roots simply as abstract objects which can be "adjoined" to the base field to provide solutions of the equation.

We can introduce symbols for roots of certain equations, e. g. $\sqrt{2}$ and i, and then express other roots in terms of those symbols. It turns out that when one adds such symbols to a field (i.e. "adjoins" them) and uses the equation they satisfy as an additional axiom, then the enlarged set also satisfies all the axioms for a field, and it is called an *extension field*.

Looking again at any polynomial equation, one finds that it can have at most n roots in any extension field, where n is the degree of the polynomial. It may have fewer distinct roots if some are repeated: $x^2 + 2x + 1 = (x + 1)^2 = 0$ has just a single root ($x = -1$).

Galois' brilliant insight was that one can know essentially "everything" there is to know about the roots of polynomial equations by considering a new object, a group, namely the group of all "reasonable" permuations of those roots. Here, "reasonable" is not a technical term, but can be explained as follows.

A permutation of the roots determines an "automorphism" of the extension field that contains those roots, that is, a map T of the extension field to itself which preserves the field structure. In partucular, $T(ab) = T(a)T(b)$ and $T(a + b) = T(a) + T(b)$. Furthermore, if a number c is in the base field (the field of the coefficients of the polynomial), then we require $T(c) = c$, i.e. T leaves the base field fixed.

Given this, not all permutations of the roots of a polynomial may be reasonable, because they don't induce an automorphism of the extension field which leaves the base field fixed. This may happen if there are polynomial relationships among the roots with coefficients in the base field.

For instance, in the polynomial

$$f(x) = (x - i)(x + i)(x - 2i)(x + 2i) = x^4 + 5x^2 + 4$$

the roots are $x_1 = i$, $x_2 = -i$, $x_3 = 2i$, $x_4 = -2i$. We have the relations $x_3 = 2x_1$ and $x_4 = 2x_2$. We can allow permutations that exchange x_1 and x_2, or x_3 and x_4, or both. But we can't allow a permutation that exchanges x_1 and x_3. Because the resulting field automorphism T would require $T(x_1) = x_3 = 2x_1 = 2T(x_3) = T(2x_3) = T(4x_1) = 4T(x_1)$

The set of "reasonable" permutations thus generates a set of automorphisms of the extension field that leaves the base field fixed. This set of automorphisms is actually a group, and it is called the *Galois group* of the extension. The Galois group is a way of encoding all available information about the relationships of the roots of polynomials with coefficients in the base field that factor completely in the extension field. So in order to study all roots of a given polynomial, it is sufficient to find an extension field that contains all of the roots and examine the Galois group.

Notice that we have managed to express one kind of mathematical problem — description of the roots of a polynomial equation — in terms of a *symmetry* group, where the symmetry in question involves permutations among the roots. Here again, symmetry operations can be used to express a concept of "similarity" or "likeness". In this case, certain roots of an equation are "like" others because they satisfy the same algebraic relations, even though they are not numerically the same. But for all algebraic purposes they are interchangeable.

For future reference, we will simply state the fundamental facts of Galois theory. We say that a (finite) field extension $E \supseteq F$ is Galois if E is the field obtained by adjoining to F all roots of some irreducible polynomial with coefficients in F. The Galois group of E over F, $Gal(E/F)$, is the group of automorphisms of E that leave F fixed (i.e., that map all elements of F to themselves). The fundamental theorem says that there is a 1–1 correspondence of intermediate fields E' such that $E \supseteq E' \supseteq F$, and subgroups H of $Gal(E/F)$, where E' is the field left fixed by H. Further, H is a normal subgroup of $Gal(E/F)$, if and only if the corresponding extension E' is Galois over F, in which case $Gal(E'/F)$ is isomorphic to the quotient group $Gal(E/F)/H$.

Group representations

We have had ample evidence that groups are very useful mathematical objects. We have seen them used to describe such diverse phenomena as geometric transformations of a shape and the roots of an algebraic equation — to say nothing of their many applications outside of pure mathematics itself.

But there is one problem in working with abstract groups, in that it is often not easy to do computations with them. Group elements hardly ever involve ordinary numbers such as people, or computers, are accustomed to computing with. They are abstract objects like geometric transformations or permutations of a set. Sometimes they are just symbols related by certain equations.

However, it was discovered long ago that it is always possible to "represent" an abstract group in terms of objects that can easily be computed with, namely matrices over a field such as $\mathbb{R}$ or $\mathbb{C}$. In fact, it can be done with matricies whose entries are members of a ring, such as the integers $\mathbb{Z}$, rather than a field. (A ring is like a field, except that nonzero elements don't necessarily have multiplicative inverses.)

The construction is very straightforward. Given an arbitrary group G and a ring R, one first constructs the "group ring" $R(G)$ which consists of finite formal "sums" of "products" of elements of R and elements of G, e. g.

$$\sum_i r_i g_i, \qquad r_i \in R,\ g_i \in G$$

The sums here are in a formal sense, unrelated to the addition operation in the ring. Addition of such sums is done in the "obvious" way, and multiplication is done using the addition and multiplication laws of the ring and the group law. For instance,

$$(r_1 g_1 + r_2 g_2)(r_3 g_3 + r_4 g_4) = (r_1 r_3)(g_1 g_3) + (r_1 r_4)(g_1 g_4) + (r_2 r_3)(g_2 g_3) + (r_2 r_4)(g_2 g_4)$$

Note that multiplication doesn't have to be commutative in either the ring or the group, though most often the ring is commutative.

This group ring is an "R module" because it also allows for multiplication by elements of R in the obvious way. In case R is a field, it is simply a "vector space", and all the usual concepts of linear algebra apply. In particular, linear maps from the group ring to itself, called endomorphisms, can be expressed as matrices with entries lying in R.

Now we just have to observe that every element g of G induces a natural endomorphism of $R(G)$ given by multiplication (on the right, in the noncommutative case). Specifically, if $g \in G$ we define

$$T_g(\sum r_i g_i) = \sum r_i g_i g$$

Endomorphisms have a natural operation, namely composition, where $(T_g \circ T_h)(x) = T_h(T_g(x)) = T_h(xg) = xgh = T_{gh}(x)$ if $x \in R(G)$. Further, since G is a group, T_g has an inverse, namely $T_{g^{-1}}$. In particular, any T_g must be 1–1, and in that case it is called an automorphism. The relationship $T_g \circ T_h = T_{gh}$ says that the group law for endomorphisms is the "same" as that of G itself, in the sense that the map $g \mapsto T_g$ is a group "homomorphism", in other words, a map that preserves the group structure.

Any map from G to a group of endomorphisms on an R-module is called a *linear group representation*, or simply a representation. When the R-module is finitely generated (as it is when G is finite) and R is a field, then we just have a finite dimensional vector space V, and we choose a basis. Given a basis, we can then express any endomorphism of V as a matrix. Hence a group representation gives us a homomorphism from G to a group of matrices. If the homomorphism is "injective" (1–1), then we can treat G as a subgroup of a matrix group. This is the case, in particular, for the representation using the group ring. (If T_g is the identity as an automorphism, g must be the identity of the group.) So computations become straightforward (though possibly tedious), for either humans or computers, since they are just matrix operations.

p-adic numbers

At this point we need to make just one more detour. In number theory it happens that it is often easier to approach problems in terms of one prime number at a time. For example, this is why we considered the points of an elliptic curve over a field F_p rather than over $\mathbb{Q}$. This is such a common occurrence that we need more powerful tools to work with. One of the most useful of such tools is the "p-adic numbers".

So pick some prime p. Whenever $n \geq m$, there is a natural map of rings $\mathbb{Z}/p^n\mathbb{Z} \to \mathbb{Z}/p^m\mathbb{Z}$ given by reduction modulo p^m. Using an abstract algebraic construction called the projective limit, one can define a "universal" object which is a ring that in some sense encompasses all the $\mathbb{Z}/p^n\mathbb{Z}$ simultaneously. This object is called the ring of "p-adic integers" and denoted by $\mathbb{Z}_p$. In particular, the ordinary integers $\mathbb{Z}$ can be viewed as a subring of $\mathbb{Z}_p$, since any integer can eventually be represented by itself in $\mathbb{Z}/p^n\mathbb{Z}$ for all n sufficiently large. The quotient field of $\mathbb{Z}_p$ is denoted by $\mathbb{Q}_p$, and $\mathbb{Q}$ may be viewed as a subfield of $\mathbb{Q}_p$. Elements of $\mathbb{Q}_p$ are called *p-adic numbers*.

There are various other ways of defining $\mathbb{Z}_p$. In particular, a topology can be defined on $\mathbb{Z}$ in which "closeness" of two elements n_1 and n_2 is measured by the power of p that divides $z_2 - z_1$. (The points are closer together the larger that power of p is.) This actually yields a metric space, and $\mathbb{Z}_p$ is just the completion of $\mathbb{Z}$ in this topology.

$\mathbb{Z}_p$ can also be defined in terms of formal power series

$$\sum_{n=0}^{\infty} a_n p^n$$

where $0 \leq a_n < p$. Such series actually converge in the p-adic topology. Any ordinary integer can be uniquely represented as a polynomial in powers of p with non-negative integral coefficients $< p$, so again $\mathbb{Z}$ is naturally included in $\mathbb{Z}_p$.

Use of p-adic numbers is ubiquitous in modern number theory. There is a well-developed theory of p-adic analysis which is analogous to classical analysis on the topologically complete field $\mathbb{C}$. (But note that $\mathbb{Q}_p$ isn't algebraically complete like $\mathbb{C}$ is.) There are even p-adic analogues of zeta and L-functions.

Galois representations and elliptic curves

We have apparently strayed quite far from the topic of elliptic curves, to say nothing of Fermat's Last Theorem. What's the connection? It is that given an elliptic curve, we can define in a fairly straightforward

way, a family of representations, one for all but a finite number of primes, of an important "universal" Galois group on a group of 2-by-2 matrices over the p-adic numbers $\mathbb{Q}_p$.

Although linear group representations can be constructed by means of the group ring, as above, they can also arise naturally in many other ways. They need not be injective, either. In the abstract, a representation of a group G is just a group homomorphism $\rho : G \to GL_n(R)$, for some $n > 0$ and some ring R. ($GL_n(R)$ is the group of invertible n-by-n matrices with entries in R.)

The group representation we are about to look at encodes a lot of information about the particular elliptic curve E on which it is based. In this case, Its purpose is as a tool for studying E rather than for the information it carries about the group.

The group in question is the Galois group of an infinite field extension. We start with the rational numbers $\mathbb{Q}$. There is a field, called the algebraic closure of $\mathbb{Q}$, $\overline{\mathbb{Q}}$, which is the smallest subfield of $\mathbb{C}$ that contains all finite extensions of $\mathbb{Q}$. Essentially, $\overline{\mathbb{Q}}$ is the field generated by all algebraic numbers, i.e. all roots of polynomial equations whose coefficients lie in a finite extension of $\mathbb{Q}$. It is quite a much larger field than $\mathbb{Q}$, though it is a small subfield of $\mathbb{C}$. With a little bit of work, one can define the Galois group of the extension $\overline{\mathbb{Q}}/\mathbb{Q}$. Needless to say, it is not a finite group. From now on, $\mathcal{G}$ will denote this group.

Let E be a particular elliptic curve. Recall from the introductory material on elliptic curves, that there are groups $E[m]$ of "m-division points" of E, i.e. the subgroup of points of E that have orders dividing m. Furthermore, $E[m]$ is isomorphic to $(\mathbb{Z}/m\mathbb{Z})^2$. The coordinates of points in $E[m]$ are actually algebraic numbers, so if $g \in \mathcal{G}$, g "acts" on points of $E[m]$. A priori the result of this action is just another point on the elliptic curve, but it isn't hard to show it is actually in $E[m]$ too. In fact, this action of $\mathcal{G}$ respects the group structure of $E[m]$. Hence g corresponds to an endomorphism of the $\mathbb{Z}/m\mathbb{Z}$-module $(Z/mZ)^2$. It's not hard to check this means we have a representation of $\mathcal{G}$ in $GL_2(\mathbb{Z}/m\mathbb{Z})$, one for each m.

If p is any prime, an important special case is $m = p^n$ for any positive integer n. We get representations of $\mathcal{G}$ in $GL_2(\mathbb{Z}/p^n\mathbb{Z})$ for all n. Using the same kind of abstract algebra as was used to construct $\mathbb{Z}_p$ itself, we can piece together the representations of $\mathcal{G}$ for each n, and the result is a single representation of $\mathcal{G}$ on $GL_2(\mathbb{Q}_p)$. This representation incorporates a great deal of information about the elliptic curve E.

What kind of information in particular? Let N be the conductor of E. Consider first the representation $\rho(E, m) : \mathcal{G} \to GL_2(\mathbb{Z}/m\mathbb{Z})$ for arbitrary m. Then for any prime number q that is prime to mN there is a simple expression for the value of $a_q = q + 1 - \#(E(\mathbb{Z}/q\mathbb{Z}))$, the q-th coefficient of the Dirichlet series of the L-function $L(E, s)$, modulo m. The p-adic representation $\rho(E, p^\infty) : \mathcal{G} \to GL_2(\mathbb{Q}_p)$ is even better. For any prime q that doesn't divide pN, we get a congruence for a_q modulo p^n for any n. Since a_q is actually an integer, that is enough to determine it exactly. The best part of this is that from just this one representation at a single prime, we can recover the actual value of a_q for almost all q.

In more detail, here's how this works for a given m. The kernel of the representation $\rho(E, m)$ (all elements that map to the identity) is an infinite subgroup of finite index in $\mathcal{G}$, so by Galois theory there is a finite extension K_m of $\mathbb{Q}$ in $\overline{\mathbb{Q}}$, and $\mathcal{G}$ modulo the kernel is isomorphic to the (finite) Galois group $\mathcal{G}_m = Gal(K_m/\mathbb{Q})$ of K_m over $\mathbb{Q}$ as well as to a subgroup of $GL_2(\mathbb{Z}/m\mathbb{Z})$. ($K_m$ is just the field generated over $\mathbb{Q}$ by adjoining the coordinates of all points in $E[m]$.) So we can regard $\mathcal{G}_m$ as a subgroup of $GL_2(\mathbb{Z}/m\mathbb{Z})$. In fact, Serre has shown that (except for the rare case where E has the property known as "complex multiplication") $\mathcal{G}_p = GL_2(\mathbb{Z}/p\mathbb{Z})$ for almost all p.

In algebraic number theory (specifically, in "class field theory"), a special element σ_p of $\mathcal{G}_m$ can be identified called the "Frobenius automorphism". This element σ_p is actually well-defined only as a member of a certain (conjugacy) class. However, viewing σ_p now as an element of $GL_2(\mathbb{Z}/m\mathbb{Z})$, its trace is well-defined and, miraculously, this value is none other (modulo m) than the coefficient $a_p = p + 1 - \#(E(\mathbb{Z}/p\mathbb{Z}))$ that occurs in the L-function of E, for all primes p that do not divide m or the conductor of E.

Galois representations and modular forms

Following the line of thinking we pursued with L-functions, it seems that there ought to be some way to define a Galois representation corresponding to a modular form. We would expect, further, that whenever a modular form is related to an elliptic curve (either because it has the same L-function or because it arises from a covering $\mathbb{X}_0(N) \to E$), the Galois representation corresponding to f should be the same as the one corresponding to E. Also, this notion of modularity should be equivalent to the others. And, lastly, every elliptic curve should be modular in this sense, if the Taniyama-Shimura conjecture is true.

Unfortunately, this part of the theory seems to involve the largest number of technicalities, so that it is especially difficult to explain the constructions and the reasoning involved. On the other hand, this is the form of the theory where it has actually been possible to prove every elliptic curve is modular (in the semistable case) and apply the result to problems like Fermat's Last Theorem.

So we're going to cop out now on explaining the gory details, and instead move on to getting an overview of the theory in action, as applied to FLT.

The Proof of Fermat's Last Theorem.

External references for this section: [Gou], [Rib], [Rih]

Contents:

Frey curves

Suppose there *were* a nontrivial solution of the Fermat equation for some number n, i.e. nonzero integers a, b, c, n such that

$$a^n + b^n = c^n$$

Then we recall that around 1982 Frey called attention to the elliptic curve

$$y^2 = x(x - a^n)(x + b^n)$$

Call this curve E. Frey noted it had some very unusual properties, and guessed it might be so unusual it could not actually exist.

To begin with, various routine calculations enable us to make some useful simplifying assumptions, without loss of generality. For instance, n may be supposed to be prime and ≥ 5. b can be assumed to be even, $a \equiv 3 (\mathrm{mod}\ 4)$, and $c \equiv 1 (\mathrm{mod}\ 4)$. a, b, and c can be assumed relatively prime.

The "minimal discriminant" of E, can be computed to be $(abc)^{2n}/2^8$ — a power of 2 times a perfect prime power. One unusual thing about E is how large the discriminant is.

The conductor is a product of primes at which E has bad reduction, which is the same as the set of primes that divide the minimal discriminant. However, the exact power of each prime occurring in the conductor depends on what type of singularity the curve possesses modulo the primes of bad reduction. The definition of the conductor provides that p divides the conductor only to the first power if $x(x - a^n)(x + b^n)$ has only a double root rather than a triple root mod p. Now, any prime can divide only a or b but not both, since otherwise it would also divide c, and we have assumed a, b, and c are relatively prime. Hence the the polynomial will have the form $x^2(x + d) (\mathrm{mod}\ p)$, where $(p, d) = 1$. Hence there is only at most a double root modulo any prime, and therefore the conductor is square-free. In other words, E is semistable.

There are other odd things about E, which have to do with specific properties of its Galois representations. Because of these, Ribet's results allow us to conclude that E cannot be modular.

Proof of Fermat's Last Theorem from the Taniyama-Shimura conjecture

After Frey drew attention to the unusual elliptic curve which would result if there were actually a nontrivial solution to the Fermat equation, Jean-Pierre Serre (who has made many contributions to modern number theory and algebraic geometry) formulated various conjectures which, sometimes alone and sometimes together with the Taniyama-Shimura conjecture, could be used to prove Fermat's Last Theorem.

Kenneth Ribet quickly found a way to prove one of these conjectures. The conjecture itself doesn't really talk about either Frey curves or FLT. Instead, it simply states that if the Galois representation associated with an elliptic curve E has certain properties, then E cannot be modular. Specifically, it cannot be modular in the sense that there exists a modular form which gives rise to the same Galois representation.

We need to introduce a little additional notation and terminology to explain this more precisely. Let $S(N)$ be the (vector) space of cusp forms for $\Gamma_0(N)$ of weight 2. "Classical" theory of modular forms shows that $S(N)$ can be identified with the space of "holomorphic differentials" on the Riemann surface $\mathbb{X}_0(N)$. Furthermore, the dimension of $S(N)$ is finite and equal to the "genus" of $\mathbb{X}_0(N)$. "Genus" is a standard topological property of surfaces, which is intuitively the number of holes in the surface. (E. g. a torus, such as an elliptic curve, has genus 1.)

But there are relatively simple explicit formulas for the genus of $\mathbb{X}_0(N)$. These formulas, developed long ago by Hurwitz in the theory of Riemann surfaces, involve the index of $\Gamma_0(N)$ in $\mathcal{G}$. A fact of crucial importance is that for $N < 11$, the genus of $\mathbb{X}_0(N)$, and hence the dimension of $S(N)$, is zero. In other words, $S(N)$ contains only the constant form 0 in that case. We shall use this fact about $S(2)$ very soon.

There are certain operators called Hecke operators, after Erich Hecke, on spaces of modular forms, and for the subspace $S(N)$ in particular, since they preserve the weight of a form. Hecke operators can be defined concretely in various ways. There is a Hecke operator $T(n)$ for all $n \geq 1$. There are formulas that relate $T(n)$ for composite n to $T(p)$ where is a prime dividing n, so $T(p)$ for prime p determine all $T(n)$.

All $T(n)$ are linear operators on $S(N)$. If there is an f in $S(N)$ that is a simultaneous eigenvector of all $T(n)$, i.e. $T(n)(f) = \lambda(n)f$, where $\lambda(n) \in \mathbb{C}$, f is called an *eigenform*. (Nontrivial eigenforms need not exist, e. g. if $S(N)$ has dimension 0.) f is said to be normalized if its leading Fourier series coefficient is 1. In that case, the eigenvalues $\lambda(n)$ turn out to be the Fourier series coefficients in the expansion

$$f(z) = \sum_{n=0}^{\infty} a_n e^{2\pi i n z}$$

It can be shown that if $f(z)$ is a cusp form which is a normalized eigenfunction for all $T(p)$, then there is an Euler product decomposition for the L-function $L(f, s)$. This is obviously of great technical usefulness in relating L-functions of forms and those of elliptic curves (which are Euler products by definition).

If $f \in S(N)$ is a normalized eigenform of all Hecke operators, it can in fact be shown that the coefficients in the Fourier expansion are all algebraic numbers and that they generate a *finite* extension K of $\mathbb{Q}$.

Prime ideals of the ring of integers of K are the analogues of prime numbers of $\mathbb{Q}$. In the case that f is a normalized eigenform it is possible to carry out the construction of a Galois representation $\rho(f, \mathcal{P})$ of $Gal(\overline{\mathbb{Q}}/\mathbb{Q})$ for any prime ideal $\mathcal{P}$ of the ring of integers of K.

At last we can describe what Ribet proved. Suppose E is a semistable elliptic curve with conductor N and that its associated Galois representation $\rho(E, p)$ for some prime p has certain properties. Suppose 2 divides N (which is true for Frey curves). If E is modular, then there is a normalized eigenform f and a prime ideal $\mathcal{P}$ lying over p (i.e. one of the prime factors of p in the extension field generated by the Fourier coefficients of f) such that the Galois representation $\rho(f, \mathcal{P})$ is $\rho(E, p)$. Ribet showed that it is possible to find an odd prime $q \neq p$ which divides N such that there is another $f' \in S(N/q)$, and a corresponding prime ideal $\mathcal{P}'$ of the ring of integers in the field generated by the coefficients of f' such that $\rho(f', \mathcal{P}')$ gives essentially the same Galois representation. This is known as the "level lowering" conjecture since it asserts that under the right conditions there is an eigenform of a lower level that gives essentially the same representation.

But this process can be repeated as long as N has any odd prime factors. It is important that the curve E is semistable so that N is square-free. This means all that all odd prime factors of N can been eliminated, so there must be a nontrivial eigenform of level 2, i.e. in $S(2)$, that gives essentially the same

Galois representation. And that is a contradiction, since $S(2)$ has dimension 0, hence contains no non-trivial forms. The contradiction means that E can't be modular.

Now we invoke the "unusual" properties of the Frey curve resulting from a solution of FLT. These properties allow it to be shown that the associated Galois representation has the properties required to apply Ribet's result. Hence the Frey curve can't be modular.

But the Frey curve is semistable, so the semistable case of the Taniyama-Shimura conjecture, which Wiles proved, implies the curve *is* modular. This contradiction means that the assumption of the existence of a nontrivial solution of the Fermat equation must be wrong, and so FLT is proved.

Proof of the semistable case of the Taniyama-Shimura conjecture

Not very surprisingly (since it was such hard work), the proof is quite technical. However, the outline of it is relatively simple. In the following, we assume that E is a semistable elliptic curve with conductor N. We have to prove E is modular.

We know we can construct a Galois representation $\rho(E, p^\infty) : \mathcal{G} \to GL_2(\mathbb{Z}_p)$ for any prime p. To show that E is modular, we have to show this representation is modular in a suitable sense. The wonderful thing is, this needs to be done for only one prime p, and we can "shop around" for whatever prime is easiest to work with.

To show $\rho(E, p^\infty)$ is modular involves finding a normalized eigenform f in $S(N)$ with appropiate properties. The properties required are that the eigenvalues of f, which are its Fourier series coefficients, should be congruent mod q to trace$(\rho(E, p^\infty)(\sigma_q))$ for all but a finite number of prime q. ($\sigma_q \in \mathcal{G}$ is the "Frobenius element".) We know that the trace is, for q prime to pN, the coefficient $a_q = q + 1 - \#(E(F_q))$ of the Dirichlet series of $L(E, s)$.

The longest and hardest part of Wiles' work was to prove a general result which is roughly that if $\rho(E, p)$ is modular then so is $\rho(E, p^\infty)$ In other words, to show that E is modular, it is actually sufficient just to show that $\rho(E, p) : \mathcal{G} \to GL_2(\mathbb{Z}/p\mathbb{Z})$ is modular. This is called the "modular lifting problem".

The problem boils down to assuming that $\rho(E, p)$ is modular and attempting to "lift" the representation to $\rho(E, p^\infty)$. This is done mainly by working with the theory of representations as much as possible, without specific reference to the curve E. The proof uses a concept called "deformation", which suggests intuitively what goes on in the process of lifting.

The outcome of this part of Wiles work is:

Theorem. *Suppose that E is a semistable elliptic curve over $\mathbb{Q}$. Let p be an odd prime. Assume that the representation $\rho(E, p)$ is both irreducible and modular. Then E is a modular elliptic curve.*

At this point, all we have to do is find a single prime p such that $\rho(E, p)$ is irreducible and modular. But Langlands and Tunnell had already proven in 1980–81 that $\rho(E, 3)$ is modular.

Unfortunately, this isn't quite enough. If $\rho(E, 3)$ is irreducible, we are done. But otherwise, one more step is required. So suppose $\rho(E, 3)$ is reducible. Wiles then considered $\rho(E, 5)$. That may be either reducible or irreducible as well. If it is reducible, Wiles proved directly that E is modular.

So the last case is if $\rho(E, 5)$ is irreducible. Wiles showed that there is another semistable curve E' such that $\rho(E', 3)$ is irreducible, and hence E' is modular by the above theorem. But Wiles could also arrange that the representations $\rho(E', 5)$ and $\rho(E, 5)$ are isomorphic. Hence $\rho(E, 5)$ is irreducible and modular, so E is modular by the theorem.

Glossary of terms for Fermat's Last Theorem.

Definitions for terms that are in boldface may be found elsewhere in the glossary.

Note that some of these "definitions" are more or less in their exact technical form, while others (due to the complexity of the concept) are only intuitive descriptions.

ABELIAN GROUP
: A **group** in which the operation is commutative.

ABELIAN VARIETY
: An **algebraic group** that is also a **complete algebraic variety**. The group is necesarily an **abelian group**.

ALGEBRAIC CLOSURE
: The smallest field that contains the roots of all polynomial equations of one variable having coefficients lying in the field.

ALGEBRAIC CURVE
: An **algebraic variety** of dimension one. More informally, it is the locus of points that satisfy a polynomial equation in two variables.

ALGEBRAIC FUNCTION
: A function whose dependent variable satisfies a polynomial relationship with one or more independent variables.

ALGEBRAIC GEOMETRY
: The study of the geometric properties of the locus of points in 2 or more dimensions that satisfy sets of polynomial equations.

ALGEBRAIC GROUP
: An **algebraic variety** that has a group structure where the multiplication and inversion mappings are **morphisms** of algebraic varieties. There are two distinct types of algebraic groups: **abelian varieties** and **linear algebraic groups**.

ALGEBRAIC NUMBER
: Any solution of a polynomial equation of one variable whose coeffients are in a specific **field** (usually the rational numbers $\mathbb{Q}$).

ALGEBRAIC VARIETY
: One of the principal objects studied in **algebraic geometry**. A generalization to higher dimensions of an **algebraic curve**. It is, essentially, the locus of points that simultaneously satisfy m polynomial equations in n variables, with $m < n$.

ALGEBRAICALLY CLOSED FIELD
: A **field** that contains all solutions of 1-variable polynomial equations with coefficients in the field.

ANALYTIC CONTINUATION
: The process of extending an **analytic function** defined on some domain (of the complex numbers) to a larger domain.

ANALYTIC FUNCTION
: A complex-valued function of a complex variable that is infinitely differentiable. Equivalently, it is a complex function that can be represented in the **neighborhood** of some point by a power series about the point.

AUTOMORPHIC FUNCTION
: A **meromorphic** function of a complex variable that is invariant under a **group** of transformations of the function's domain. i.e., f is automorphic if $f(Tz) = f(z)$ for every transformation T in a given group.

BIRCH AND SWINNERTON-DYER CONJECTURE
: The conjecture that the **rank** of the **group** of rational points of an **elliptic curve** E is equal to the order of the zero of the **L-function** $L(E, s)$ of the curve at $s = 1$.

CLOSED SET
: The set complement of an **open set** in a **topological space**. i.e., a set of all points of the space not in some open set.

COMPLETE ALGEBRAIC VARIETY
: An **algebraic variety** with the property that for any other variety Y, the projection $X \times Y \to Y$ maps **closed sets** to closed sets.

CONDUCTOR
: A integer associated with an **elliptic curve** that contains information about the **reduction** of the curve at any prime number.

COVERING
: A mapping of **topological spaces**, $p : X \to Y$ such that for any y in Y there is an **open set** containing y whose pre-image (i. e., set of points that map to it) is a union of open sets of X.

DIFFERENTIAL FORM
: A fundamental object related to the **differential geometry** of a **manifold**. It is a way to define "partial differentiation" of a function on a manifold in a manner that takes account of the geometry of the manifold.

DIFFERENTIAL GEOMETRY
: The study by techniques of differential calculus of the geometric properties of **manifolds**.

DIOPHANTINE EQUATION
: A polynomial equation with rational coefficients and one or more variables for which integral or rational solutions are sought.

DIRICHLET SERIES
: An infinite series of the form $F(s) = \sum_{n=1}^{\infty} a_n/n^s$. Typically such a series converges in some part of the complex plane where $Re(s)$ is sufficiently large, and it defines an **analytic function** there.

DISCRIMINANT
: A function of the coefficients of a polynomial of one variable. It is essentially the product of all possible differences of roots of the polynomial. Therefore the discriminant is zero if and only if the polynomial has repeated roots.

EIGENFUNCTION
: A function in a **vector space** of functions whose image under some linear transformation on the space is a constant multiple of itself. i.e., there is some transformation T such that $T(f) = cf$ for some constant c.

EIGENVALUE
: The constant multiple associated with a particular linear transformation and **eigenfunction**. i.e. the constant c such that $T(f) = cf$.

ELLIPTIC CURVE
: A non-singular complete **algebraic curve** of **genus** one. In more elementary terms, it is the locus of points satisfied by an equation of the form ... where the right hand side of the equation has no repeated roots.

ELLIPTIC FUNCTION
: A doubly-periodic **meromorphic** function. i.e., there are two periods ω_1 and ω_2, whose ratio isn't a real number, such that $f(z + \omega_1) = f(z + \omega_2) = f(z)$ for all complex z.

EQUIVALENCE RELATION
: A 2-place set-theoretic relation R that is reflexive (i.e. xRx), symmetric (i.e. xRy if and only if yRx), and transitive (i.e. xRy plus yRz imply xRz). An equivalence relation on a set partitions the set into disjoint equivalence classes (subsets of elements which are all equivalent to each other).

EULER PRODUCT
: An infinite product of the form $F(s) = \prod_p 1/(1-a_p p^{-s})$ for a complex variable s and complex values a_p, with p a prime number. When such a product converges, it can be represented by a **Dirichlet series**. There are generalizations whose factors are more complicated expressions involving prime numbers.

EXTENSION FIELD
: A **field** that contains a smaller field. Usually it consists of a given "base" field to which has been "adjoined" one or more roots of a set of polynomial equations of one variable.

FIELD
: A mathematical system that has two distinct operations, where both operations satisfy the axions of an **abelian group**. Usually the operations are expressed as addition and multiplication. (Zero is excluded from the multiplicative group since it has no inverse.) The most common examples of fields are the rational numbers $\mathbb{Q}$, the real numbers $\mathbb{R}$, the complex numbers $\mathbb{C}$, and **finite fields**.

FINITE FIELD
: A **field** with a finite number of elements. The simplest example is $\mathbb{Z}/p\mathbb{Z}$, also written F_p the integers modulo a prime p. All other finite fields are finite extensions of F_p.

FOURIER SERIES
An infinite series of the form $f(z) = \sum_{n=-\infty}^{\infty} c_n e^{2\pi i n z}$ Such series represent singly periodic **meromorphic functions** $f(z)$, where $f(z+1) = f(z)$ for all z. There is an extensive theory developed around the properties of such series, having many uses in both theoretical and applied mathematics.

FRACTIONAL LINEAR TRANSFORMATION
A transformation of the complex plane of the form $T(z) = (az+b)/(cz+d)$ for a, b, c, d in $\mathbb{C}$. The set of all fractional linear transformations with coefficients in $\mathbb{Z}$ forms a group called the **modular group**.

FUNCTIONAL EQUATION
An equation, which can be of many different forms, that prescribes certain properties of a function. The solutions of the equation (if any) are the functions having the property. Examples include differential equations, the periodicity property $f(z+t) = f(t)$, the symmetry property of an **automorphic function**, and relationships between function values at different points such as the functional equations of the **Riemann zeta function** and L**-functions**.

FUNDAMENTAL DOMAIN
A connected region of the complex plane that contains exactly one representative of each **orbit** under the action of some subgroup of the **modular group**. For any given subgroup, there is usually an obvious choice for the fundamental domain.

GALOIS GROUP
A group of permutations of the roots of a polynomial equation of one variable over some **field**.

GALOIS REPRESENTATION
A **group representation** of the **Galois group** of all **algebraic numbers** over the rationals. Galois representations can be constructed using any **elliptic curve**.

GALOIS THEORY
The theory of solutions of polynomial equations over a **field**. The theory uses **Galois groups** to describe all possible **extension fields** of a given field by means of a correspondence with subgroups of the Galois group.

GENERAL LINEAR GROUP
The **group** of all invertible n-by-n matrices with coefficients in a **ring** or, more usually, a **field**. Such a group is usually denoted by $GL_n(R)$ or $GL(n, R)$.

GENUS
A numerical integer invariant of an **algebraic curve**. As applied to a topological object such as 2-dimensional **manifold**, it can be interpreted as the number of "handles" the object has. E. g. a sphere has genus 0, while a torus (donut shape) has genus 1. There are various other definitions, such as the dimension of the space of **differential (1-)forms**.

GROUP
A mathematical system consisting of a set with an operation between elements of the set and the properties that the operation is associative (i. e. $(ab)c = a(bc)$), has an "identity element" (i. e. $1a = a$ for all a), and all elements have inverses (i. e. an a^{-1} with $aa^{-1} = 1$). Groups are used pervasively in mathematics, and they often express symmetry properties of other sets or objects.

GROUP REPRESENTATION
A **homomorphism** from an abstract **group** to a **general linear group**. In general, it need not be either injective (1–1) or surjective (onto). Group representations have numerous theoretical and applied uses, since matrix groups have well-known properties and are easy to compute with.

HOLOMORPHIC FUNCTION
An **analytic function**. The terms are used interchangably.

HOMEOMORPHISM
A 1–1 mapping between **topological spaces** that preserves the topological structure, i. e. the open sets. Homeomorphic spaces have essentially the same topological properties.

HOMOMORPHISM
A mapping between mathematical structures of the same type (e. g. **groups** or **rings**) that preserves the structure, i. e. $f(ab) = f(a)f(b)$.

ISOMORPHISM
A mapping between mathematical strucures of the same type that preserves the structure and is both injective (1–1) and surjective (onto). Isomorphic objects are essentially the same with respect to the preserved structure.

KERNEL
: A subset of the domain of a **homomorphism** consisting of all elements that map to the identity element. The kernel is always a sub-object of appropriate type. E. g. the kernel of a **group** homomorphism is a subgroup.

LATTICE
: A discrete subgroup of the additive **group** of complex numbers. Concretely, it is the set of all complex numbers of the for $n_1\omega_1 + n_2\omega_2$ for integers n_1, n_2 and "periods" ω_1 and ω_2 (whose ratio is not a real number).

L-FUNCTION
: A complex function which can usually be represented as a **Dirichlet series** or Euler product and which expresses arithmetic properties of some mathematical construct such as an **elliptic curve** or **modular function**. L-functions are a powerful theoretical tool for "encoding" arithmetic information in a single object.

LIE GROUP
: An analytic **manifold** G that has a **group** structure such that that the map $(x, y) \to xy^{-1}$ from $G \times G$ to G is analytic (i. e. infinitely differentiable). The **general linear groups** $GL_n(\mathbb{R})$, $GL_n(\mathbb{C})$, and their subgroups, are the most common examples. There is an extensive and deep theory of Lie groups, with many theoretical and applied uses.

LINEAR ALGEBRAIC GROUP
: A group that is isomorphic to a subgroup of a **general linear group**. It is one of the two distinct types of **algebraic groups**.

MANIFOLD
: A topological space that is "locally Euclidean" in the sense that every point is contained in an **open set** that is homeomorphic to an open set of Euclidean n-space $\mathbb{R}^n$. Additional conditions such as differentiability are often imposed. The manifold is the primary object of study in **differential geomentry**.

MEROMORPHIC FUNCTION
: An function that is an **analytic function** except at a discrete set of points where it has singularity called "poles". At such a point, the power series expansion of the function has a finite number of terms with negative powers of z.

MODULAR CURVE
: A **complete algebraic variety** which is an **algebraic curve** that is essentially the **quotient space** of the upper half of the complex plane by the action of a subgroup of finite index of the **modular group**. This space is "compactified" by the addition of a finite number of points in the same way as the **Riemann sphere** is constructed.

MODULAR ELLIPTIC CURVE
: An **elliptic curve** E for which there is a **modular curve** X of a certain kind and a surjective map $X \to E$. Such an elliptic curve is said to have a "parameterization by modular functions". There are equivalent definitions, the simplest of which is that there exists a **modular form** whose L**-function** is the same as that of E. The **Taniyama-Shimura conjecture** states that every elliptic curve is modular.

MODULAR FORM
: A holomorphic **modular function**. The term is usually applied in a more general sense in the same way as with modular functions, i. e. including functions with non-zero weight and with respect to subgroups of finite index in the **modular group**.

MODULAR FUNCTION
: A special type of **automorphic function** where the group involved is the **modular group**. The term is usually applied in a more general way. First, the automorphicity condition is relaxed to $f((az + b)/(cz + d)) = (cz + d)^k f(z)$, where the integer k is called the "weight" of f. Second, the condition may be applied only for certain subgroups of finite index in the modular group.

MODULAR GROUP
: The group of all **fractional linear transformations** of the complex plane with coefficients in $\mathbb{Z}$. This is essentially the same, up to a factor of ± 1, as $SL_2(\mathbb{Z})$, the group of 2-by-2 matrices with entries in $\mathbb{Z}$ and determinant 1, so sometimes $SL_2(\mathbb{Z})$ is referred to as the modular group.

MORPHISM
: A mapping between two mathematical objects of the same type, such as **topological spaces** or **groups**, that preserves the essential structure of the object.

NEIGHBORHOOD
An **open set** in a **topological space** that contains a specific point.

NORMAL SUBGROUP
A subgroup H of a **group** G with the property that $H = xHx^{-1}$ for any x in G. The **kernel** of a group homomorphism is always normal, and a normal subgroup is the kernel of the projection map $G \to G/H$ onto the **quotient group**.

OPEN SET
A member of a class of subsets of a **topological space** that satisfy certain axioms and define the topology.

ORBIT
A set of points of the complex plane containing all points which are equivalent under all transformations in some subgroup of the **modular group**.

p-ADIC NUMBERS
A **field** containing the ordinary rational numbers, defined for any prime p. There are several ways the construction can be performed, such as formal power series in p, as the "completion" of $\mathbb{Q}$ with respect to a metric based on divisibility by p, or as the field of quotients of a "projective limit" of the sequence of groups $\mathbb{Z}/p^n\mathbb{Z}$. There are p-adic analogues of many concepts of complex analysis.

PROJECTIVE PLANE
A geometric construct frequently used in **algebraic geometry** to make the equations easier to deal with and avoid having to treat the "point at infinity" as a special case.

QUOTIENT GROUP
The **group** obtained from a group G with **normal subgroup** H by putting a natural group structure on the set of equivalence classes of elements of G under the **equivalence relation** that $x \equiv y$ if and only $xy^{-1} \in H$. The quotient group is denoted by G/H.

QUOTIENT SPACE
A **topological space** obtained in a natural way from another space on which there is an **equivalence relation**. The points of the quotient space are equivalence classes, and the topology is the strongest one such that the projection map is continuous.

RANK
A numerical invariant of a finitely generated **abelian group**. Such a group is isomorphic to a product of a finte group and a finite number of infinite cyclic groups. The rank is the number of infinite cyclic groups in this product.

REDUCTION
The process of viewing an **algebraic curve** defined by a polynomial with integral coefficients as a curve over the **finite field** F_p for some prime p.

RIEMANN SPHERE
The **Riemann surface** obtained by "compactifying" the complex plane by adding a "point at infinity" and appropriate **neighborhoods** of that point.

RIEMANN SURFACE
A topological **manifold** that serves as the domain of definition of a single-valued **algebraic function**. The precise construction is rather technical, but greatly simplifies many ideas in complex function theory.

RIEMANN ZETA FUNCTION
The **Dirichlet series** $\zeta(s) = \sum_{n=1}^{\infty} 1/n^s$. The zeta function has many surprising analytic and number theoretic properties, and has been one of the central objects of study in analytic number theory. Many of its properties can be generalized to L-**functions**.

RING
A mathematical system that has two operations, usually called addition and multiplication. A ring is an **abelian group** with respect to addition. Multiplication is associative and distributive with respect to addition.

SPECIAL LINEAR GROUP
A subgroup of the **general linear group** consisting of matrices having determinant 1. It is usually denoted $SL_n(R)$ or $SL(n, R)$ for some **ring** R.

TANIYAMA-SHIMURA CONJECTURE
The conjecture that every **elliptic curve** is actually a **modular elliptic curve**. It has now been proven for the case where the **conductor** of the curve is square-free (the "semistable" case).

TOPOLOGICAL SPACE
: A set together with a collection of subsets, called **open sets** that satisfy certain axioms. The open sets endow the space with a concept of "nearness" between any two points. This is a generalization of the concept of nearness obtained from a numerical measure of "distance" between two points.

VECTOR SPACE
: A mathematical system consisting of a set of points ("vectors") that form an **abelian group** and which allow for "multiplication" by elements of some "field". Examples of vector spaces include n-tuples of elements of a field and various classes of **analytic functions**. A vector space is the fundamental object of study in linear algebra.

References for Fermat's Last Theorem.

[Cox] D. A. Cox, *Introduction to Fermat's Last Theorem*, Amer. Math. Monthly 101(1994) 3–14

[Gou] F. Q. Gouvea, *"A Marvelous Proof"*, Amer. Math. Monthly 101(1994) 203–222

[Hus] D. Husemoller, *Elliptic Curves*, Springer-Verlag (1987)

[Kob] N. Koblitz, *Introduction to Elliptic Curves and Modular Forms*, Springer-Verlag (1984)

[Maz] B. Mazur, *Number Theory as Gadfly*, Amer. Math. Monthly 98(1991) 593–610

[Rib] K. A. Ribet, *Galois Representations and Modular Forms*, Bull. AMS 32,4(1995) 375–402

[Rih] K. A. Ribet, B. Hayes, *Fermat's Last Theorem and Modern Arithmetic*, Amer. Sci. 82(1994) 144–156

[Rus] K. Rubin, A. Silverberg, *A Report on Wiles' Cambridge Lectures*, Bull. AMS 31,1(1994) 15–38

[Ser] J.-P. Serre, *A Course in Arithmetic*, Springer-Verlag (1973)

[Sil] J. H. Silverman, *The Arithmetic of Elliptic Curves* Springer-Verlag (1986)