The five fundamental operations of mathematics: addition, subtraction, multiplication, division, and modular forms

Kenneth A. Ribet

UC Berkeley

Trinity University
March 31, 2008

This talk is about counting, and it's about solving equations.

Counting is a very familiar activity in mathematics. Many universities teach sophomore-level courses on discrete mathematics that turn out to be mostly about counting. For example, we ask our students to find the number of different ways of constituting a bag of a dozen lollipops if there are 5 different flavors. (The answer is 1820, I think.)

Solving equations is even more of a flagship activity for mathematicians. At a mathematics conference at Sundance, Robert Redford told a group of my colleagues “I hope you solve all your equations”!

The kind of equations that I like to solve are *Diophantine* equations.

Diophantus of Alexandria (third century AD) was Robert Redford’s kind of mathematician. This “father of algebra” focused on the solution to algebraic equations, especially in contexts where the solutions are constrained to be whole numbers or fractions.

Here's a typical example. Consider the equation $y^2 = x^3 + 1$. In an algebra or high school class, we might graph this equation in the plane; there's little challenge. But what if we ask for solutions in integers (i.e., whole numbers)? It is relatively easy to discover the solutions $(0, \pm 1)$, $(-1, 0)$ and $(2, \pm 3)$, and Diophantus might have asked if there are any more.

There aren't, but this is far from obvious.

Which positive integers can be written as sums of squares? If m is a positive integer, we consider the Diophantine equation $m = x^2 + y^2$. We can ask:

- Does the equation have any solutions at all?
- How many solutions does the equation have?
- Can we find all solutions?

Let's look at three examples where we can answer the first question. We'll take three consecutive 6-figure prime numbers as values of m:

- If m is the prime number 144169, there's a positive answer because $m = 315^2 + 212^2$.
- If m is the prime 144173, we have similarly $m = 338^2 + 173^2$.
- If $m = 144203$, the equation has no solutions.

Fermat (1601–1665) solved the first problem definitively:

- The prime 2 is the sum of two squares.
- All primes that are 1 mod 4 are sums of two squares
- All other primes are not sums of two squares

If *m* is a given positive integer, factor *m* as a product of primes. For example,

$$123456 = 2^6 \cdot 3 \cdot 643, \qquad 1234567 = 127 \cdot 9721.$$

The rule is that *m* is the sum of two squares if and only if every prime that's 3 mod 4 appears in its factorization with an even exponent. Neither of 123456, 1234567 is a sum of two squares.

Fermat also solved the second problem by giving a formula in terms of m for the number of solutions to $m = x^2 + y^2$. Call this number $N(m)$. Thus $N(123456) = N(1234567) = 0$, and $N(144169)$ is non-zero.

In fact, it's pretty clear that $N(144169)$ is at least 8 because in the equation $144169 = 315^2 + 212^2$ we can change the signs of both 212 and 315 and we an flip the order of the two terms. For example, $144169 = (-212)^2 + 315^2$.

Fermat showed that $N(144169) = 8$, which we can paraphrase as the statement that 144169 is the sum of two squares in *essentially one way*.

For a slightly more complicated example, consider 65, which is both $1 + 64$ and $49 + 16$. By the same reasoning as for 144169, $N(65)$ is at least 16. Fermat showed that it's exactly 16.

Fermat's general formula for $N(m)$ is that it's 4 times the difference between the number of positive divisors of m that are 1 mod 4 and the number of divisors of m that are 3 mod 4.

For example, if p is a prime that's 3 mod 4, p has two divisors, namely 1 and p. One of them is 1 mod 4 and the other is 3 mod 4, so the difference between the counts is 0, and $N(m) = 0$.

If instead p is a prime that's 1 mod 4, the difference is 2, so $N(m) = 8$.

The number 65 has 4 divisors (1, 5, 13, 65), all congruent to 1 mod 4. Since the difference is 4, $N(65) = 16$.

Fermat's Last Theorem

It's time now to talk about the most notorious of Diophantine problems. Until its resolution in 1994, it was perhaps the single famous problem in mathematics.

To explain Fermat's Last Theorem, we can begin with the perfect squares: the numbers 0, 1, 4, 9, 16, and so on. Add two of them and you're unlikely to end up with a third; for example, $4 + 9 = 13$, a non-square.

"Unlikely" does not mean impossible, since some sums of two perfect squares *are* perfect squares: $3^2 + 4^2 = 5^2$, $5^2 + 12^2 = 13^2$, so on. If $a^2 + b^2 = c^2$, (a, b, c) is called a *Pythagorean triple*.

The ancient Greeks gave a general recipe for finding all Pythagorean triples. If n and m are positive integers with $n > m$, then

$$(n^2 - m^2)^2 + (2nm)^2 = (n^2 + m^2)^2.$$

For example, if $n = 3$, $n = 2$, we get the familiar triple $(5, 12, 13)$.

This recipe really does give all Pythagorean triples, up to the operations of scaling a triple and exchanging the first two entries.

What happens if we replace squares by cubes, fourth powers, and so on? Fermat proved that the sum of two non-zero perfect fourth powers is never a perfect fourth power.

After his death, Fermat's son discovered a marginal note by his father in which Fermat claimed to have proved that the sum of two non-zero perfect *n*th powers is never a perfect *n* power, when *n* is an exponent bigger than 2.

Most mathematicians believe that Fermat realized later in his life that his "proof" was mistaken. We don't know this for sure, and we have little idea of what his proof might have been.

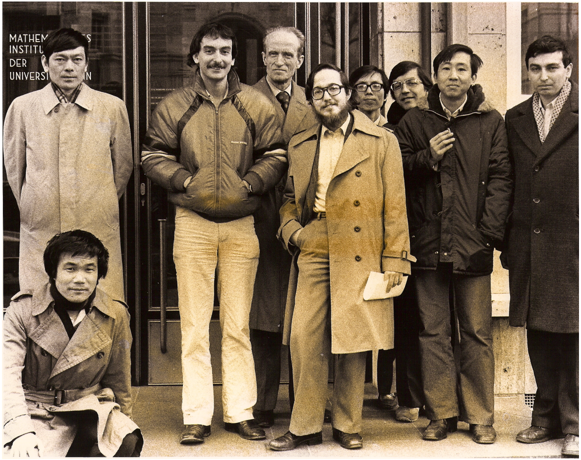
Euler (1707–1783) proved that the sum of two non-zero perfect cubes is never a perfect cube.

After Euler and Fermat, countless professional and amateur mathematicians attempted to discover a simple proof of the statement in Fermat's marginal note. This statement—Fermat's Last Theorem—was proved in 1994 by an argument engineered by Andrew Wiles and carried out partially by Wiles and partially in joint work by Wiles and Richard Taylor.

Andrew Wiles announced a proof of Fermat's Last Theorem in late June, 1993. A “gap” was found in the proof by Nicholas Katz soon after. The proof was revived in Fall, 1994 by a new argument that was crafted by Richard Taylor and Andrew Wiles.

The argument by Wiles and Taylor–Wiles made crucial use of *modular forms*, a magic power of number theory that was described by Martin Eichler (1912–1992) as the fifth fundamental operation of number theory.

A gathering at the University of Bonn around 1980.



Standing: Koji Doi, Ken Ribet, Martin Eichler, Don Zagier, Tsuneo Arakawa, Carlos Moreno, Masami Ohta, Yevsey Nisnevich. Kneeling: Hiroyuki Yoshida.

Modular Forms

Modular forms seem to come up everywhere in mathematics where one does systematic counts.

For example, consider the infinite series whose coefficients are the numbers $N(m)$ that we discussed before:

$$1 + 4q + 4q^2 + 4q^4 + 8q^5 + 4q^8 + 4q^9 + 8q^{10} + 8q^{13} + \cdots .$$

This infinite series, viewed in the right way, *is* a modular form.

In fact, this modular form can be used to give a short proof of Fermat's formula for $N(m)$.

Modular forms are special functions that are analogous to the trigonometric functions like sin, cos, tan,$\ldots$ in that they are periodic in the same way that sin is periodic. (Recall the formula $\sin(x + 2\pi) = \sin(x)$.) Modular forms have the periodicity of the trigonometric functions *plus* enough extra symmetries that they are essentially unchanged under a large group of substitutions.

Because of the symmetries, it is possible to write modular forms as Fourier series $\sum_{m=0}^{\infty} a_m q^m$, where the "$q$" here is a shorthand for $e^{2\pi iz}$.

When we say that the formal series $1 + 4q + 4q^2 + \cdots$ is a modular form, we mean that it becomes a modular form when we substitute $q = e^{2\pi iz}$.

To prove that $1 + 4q + 4q^2 + \cdots$ is a modular form is not too difficult. It's the square of a modular form that was studied by Carl Gustav Jacob Jacobi (1804–1851). It belongs to a space of modular forms that is so constrained that all non-zero modular forms in the space are multiples of each other. Also in this same space is the modular form corresponding to the series

$$1 + N'(1)q + N'(2)q^2 + \cdots ,$$

where the coefficients $N'(m)$ are as described before: $N'(m)$ is the number of divisors of m that are 1 mod 4 less the number that are 3 mod 4. Because of the constraint, the two modular forms are automatically multiples of each other. It follows they are equal because they begin with the same constant term, namely 1. This gives the "modular" proof of Fermat's formula for the number of solutions to $x^2 + y^2 = m$.

Now we can begin to talk about the way in which the proof of Fermat's Last Theorem uses modular forms. Here's a skeletal outline:

- You want to prove that there is no solution to $a^p + b^p = c^p$ when p is a prime number bigger than 2. Assume there is such a solution, (a, b, c), and try to get a contradiction.
- Using (a, b, c), set up an auxiliary Diophantine counting problem.
- Show that this problem yields a modular form.
- Pinpoint the modular form as a member of a highly constrained space of modular forms.
- Show in fact, that this space has no non-zero elements!
- That's a contradiction, so the proof is complete.

The auxiliary counting problem is associated to the simple-looking equation $y^2 = x(x - a^p)(x + b^p)$, where (a, b, c) is a Fermat counterexample.

Of course, it's impossible to give an actual example because there are no solutions to $a^p + b^p = c^p$! As a proxy, we can consider the even more simple equation

$$y^2 = x(x - 1)(x + 1) = x^3 - x.$$

It may recall the equation $y^2 = x^3 + 1$ that we encountered before. Both are examples of *elliptic curves*. Elliptic curves are basically just fancy names for cubic equations in two variables.

Elliptic curves give rise to counting problems (as I'll explain) and hence to infinite series $q + a(2)q + a(3)q^2 + \cdots$. For $y^2 = x^3 - x$, the series begins

$$q - 2q^5 - 3q^9 + 6q^{13} + 2q^{17} - q^{25} - 10q^{29} - 2q^{37} \cdots .$$

A famous conjecture (known as the modularity conjecture, the Taniyama–Shimura conjecture, the Shimura–Taniyama conjecture, Weil's conjecture, the Shimura–Taniyama–Weil conjecture,…) predicted that these infinite series are all modular forms.

Goro Shimura and Ken Ribet, summer 1973

Wiles's essential contribution was to break open this conjecture by finding a revolutionary method to prove modularity. He, along with Taylor, proved the modularity of lots of elliptic curves in 1994. A series of authors completed the proof of the modularity conjecture in 1999. The method launched by Taylor and Wiles has found numerous applications to other Diophantine problems.

My contribution was to show that if the counting problem is modular, then the associated modular form can be shoe-horned down into a space with no non-zero elements.

My work on this problem was done around 20 years ago. It provided the motivation for Wiles to work on the modularity of elliptic curves.

To conclude, I need to explain the counting problem for $y^2 = x^3 - x$. It will produce a number $a(\ell)$ for every prime number ℓ; this number becomes the coefficient of q^ℓ in the series for the curve. For example, $a(5) = -2$.

The remaining coefficients $a(m)$ are calculated from the prime-indexed coefficients $a(ell)$ by a relatively uninteresting combinatorial formula. Only the prime-indexed coefficients hold interest.

To calculate $a(\ell)$, we have to view $y^2 = x^3 - x$ as a congruence modulo ℓ. This just means systematically throwing out all integers that are multiples of ℓ. (Divide by ℓ and consider only the remainders.) The only relevant xs and ys are the $0, 1, 2, \ldots, \ell - 1$. Out of the ℓ^2 possibilities for (x, y), count the number of them that are solutions to $y^2 = x^3 - x$ as a congruence modulo ℓ. Call this number $C(\ell)$. Then $a(\ell) = \ell - C(\ell)$.

For $\ell = 5$ and $y^2 = x^3 - x$, the values 0, 1 and 4 for x give 0 mod 5 on the right-hand side, so there is only one y that satisfies the congruence, namely 0. The values 2 and 3 for x give 1 mod 4 and 4 mod 4 on the right-hand side; each of these correspond to two possible values of y. Altogether there are 7 possible (x, y), so $C(5) = 7$ and $a(5) = -2$, as announced.