Elliptic Curves

Kenneth A. Ribet

UC Berkeley

PARC Forum
October 17, 2008

In talking about elliptic curves, one can do no better than to quote the Yale mathematician Serge Lang (1927–2005), who began one of his many monographs as follows:

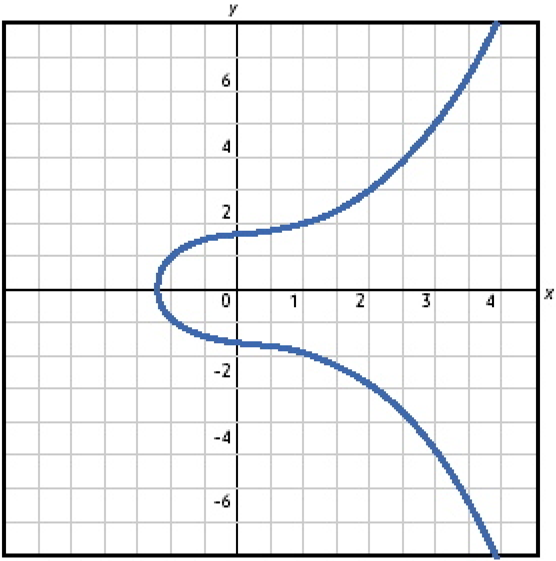
It is possible to write endlessly on elliptic curves. (This is not a threat.)

Although I have no intention of talking endlessly today, my goal is to recall how elliptic curves intervened in the proof of Fermat's Last Theorem to discuss some subsequent developments.

Elliptic curves are initially cubic equations such as

$$y^2 = x^3 + 1, \quad y^2 = x^3 - x, \quad y^2 = x^3 + x.$$

They become curves when we draw their graphs.


y
6
4
2
0
1
2
3
4
x
-2
-4
-6

In antiquity, elliptic curves arose in connection with Diophantine problems. (Diophantus of Alexandria lived in the third century AD.) For a simple example, we might consider solutions to $y^2 = x^3 + 1$ when x and y are *integers* (or whole numbers). For $x, y \geq 0$, there are the solutions $1^2 = 0^3 + 1$ and $3^2 = 2^3 + 1$. If we allow y to be negative, we get a few more solutions by flipping the sign of y. If we allow x to be negative as well, then we get the genuinely new solution $0^2 = (-1)^3 + 1$.

Are there any more solutions with x and y integral?

In turns out in fact that there are no other solutions even if we allow x and y to be fractions (rational numbers).

The embarrassing fact is that problems like this are still *hard*, even 1700 years after Diophantus.

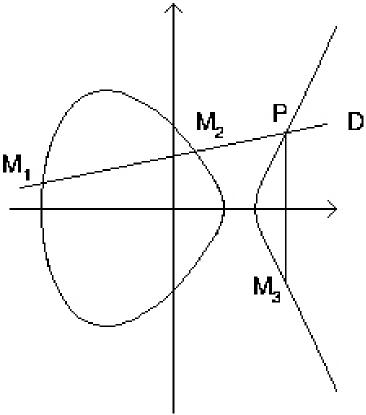
Indeed, if a and b are integers, it is nowadays not hard to figure out whether there are finitely many or infinitely many solutions to $y^2 = x^3 + ax + b$ in rational numbers. However, the methods that are used do not yield an algorithm that decides between “finite” and “infinite” in a predictable amount of time.

Fortunately, a tremendous amount is known about the set of solutions to $y^2 = x^3 + ax + b$ in rational numbers.

A basic fact is that this set is actually an abelian group! To make up the group structure, it is necessary to add to the set of points one additional point—the point at infinity. This point is the one extra point that we discover when we consider solutions to $y^2 = x^3 + ax + b$ in the projective plane, not just in the regular (affine) plane.

If one considers this extra point to be the 0-point of the abelian group, we get a group structure by declaring that three points P, Q and R sum to the 0-point whenever they are distinct and all lie on a straight line.

Equivalently, to get the sum M_3 of two distinct points M_1 and M_2, draw a line through them and find the three points where this line intersects the elliptic curve. Two of the three are M_1 and M_2; the third (P, in this diagram that I grabbed off the web) is $-(M_1 + M_2)$ in the group law.



The process of adding points together is called the *chord and tangent* process. Tangents arise if you add a point to itself: the chord connecting two close points on the curve becomes a tangent line in the limiting case when the points coalesce.

A fundamental theorem of L. E. Mordell (1922) states that the group of rational points on an elliptic curve is *finitely generated* in the sense that there's a finite set of points $P_1, \ldots, P_t$ such that each point on the curve with rational x and y can be gotten from this initial stock of points by repeated application of the geometric chord and tangent process.

What we don't have is a finitely terminating algorithm that inputs rational numbers a and b and outputs points $P_1, \ldots, P_t$ that generate all rational points on $y^2 = x^3 + ax + b$.

Perhaps because the group of rational points on an elliptic curve is so intractable, one is led to consider the easier problem is counting the number of solutions to equations $y^2 = x^3 + ax + b$ modulo prime numbers. The idea is the following: We fix a and b and let p be a (varying) prime number. Instead of considering $y^2 = x^3 + ax + b$ as a literal equation, we think of it as defining a *congruence modulo p* and ask for the number of solutions of this congruence.

The number is finite because there are only a finite number of x and y modulo p. In fact, since there are only p possibilities for each of x and y, the number of solutions to the congruence is at most $p \cdot p = p^2$. (If we want to count the point at infinity, we should add 1 to get the total number of projective points.)

One might imagine that the number of solutions is close to p^2, but in fact a theorem of H. Hasse (c. 1937) states that the number is approximately p. If N_p is the number of (affine) solutions to $y^2 = x^3 + ax + b$ mod p, then we have

$$|N_p - p| \leq 2\sqrt{p}.$$

It is traditional to define an error term a_p by the equation

$$N_p = p - a_p,$$

so that a_p is an integer satisfying $|a_p| \leq 2\sqrt{p}$.

For an especially nice example, consider the elliptic curve $y^2 = x^3 - x$ and take p to be a prime number other than 2 or 3. We have

p	3	5	7	11	13	17	19	23	29	31	37	41	$\cdots$
a_p	0	-2	0	0	6	2	0	0	-10	0	-2	10	$\cdots$.

The table suggests that the error term is 0 for primes that are 3 mod 4 and is twice an odd number for the primes that are 1 mod 4.

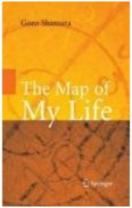
The first observation can be explained by a relatively elementary argument. The second is explained by Gauss's beautiful formula for the error term a_p when p is 1 mod 4. Such primes can be written $r^2 + t^2$ where r and t are positive integers; r and s are unique once we agree that r will be odd and s will be even. What Gauss showed is that $a_p = \pm 2r$; the sign is determined by the requirement that $\pm r + s - 1$ should be divisible by 4.

For example, $29 = 5^2 + 2^2$, while $41 = 5^2 + 4^2$. To make $\pm 5 + 2 - 1$ be divisible by 4, we take the minus sign; to make $\pm 5 + 4 - 1$ be divisible by 4, we take the plus sign.

That there's an explicit recipe for the error terms a_p as p varies turns out to be an accident that can be traced to the extra symmetry of the equation $y^2 = x^3 - x$. For random integers a and b, the elliptic curve $y^2 = x^3 + ax + b$ has no symmetry and there is no explicit formula for the error terms a_p.

On the other hand, the link between elliptic curves and modular forms (established in the 1990s) provides an amazing far-reaching generalization of Gauss's formula.

The *modularity of elliptic curves* was first stated as a conjecture in the middle of the last century. There is some dispute as to the origin of the conjecture, but there is no doubt that Goro Shimura was one of the first people to understand that every elliptic curve would be linked to modular forms.



Shimura's "The Map of My Life" was published several weeks ago. It's a slim volume that's about Shimura's early life in Japan and about his mathematical career in Princeton.

The proof that elliptic curves are modular was initiated in Andrew Wiles's 1995 "Fermat" article and completed (for a large class of curves) in the companion article by Taylor and Wiles.

Over the next several years, the class of elliptic curves for which modularity could be proved was enlarged in a series of articles that culminated in the 2001 article “On the modularity of elliptic curves over $\mathbf{Q}$: Wild 3-adic exercises” by Christophe Breuil, Brian Conrad, Fred Diamond, and Richard Taylor.

The idea of modularity is that there is a stock of special complex functions—modular forms—that are defined in a branch of mathematics that might not seem to be related to number theory. The modular forms are described by formal power series (Fourier series)

$$f = c_1 q + c_2 q^2 + c_3 q^3 + \cdots$$

in which $c_1 = 1$ and the other coefficients are complex numbers. If it so happens that the cs are all integers, then there is an associated elliptic curve $y^2 = x^3 + ax + b$ whose arithmetic is linked to f via the relation $c_p = a_p$ for all prime numbers p.

The conjecture, now a theorem, is that the process could be reversed: for each elliptic curve there is a form f so that the c_p for f are the same numbers as the a_p for the curve.

When we study elliptic curves, we often list them in order of their *conductor*, a positive integer whose divisors are the prime numbers p for which the equation defining the curve leads to a singularity mod p. For such p, a_p cannot be defined by the rule that I explained, and there's an alternative definition that yields one of the three values -1, 0, $+1$ for a_p.

In any event, the smallest possible conductor is 11. For the curves with this conductor, the associated modular form is the infinite series obtained by expanding out the product

$$q \prod_{m=1}^{\infty} (1 - q^m)^2 (1 - q^{11m})^2.$$

In this (rather special) case, the product can be viewed as yielding a formula for the a_p, but the formula is different-looking from the formula that Gauss gave.

Before recalling the relation between modularity and Fermat's Last Theorem, we should pause to review the history of this result.

The story starts with Pierre de Fermat's marginal note to the effect that $a^n + b^n = c^n$ has no solutions in non-zero integers a, b and c when n is an integer ≥ 2. This assertion was actually proved by Fermat himself for $n = 4$; Fermat proved that a fourth power plus a square can never be a fourth power, which is more than enough to prove what is wanted for $n = 4$.

In the 17th century, Euler treated the case $n = 3$ by a method that is recounted in many textbooks; see for example “A Classical Introduction to Modern Number Theory” by K. Ireland and M. Rosen. Both Fermat and Euler implicitly considered elliptic curves!

Once one knows the cases $n = 4$ and $n = 3$, one can concentrate on the case $n = r$, where r is a prime number ≥ 5.

There is a well known false proof for this case that relies on the implicit assumption that cyclotomic integers (expressions involving $e^{2\pi i/r}$ as well as ordinary integers) factor uniquely in the same way that ordinary whole numbers factor. Unique factorization is true only for primes < 23. Kummer set everyone straight by showing that a modified form of unique factorization is all we need and by giving a simple criterion to decide whether or not this weak unique factorization is true for r. It's true, for example, for all primes less than 37, but false if $r = 37$. We believe that it's true for about 67% of all primes, but no one knows how to prove this!

The modern subject of algebraic number theory grew out of attempts to prove FLT by modifications of Kummer's techniques. Two books by Paulo Ribenboim make good reading: one is his "13 Lectures on Fermat's Last Theorem" and the other is "Fermat's Last Theorem for Amateurs."

It seems to be abundantly clear that the techniques of algebraic number theory alone are not sufficient to prove the Theorem, but people keep trying. I receive many manuscripts and e-mail messages from amateur mathematicians who believe that they have found direct proofs of FLT.

The proof that works—the only one that we have so far–arose as follows. If $a^r + b^r = c^r$ (where a, b and c are non-zero, and r is a prime ≥ 5), G. Frey promoted the idea that it should be possible to prove that no modular form could be associated with the elliptic curve

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

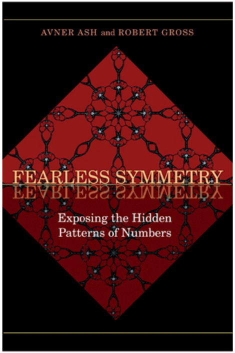
After Frey's idea has circulated for some months, I proved (in 1986) that his idea was indeed correct.

Once the modularity of elliptic curves was established, we had a proof of Fermat by contradiction: since a solution to Fermat's equation would give a non-modular elliptic curve, there are no solutions to Fermat's equation because there are no non-modular curves.

In the 15 years since the proof of FLT was completed (Fall, 1994), we have seen increasing generalizations of and applications of the modularity techniques that Taylor and Wiles pioneered in their article.

As we stressed in the 1990s, what Taylor and Wiles did in their article was to link modular forms to *Galois representations*. Since elliptic curves are tightly associated with Galois representations, it is nearly a tautology that modular forms are linked with elliptic curves whenever they are linked with the Galois representations that one attaches to elliptic curves.

Basically, Galois representations are all about symmetry. When I thought about explaining things that way, I remembered that there's a book by Avner Ash and Rob Gross that attempts exactly this approach.



Lots of books have been written about Fermat's Last Theorem. This one is written for a general reader and tries to do justice to the actual proof.

Maybe it's fruitful to give the general idea. We need to introduce some *fields*, systems of numbers that are closed under addition, subtraction, multiplication and division (by non-zero numbers). The set of rational numbers forms a field, **Q**. The set of complex numbers forms a field **C**. In between them is the field of algebraic numbers, $\overline{\mathbf{Q}}$. (A number is algebraic if it's a root of a a non-zero polynomial with rational coefficients.)

The Galois group of **Q** is the group $\mathcal{G}$ of symmetries of the field $\overline{\mathbf{Q}}$. This group is infinite and fear-inducing. People study it by looking at finite images of $\mathcal{G}$.

In the old days (starting with Galois and running through the mid-20th century) people explored $\mathcal{G}$ by starting with a polynomial like $x^4 - 3x^2 + 2x + 1$ and looking at the smallest field that contains all the roots of the equation in $\overline{\mathbf{Q}}$. Taniyama, Shimura, Serre, Tate and others stressed the importance of considering objects other than polynomials—things like elliptic curves.

If you take an elliptic curve and fix an auxiliary prime number ℓ, you end up with a finite image of $\mathcal{G}$ inside a group of 2×2 matrices, whose entries are integers mod ℓ. The association that yields a matrix for each symmetry in $\mathcal{G}$ is a Galois representation; it's the mod ℓ representation attached to the curve.

What is new is that you don't get finite images of $\mathcal{G}$ by starting with a polynomial and looking at its roots. Instead you start with an object coming from algebraic geometry and look at the symmetries of a finite set that's associated with the object. There's a famous 1967 article of Shimura that studies the mod ℓ Galois representations attached to the elliptic curves of conductor 11 that I mentioned before.

The proof of Fermat's last theorem leverages the fact that the various numbers a_p mod ℓ are visible from the mod ℓ Galois representation attached to an elliptic curve. If there's a modular form f whose prime-indexed coefficients agree with the a_p mod ℓ, we say that the representation is modular.

When Wiles announced his proof in 1993, he had at his disposal at theorem of Langlands to the effect that the mod 3 Galois representations attached to elliptic curves are all modular. The Taylor–Wiles article and its successors parlayed this input into the conclusion that elliptic curves are modular.

There's an increasingly sophisticated technology that enables one to prove that "big things" are modular if one knows that little pieces of the big things are modular. The buzzword for this technique is "relative modularity"; relative modularity is an outgrown of Taylor–Wiles.

Amazingly, there's a companion technique that tends to prove that the little pieces are modular. I'm thinking of the recent work by Khare and Wintenberger that establishes the modularity of mod ℓ representations like those coming from elliptic curves. How do they do this—you can't get something from nothing?!

Their work builds very elaborate bridges from the initial starting point back to a representation that is known already to be modular. It's as is Langlands' theorem about mod 3 representations is a sourdough starter that has enabled bakers around the world to make new bread.