The Proof of Fermat's Last Theorem by R. Taylor and A. Wiles

Gerd Faltings

The proof of the conjecture mentioned in the title was finally completed in September of 1994. A. Wiles announced this result in the summer of 1993; however, there was a gap in his work. The paper of Taylor and Wiles does not close this gap but circumvents it. This article is an adaptation of several talks that I have given on this topic and is by no means about my own work. I have tried to present the basic ideas to a wider mathematical audience, and in the process I have skipped over certain details, which are in my opinion not so much of interest to the nonspecialist. The specialists can then alleviate their boredom by finding those mistakes and correcting them.

Elliptic Curves

For our purposes an elliptic curve E is given as the set of solutions $\{x, y\}$ of an equation $y^2 = f(x)$, where $f(x) = x^3 + \ldots$ is a polynomial of degree three. Usually E is defined over the rational numbers $\mathbf{Q}$; that is, the coefficients of f are in $\mathbf{Q}$. We also demand that all three zeros of f are distinct (E is "nonsingular"). We may consider E as those solutions in $\mathbf{Q}$, $\mathbf{R}$, or $\mathbf{C}$, denoted, respectively, $E(\mathbf{Q})$, $E(\mathbf{R})$, and $E(\mathbf{C})$. One usually includes in this set an infinitely distant point, denoted ∞. With this addition, the solution set has the structure of an abelian group, with ∞ as the neutral element. The inverse of (x, y) is $(x, -y)$, and the sum of three points vanishes if they lie on a line. The group addition is given by algebraic functions. As a group $E(\mathbf{Q})$ is finitely generated (Mordell's Theorem), $E(\mathbf{R})$ is isomorphic to $\mathbf{R}/\mathbf{Z}$ or to $\mathbf{R}/\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z}$, and $E(\mathbf{C}) \cong \mathbf{C}/\text{lattice}$ (for example, $y^2 = x^3 - x$ yields the lattice $\mathbf{Z} \oplus \mathbf{Z}i$). For an integer n let $E[n]$ denote the n-division points, that is, the kernel of multiplication by n. Over $\mathbf{C}$ these are isomorphic to $(\mathbf{Z}/n\mathbf{Z})^2$, and the coordinates are algebraic numbers. For example, the 2-division points are exactly ∞ and the three zeros of f (where $y = 0$). The absolute Galois group $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ acts on

Gerd Faltings is affiliated with the Max-Planck-Institute für Mathematik in Bonn, Germany.

Translated from Testausdruck DMV Mitteilungen 27. März 1995 für 2/95 *by Uwe F. Mayer, University of Utah.*

Andrew Wiles

them, since the determining equations have coefficients in $\mathbf{Q}$. This yields Galois representations $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_2(\mathbf{Z}/n\mathbf{Z})$. Using a change of coordinates, one can arrange that f has integer coefficients. If one then reduces modulo a prime number p, one obtains a polynomial over the finite field $\mathbf{F}_p$. If the zeros of the reduced polynomial are distinct then this yields an elliptic curve over $\mathbf{F}_p$. This is true for all prime numbers p except for the finitely many prime divisors of the discriminant of f. Also the choice of f is not unique, but we do say that E has *good reduction at* p if we can find an f such that the zeros modulo p are distinct. (These observations are not completely true at $p = 2$ because of the term y^2 .) Otherwise E has *bad reduction at* p. If in this case only two zeros of f modulo p coincide, one says E has *semistable bad reduction*. E is called *semistable* if at all p it has either good or semistable reduction. The curve $y^2 = x^3 - x$ is not semistable at $p = 2$ (no CM-curve is semistable).

An example (which in the end will not exist) of a semistable curve is the Frey curve. To a solution of Fermat's equation $a^l + b^l = c^l$ (where a, b, c are relatively prime, and $l \geq 3$ is prime) one associates the curve

$$E : y^2 = x(x - a^l)(x - c^l).$$

This curve has bad reduction exactly at the prime divisors of abc. It has the following noteworthy property: Consider the associated Galois representation $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_2(\mathbf{F}_l)$. This representation is unramified (the analog of "good reduction") at all prime numbers p at which E has good reduction. Here one might have to say "crystalline" for "unramified" if $p = l$. Because of the particular form of the equation for E, this is also true at all prime divisors $p > 2$ of abc. Therefore the l-division points behave as if E had good reduction at all $p > 2$. However, as we shall see, there are no semistable elliptic curves over $\mathbf{Q}$ with this property, and this is the desired contradiction.

In order to reach the goal this way, one has to replace elliptic curves by modular forms. That this can be done follows from the conjecture of Taniyama-Weil (which essentially is due to Shimura). If E satisfies the conclusion of this conjecture, that is, if E is "modular", then according to a theorem of K. Ribet one can find a modular form for $\Gamma_0(2)$ which corresponds to the representation of $E[l]$. However, there are no such modular forms. The content of the papers by R. Taylor and A. Wiles is exactly the proof of the Taniyama-Weil conjecture for semistable elliptic curves over $\mathbf{Q}$. To explain this we need a few basic facts about modular forms.

Modular Forms

Let $\mathbf{H} = \{\tau \in \mathbf{C} \mid \mathrm{Im}(\tau) > 0\}$ be the upper half plane, on which $\mathrm{SL}(2, \mathbf{R})$ acts by the usual $(a\tau + b)/(c\tau + d)$-rule. The subgroup $\Gamma_0(N)$ of $\mathrm{SL}(2, \mathbf{Z})$ consists of those matrices

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

with $c \equiv 0 \mod N$. A modular form (of weight 2) for $\Gamma_0(N)$ is a holomorphic function $f(\tau)$ on $\mathbf{H}$ with

$$f((a\tau + b)/(c\tau + d)) = (c\tau + d)^2 f(\tau)$$

for

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N)$$

and $f(\tau)$ "holomorphic at the cusps". This last statement means in particular for the Fourier series (since $f(\tau + 1) = f(\tau)$)

$$f(\tau) = \sum_{n \in \mathbf{Z}} a_n \cdot e^{2\pi i n \tau}$$

that all a_n vanish for $n < 0$. If additionally $a_0 = 0$, then f is called a *cusp form*. The Hecke algebra $\mathbf{T}$ acts on the space of cusp forms. It is generated by Hecke operators T_p (for $p \nmid N$ prime) and U_p (for $p \mid N$). For the Fourier coefficients one has

$$a_n(T_p f) = a_{np}(f) + p a_{n/p}(f),$$
$$a_n(U_p f) = a_{np}(f).$$

An eigenform is a common eigenform of all Hecke operators. One can always normalize it so that $a_1(f) = 1$; then $a_p(f)$ is the corresponding eigenvalue of T_p or U_p. The above equations allow one to determine all a_n recursively, and therefore one can determine the eigenform f. Conversely, one can construct for a given sys-

tem $\{a_p\}$ of eigenvalues a Fourier series $f(\tau) = \sum a_n e^{2\pi i n\tau}$. According to a theorem of A. Weil this is a modular form if and only if the L-series $L(s,f) = \sum_{n=1}^{\infty} a_n n^{-s}$ has a holomorphic extension to the full s-plane and satisfies a suitable functional equation. (This also must be true for twists by Dirichlet characters.)

In case all a_p are in $\mathbf{Q}$, the eigenform f has an associated elliptic curve E with good reduction outside the prime divisors of N. For $p \nmid N$ the number of the $\mathbf{F}_p$-rational points $E(\mathbf{F}_p)$ is equal to $\sharp E(\mathbf{F}_p) = p + 1 - a_p$. Conversely one can define for each elliptic curve E over $\mathbf{Q}$ a Hasse-Weil L-series $L(s,E)$, and it is conjectured that it has the nice properties from above. According to the theorem of A. Weil it should thus belong to an eigenform with rational eigenvalues. This is the content of the Taniyama-Weil conjecture.

Even when the coefficients a_p are not in $\mathbf{Q}$, one can construct a Galois representation associated to the eigenform.

The Hecke algebra $\mathbf{T}$ is a finitely generated $\mathbf{Z}$-module. We now replace it by the completion $\hat{\mathbf{T}}$ at a suitable maximal ideal $\mathbf{m}$ (a "non-Eisenstein ideal"), with $\kappa = \mathbf{T}/\mathbf{m}$ denoting the residue class field of characteristic l. Then there is a two-dimensional Galois representation

$$\rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_2(\hat{\mathbf{T}}),$$

which is unramified (or crystalline, respectively) at $p \nmid N$, with

$$\mathrm{trace}(\rho(\mathrm{Frob}_p)) = T_p$$
$$\det(\rho(\mathrm{Frob}_p)) = p.$$

An eigenform with rational eigenvalues yields a homomorphism $\hat{\mathbf{T}} \to \mathbf{Z}_l$, and ρ induces the l-adic representation that is given by the associated elliptic curve E, describing the Galois action on all l^n-division points of E. Conversely, it is possible to show that E is modular if and only if the associated l-adic representation can be constructed in this manner.

Deformations

The l-adic representation is constructed for $l = 3$, starting with the representation on the 3-division points. This is known to be congruent to a modular representation, and then the universal lifting of this representation is proven modular, which is the core of the proof. The prime 3 is very special here. So one starts with the consideration of $l = 3$.

One can restrict to the case that the 3-division points yield a surjective map

$$\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}(2, \mathbf{F}_3)$$

(in this argument 5-division points are also used once). As $\mathrm{PGL}(2,\mathbf{F}_3) \cong S_4$ (the symmetric group on the four elements of $\mathbf{P}^1(\mathbf{F}_3)$) is solvable, the representation on the 3-division points is already modular according to ("lifting") theorems by Langlands and Tunnell. This uses intensively the special properties of the prime number $l = 3$. For $l = 2$ the general theory does not work well for various reasons, and for $l \geq 5$ this beginning is impossible. We now look for a deformation argument for the representations modulo 9, 27, 81, 243, 729, etc., to be successively recognized as being modular. For this one uses the universal deformation of the representation modulo 3: There is a $\mathbf{Z}_3$-algebra $\mathcal{R}$ of the form $\mathcal{R} = \mathbf{Z}_3[[T_1, \ldots, T_r]]/I$ (I is an ideal), and a "universal" Galois representation

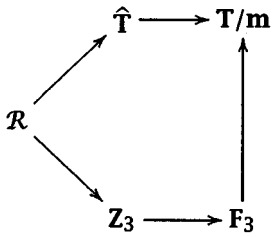
$$\rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_2(\mathcal{R})$$

with these properties:

1. ρ is unramified (or crystalline, respectively) for $p \nmid N$ (that is, E has good reduction at p);
2. ρ has certain local properties at $p \mid N$ ("certain" will not be discussed here);
3. $\det(\rho(\mathrm{Frob}_p)) = p$ for $p \nmid N$;
4. $\rho \mod (3, T_1, \ldots, T_r)$ is our given representation on $E[3]$;
5. any other representation $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_2(\mathcal{A})$ with the properties 1)–4) arises in a unique way via a homomorphism $\mathcal{R} \to \mathcal{A}$.

The construction of $\mathcal{R}$ follows general principles. Basically, one takes a set of generators $\{\sigma_1, \ldots, \sigma_s\}$ of the Galois group, and considers the ring of power series in $4s$ variables and divides by the smallest ideal I such that modulo I one obtains a representation with 1), ...,4), provided one assigns to σ_i the 2×2-matrix which has the four unknowns corresponding to σ_i as coefficients.

After the construction we get the following commutative diagram



where the two left mappings arise from the modular Galois representation and from the one of E. Wiles's idea is now to show that $\mathcal{R}$ is isomorphic to $\hat{\mathbf{T}}$, because then the elliptic Galois representation is automatically modular.

For this, naturally, one needs information on $\mathcal{R}$ that is not supplied by the general construction. Let W_n denote the adjoint Galois representation of $\mathrm{sl}(2, \mathbf{Z}/3^n\mathbf{Z})$ (2×2-matrices with

trace zero). Then, for example, the minimal number of generators $r(\mathcal{R} = \mathbf{Z}_3[[T_1, \dots, T_r]]/I)$ is given by $\dim_{\mathbf{F}_3} H^1_f(\mathbf{Q}, W_1)$, where H^1_f denotes a cohomology group satisfying certain local conditions corresponding to 1), 2) from above. This is also called a Selmer group. One sees this by setting $\mathcal{A} = \mathbf{F}_3[T]/(T^2)$ in the definitions. One can show (M. Flach) that the orders of $H^1_f(\mathbf{Q}, W_n)$ are uniformly bounded in n. These orders occur in the following numerical criterion for the equality $\mathcal{R} = \hat{\mathbf{T}}$: there is a $\mathbf{Z}_3$-homomorphism $\hat{\mathbf{T}} \to \mathcal{O}$, where $\mathcal{O}$ is the integral closure of $\mathbf{Z}_3$ in a finite extension of $\mathbf{Q}_3$. For simplicity we will assume that $\mathcal{O} = \mathbf{Z}_3$. It is known that $\hat{\mathbf{T}}$ is Gorenstein; that is, $\mathrm{Hom}_{\mathbf{Z}_3}(\hat{\mathbf{T}}, \mathbf{Z}_3)$ is a free $\hat{\mathbf{T}}$-module. The surjection $\hat{\mathbf{T}} \to \mathbf{Z}_3$ then has an adjoint $\mathbf{Z}_3 \to \hat{\mathbf{T}}$, and the composition of these two maps is multiplication by an element $\eta \in \mathbf{Z}_3$, which is well defined up to a unit. Furthermore, $\eta \neq 0$. On the other hand, let $\mathfrak{p} \subseteq \mathcal{R}$ be the kernel of the surjection $\mathcal{R} \to \hat{\mathbf{T}} \to \mathbf{Z}_3$. Then one has ("$\sharp$" = order) $\sharp\mathfrak{p}/\mathfrak{p}^2 \geq \sharp\mathbf{Z}_3/\eta \cdot \mathbf{Z}_3$ and equality if and only if $\mathcal{R} = \hat{\mathbf{T}}$ and this is also a complete intersection (I can be generated by r elements). The left-hand side $\sharp\mathfrak{p}/\mathfrak{p}^2$ is identical to the order of the Selmer group $H^1_f(\mathbf{Q}, W_n)$, for $n > 0$. The first attempt tried to establish equality by using Euler systems (invented by Kolyvagin). However, it was only possible to show that $\mathfrak{p}/\mathfrak{p}^2$ is annihilated by η. This is the content of the theorem of M. Flach. The higher levels of the Euler system, however, could not be constructed.

The Proof

One first shows the *minimal case* and then reduces to it. By the minimal case we mean that all primes of bad reduction occur already modulo 3 (and not only modulo higher powers). According to the theorem of Ribet and others (used for $l = 3$ and not for l the exponent of Fermat's equation), the Galois representation belonging to the curve modulo 3 is modular of level 3. In the minimal case the computation of Euler characteristics (Poitou-Tate) shows that $H^1_f(\mathbf{Q}, W_1)$ and $H^2_f(\mathbf{Q}, W_1)$ have the same dimension r. For each n one chooses r prime numbers $q_1, \dots, q_r \equiv 1 \mod 3^n$. Then one proceeds to use a subgroup of $\Gamma_0(N)$. This subgroup contains the intersection with $\Gamma_1(q_1 \cdots q_r)$, and the quotient is isomorphic to $G = (\mathbf{Z}/3^n\mathbf{Z})^r$. The associated Hecke algebra $\hat{\mathbf{T}}_1$ is a free module over $\mathbf{Z}_3[G]$, with G-coinvariants $\hat{\mathbf{T}}$, and is the quotient of a representation ring $\mathcal{R}_1 = \mathbf{Z}_3[[T_1, \dots, T_r]]/I_1$, which again can be generated by r elements. The ideal I_1 is small due to the free action of the group G. Now one takes a limit $n \to \infty$, and in the limit $\mathcal{R}_1$ and $\hat{\mathbf{T}}_1$ become rings of power series and equal. Furthermore, one obtains $\mathcal{R}$ from $\mathcal{R}_1$ and $\hat{\mathbf{T}}$ from $\hat{\mathbf{T}}_1$ in both cases by putting in the additional r relations "$\sigma_i = 1$", where $\sigma_1, \dots, \sigma_r$ are generators of G. Finally, $\mathcal{R} = \hat{\mathbf{T}}$, and this is a complete intersection.

To reduce to the minimal case, one estimates how both sides of the inequality

$$\sharp\mathbf{p}/\mathbf{p}^2 \geq \sharp\mathbf{Z}_3/\eta \cdot \mathbf{Z}_3$$

change as one proceeds from level M to a higher level N ($M|N$). For the left-hand side $\sharp H^1_f(\mathbf{Q}, W_n)$ certain local conditions are weakened, and one obtains an upper bound. For the right-hand side there is the phenomenon of "fusion", that is, of congruences between oldforms and newforms. Here a lower bound has been constructed by Ribet and Ihara. Luckily the two bounds agree, and thus everything is shown.