



HOW TO PROVE FERMAT'S LAST THEOREM

R. A. MOLLIN

Department of Mathematics and Statistics
University of Calgary
Calgary, Alberta, Canada, T2N 1N4
e-mail: ramollin@math.ucalgary.ca

Abstract

We provide an overview of the proof of Fermat's last theorem (FLT), by developing some very basic notions surrounding the theory of elliptic curves and modular forms. The actual proof is presented at the end, in one paragraph, known as the Frey-Serre-Rabin result.

1. Introduction

The proof of FLT, namely that $x^n + y^n = z^n$ has no nontrivial integer solutions x, y, z for $n \in \mathbb{N}$ with $n > 2$, is one of the most outstanding achievements of modern mathematics. Naturally, the proof of this centuries-old problem is difficult and lengthy, pulling on many areas for its conclusion. The aim of this note is to develop concepts to be able to state the *Shimura-Taniyama-Weil* (STW) conjecture and in one paragraph show how FLT follows from it. The STW conjecture was affirmatively settled and is arguably the most striking and important mathematical development of the twentieth century.

The STW conjecture involves certain elliptic curves and relations to modular forms. FLT would seem on the face of it to have no connections with elliptic curves

2000 Mathematics Subject Classification: Primary 11D41; Secondary 11-02, 11G05.

Keywords and phrases: Fermat's last theorem, Shimura-Taniyama-Weil conjecture, reduction index.

The author's research is supported by NSERC Canada grant # A8484.

Received April 6, 2009

since $x^n + y^n = z^n$ is not a cubic equation. However, in 1986 Gerhard Frey published [3], which associated, for a prime $p > 5$, the elliptic curve

$$y^2 = x(x - a^p)(x + b^p) \quad (1)$$

with nontrivial solutions to $a^p + b^p = c^p$. We call elliptic curves, given by equation (1), *Frey curves*. It turns out that this curve is of the type mentioned in the STW conjecture. In other words, existence of a solution to the Fermat equation would give rise to elliptic curves which would contradict STW. The curves in the STW conjecture are intimately related to certain modular forms, so now we need to describe the technical details. Some of the following is adapted from [5].

2. Elliptic Curves and Modular Forms

First, we let $SL(2, \mathbb{R})$ be the group of 2×2 -matrices with coefficients in \mathbb{R} and determinant 1. Then we let $\tilde{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, called the *Riemann sphere*. We begin with the following.

Definition 1 [Möbius Transformations]. Define an action of $SL(2, \mathbb{R})$ on $\tilde{\mathbb{C}}$ via the *fractional linear transformation*, also called a *Möbius transformation*, where

$$\alpha = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{R}):$$

$$\sigma : z \mapsto \alpha z = \sigma(z) = \begin{cases} (az + b)/(cz + d) & \text{if } z \in \mathbb{C} \text{ and } z \neq -d/c, \\ \infty & \text{if } z = -d/c, \\ a/c & \text{if } z = \infty \text{ and } c \neq 0, \\ \infty & \text{if } z = \infty \text{ and } c = 0. \end{cases}$$

A value $\sigma(\infty) = a/c \neq \infty$ is called a *cusp* of α .

It can be shown that the imaginary part of $\alpha z \in \mathbb{C}$ is given by

$$\Im(\alpha z) = \frac{\Im(z)}{|cz + d|^2}. \quad (2)$$

Now set

$$\mathfrak{H} = \{z \in \mathbb{C} : \Im(z) > 0\},$$

namely the complex upper half plane. Thus, by (2), the Möbius transformation σ maps $\mathfrak{H} \mapsto \mathfrak{H}$, which says that \mathfrak{H} is *stable*, meaning \mathfrak{H} is *preserved* under the

action of $SL(2, \mathbb{R})$. Also, since $\sigma(z) = \alpha z = -\alpha z$, namely α and $-\alpha$ represent the same transformation, $-1 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$ acts trivially on \mathfrak{H} , so the group

$$PSL(2, \mathbb{R}) = SL(2, \mathbb{R})/\{\pm 1\},$$

called the *projective special linear group*, is actually isomorphic to the group of fractional linear transformation. When we specialize to \mathbb{Z} , we have the following.

Definition 2 [The Modular Group]. The group

$$\Gamma = PSL(2, \mathbb{Z}) = SL(2, \mathbb{Z})/\{\pm 1\}$$

is called the *modular group*.

We now build upon the modular group Γ by presenting and studying forms related to it.

Definition 3 [Modular Forms and Functions]. A function $f(z)$ defined for $z \in \mathfrak{H}$ is called a *modular function of weight* $k \in \mathbb{Z}$ associated with the modular group Γ if the following properties hold:

- (a) f is analytic in \mathfrak{H} .
- (b) f satisfies the *functional equation*:

$$f(z) = (cz + d)^{-k} f\left(\frac{az + b}{cz + d}\right) = (cz + d)^{-k} f(\gamma z),$$

with $z \in \mathfrak{H}$ and $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$.

- (c) The Fourier series of f in the variable $q = \exp(2\pi iz)$ is given by

$$f(z) = \sum_{n=n_0(f)}^{\infty} c_n q^n, \quad (3)$$

where $n_0(f) \in \mathbb{Z}$.

A modular function of weight k is called a *modular form of weight* k if, in addition, $n_0(f) = 0$. In this case, we say that f is analytic at ∞ and write $f(\infty) = c_0$. In the case where $f(\infty) = c_0 = 0$, we say that f is a *cuspidal form*.

In the literature modular functions of weight k are sometimes called *weakly modular functions of weight k* or an *unrestricted modular form of weight k* . However, the definition of *modular form* or *cuspidal form* of weight k appears to be uniform. Sometimes the cuspidal form is referenced as a *parabolic form*.

Remark 1. If $\gamma = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$ in Definition 3, then $\gamma z = z$ for all $z \in \mathfrak{H}$.

Therefore, if f is a modular form of weight $k = 2m + 1$ for $m \in \mathbb{Z}$, then

$$f(z) = (-1)^{-k} f(\gamma z) = -f(z),$$

so if $f(z) \neq 0$, then dividing through the equation by $f(z)$, we get $1 = -1$, a contradiction. Thus, f is just the zero map, sometimes called *identically zero*. Hence, a nontrivial modular form on Γ must necessarily be of even weight. Also, by taking

$\gamma = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = T$ in Definition 3, we obtain that

$$f(z + 1) = f(z), \tag{4}$$

namely f is invariant under the transformation $z \mapsto z + 1$. This is what allows us to expand f into the expansion (3), which is called the *q-expansion of f* . (If we went into the details, we could invoke the Cauchy integral theorem using (4) to show symmetry in a certain line integral on $f(z) \exp(-2\pi iz)$, and the interested reader with knowledge of this area can derive the q-expansion in this fashion.) Note that condition (c) implies that if $z = x + yi$ and $y \rightarrow \infty$, then $q \rightarrow 0$ as $y \rightarrow \infty$. Thus the q-expansion (3), may be considered as an expansion about $z = \infty$, which justifies the reference to f being called *holomorphic at ∞* . The condition above for a cuspidal form tells us, therefore, that f vanishes as $y \rightarrow \infty$.

Example 1. The Eisenstein series of weight $k \geq 2$ are defined by the infinite series

$$G_{2k}(z) = \sum_{m, n \in \mathbb{Z} - (0, 0)} (nz + m)^{-2k}, \text{ for } \Im(z) > 0, \tag{5}$$

where the notation $m, n \in \mathbb{Z} - (0, 0)$ means that m and n run over all integers except that $m = n = 0$ is not allowed. The Eisenstein series of even weight are the first nontrivial examples of modular forms on Γ . Indeed, the following, which establishes this fact, is of interest from the viewpoint of arithmetic functions.

Theorem 1 [Eisenstein Series as Modular Forms]. For $q = \exp(2\pi iz)$ and $\Im(z) > 0$, the Eisenstein series given in (5) has Fourier expansion given by

$$G_{2k}(z) = 2\zeta(2k) + 2 \frac{(2\pi i)^{2k}}{(2k-1)!} \sum_{n=1}^{\infty} \sigma_{2k-1}(n) q^n,$$

where $k \geq 2$, $\zeta(s)$ is the Riemann ζ -function, and $\sigma_a(n) = \sum_{d|n} d^a$ is a sum of a -th powers of positive divisors of n . Accordingly, $G_{2k}(z)$ is a modular form of weight $2k$.

Next we need the following.

Definition 4 [Modular Discriminant Function and j -invariant]. Let $g_2 = 60G_4$ and $g_3 = 140G_6$. Then the function $\Delta : \mathfrak{H} \mapsto \mathbb{C}$ given by

$$\Delta = g_2^3 - 27g_3^2$$

is called the *discriminant function*, and the j -invariant is given by

$$j(\Delta) = \frac{1728g_2^3}{\Delta}.$$

The need for the coefficients in the definition of g_2 and g_3 will become clear when we link modular forms to elliptic curves later.

Remark 2. In the area of algebraic geometry, most of the interesting entities come into view when we look at arithmetically defined subgroups of finite index in Γ . One such class of groups is called *Hecke congruence subgroups* denoted by $\Gamma_0(n)$ for any $n \in \mathbb{N}$, defined by

$$\Gamma_0(n) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma : c \equiv 0 \pmod{n} \right\}.$$

It is known that the *index* of $\Gamma_0(n)$ in Γ is given by

$$|\Gamma : \Gamma_0(n)| = n \prod_{\substack{p|n \\ p=\text{prime}}} \left(1 + \frac{1}{p}\right),$$

the product over *distinct* primes dividing n .

An example of a modular form related to $\Gamma_0(n)$ is given by

$$f(z) = \eta(z)^2 \eta(11z)^2, \quad (6)$$

which is a cusp form of weight 2 related to the group $\Gamma_0(11)$. Here η is the Dedekind- η function

$$\eta(z) = q^{1/24} \prod_{n=1}^{\infty} (1 - q^n),$$

where $q = \exp(2\pi iz)$ and $q^{1/24} = \exp(\pi i/12)$.

Hecke groups defined in Remark 2, allow us to add another “level” to the notion of a modular form.

Definition 5 [Levels of Modular Forms]. If f is an analytic function on \mathcal{H} with

$$f(\gamma z) = (cz + d)^k f(z) \quad \text{for all } \gamma \in \Gamma_0(n),$$

and has a q -expansion

$$f(z) = \sum_{j=n_0(f)}^{\infty} a_j(f) q^j, \quad \text{where } q = \exp(2\pi iz) \text{ with } n_0(f) \in \mathbb{Z}, \quad (7)$$

then f is called a *modular function of weight k and level n* . A modular function of weight k and level n is called a *modular form of weight k and level n* if $n_0(f) = 0$. Moreover, if $a_0(f) = 0$, we call f a *cusp form of weight k and level n* . When $a_1(f) = 1$, and $a_0(f) = 0$, we say that f is a *normalized cusp form of weight k and level n* .

Spaces of modular, and cusp forms of weight k and level n are denoted by $M_k(\Gamma_0(n))$, respectively $S_k(\Gamma_0(n))$.

Example 2. It can be shown that $S_2(\Gamma_0(11))$ is a one-dimensional space spanned by equation (6), see [7, Remark 12.17, p. 351]. This example will have significant implications for a celebrated conjecture, see Example 6. Also, $S_2(\Gamma_0(2))$ is the zero space and this too will have implications for the proof of FLT, see Theorem 3.

Now we set the stage for bringing in elliptic curves.

Definition 6 [Elliptic Modular Functions]. If f is a function analytic on \mathbb{C} such that for $n \in \mathbb{N}$ and $z \in \mathbb{C}$,

$$f(\gamma z) = f(z) \text{ for all } \gamma \in \Gamma(n),$$

then f is called an *elliptic modular function*, where

$$\Gamma(n) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma : b \equiv c \equiv 0 \pmod{n} \right\}$$

is called the *principal congruence subgroup* of Γ .

Note that $\Gamma(n) \subseteq \Gamma_0(n) \subseteq \Gamma$. In general, any analytic function that is invariant under a group of linear transformations is called an *automorphic function*.

Example 3. The j -invariant

$$j(\Delta) = \frac{1728g_2^3}{\Delta} = \frac{1}{q} + 744 + \sum_{n=1}^{\infty} c_n q^n,$$

where $z \in \mathfrak{H}$ and $q = \exp(2\pi iz)$ is an elliptic modular function.

The j -invariant is linked to elliptic curves in a natural way as follows.

Definition 7 [Weierstrass Equations for Elliptic Curves]. If F is a field of characteristic different from 2 or 3, where $g_2, g_3 \in F$, with $\Delta = g_2^3 - 27g_3^2 \neq 0$, then the elliptic curve over F of

$$y^2 = 4x^3 - g_2x - g_3 \tag{8}$$

denoted by $E(F)$ is the set of points (x, y) with $x, y \in F$ satisfying (8) together with a point \mathfrak{o} , called the point at infinity. Equation (8) is called the *Weierstrass equation for E* , and $\Delta(E(F)) = -16(4g_2^3 + 27g_3^2)$ is known as the *discriminant of $E(F)$* .

In order to give our first example of Weierstrass equations, we need the following concept.

Definition 8 [Lattices in \mathbb{C} and Elliptic Functions]. A *lattice in \mathbb{C}* is an additive subgroup of \mathbb{C} which is generated by two complex numbers ω_1 and ω_2 that are linearly independent over \mathbb{R} , denoted by $L = [\omega_1, \omega_2]$. Then an *elliptic*

function for L is a function f defined on \mathbb{C} , except for isolated singularities, satisfying the following two conditions:

- (a) $f(z)$ is meromorphic on \mathbb{C} .
- (b) $f(z + \omega) = f(z)$ for all $\omega \in L$.

Remark 3. Condition (b) in Definition 8 is equivalent to

$$f(z + \omega_1) = f(z + \omega_2) = f(z),$$

for all z , a property known as *doubly periodic*. Hence, an elliptic function for a lattice L is a doubly periodic meromorphic function and the elements of L are called *periods*.

Definition 9 [Lattice Discriminant and Invariant]. The *j-invariant of a lattice L* is the complex number

$$j(L) = \frac{1728g_2(L)^3}{g_2(L)^3 - 27g_3(L)^2}, \quad (9)$$

where

$$g_2(L) = 60 \sum_{w \in L - \{0\}} \frac{1}{w^4},$$

and

$$g_3(L) = 140 \sum_{w \in L - \{0\}} \frac{1}{w^6}.$$

The *discriminant of a lattice L* is given by

$$\Delta(L) = g_2(L)^3 - 27g_3(L)^2.$$

One of the most celebrated of elliptic functions is the following.

Definition 10 [Weierstrass \wp -functions]. Given $z \in \mathbb{C}$ such that $z \notin L = [\omega_1, \omega_2]$, the function

$$\wp(z; L) = \frac{1}{z^2} + \sum_{\omega \in L - \{0\}} \left(\frac{1}{(z - \omega)^2} - \frac{1}{\omega^2} \right) \quad (10)$$

is called the *Weierstrass \wp -function for the lattice L* .

Remark 4. The Weierstrass \wp -function is an elliptic function for L whose singularities can be shown to be double poles at the points of L . This is done by showing that $\wp(z)$ is holomorphic on $\mathbb{C} - L$ and has a double point at the origin. Then one may demonstrate that since

$$\wp'(z) = -2 \sum_{\omega \in L} \frac{1}{(z - \omega)^3},$$

which can be shown to converge absolutely, $\wp'(z)$ is an elliptic function for $L = [\omega_1, \omega_2]$. Since $\wp(z)$ and $\wp(z + \omega_j)$ have the same derivative, given that $\wp'(z)$ is periodic, they differ by a constant which can be shown to be zero by the fact that $\wp(z)$ is an even function. This demonstrates the periodicity of $\wp(z)$ from which it follows that the poles of $\wp(z)$ are double poles and lie in L .

Example 4. It can be shown that the *Laurent series expansion* (generally one of the form $\sum_{n=-\infty}^{\infty} a_n z^n$) for $\wp(z)$ about $z = 0$ is given by

$$\wp(z) = \frac{1}{z^2} + \sum_{n=1}^{\infty} (2n+1)G_{2n+1}(L)z^{2n}, \quad (11)$$

where for a lattice L , and an integer $r > 2$,

$$G_r(L) = \sum_{\omega \in L - \{0\}} \frac{1}{\omega^r}.$$

From this it follows that if $x = \wp(z; L)$ and $y = \wp'(z; L)$,

$$y^2 = 4x^3 - g_2(L)x - g_3(L), \quad (12)$$

where $g_j(L)$ for $j = 2, 3$ are given in Definition 9.

Remark 5. If E is an elliptic curve over \mathbb{C} given by the Weierstrass equation

$$y^2 = 4x^3 - g_2x - g_3,$$

with $g_1, g_2 \in \mathbb{C}$ and $g_2^3 - 27g_3^2 \neq 0$, then there is a unique lattice $L \subseteq \mathbb{C}$ such that

$$g_2(L) = g_2 \quad \text{and} \quad g_3(L) = g_3.$$

The j -invariant may be used with elliptic curves as follows.

Definition 11 [j -invariants for Elliptic Curves]. If E is an elliptic curve defined by the Weierstrass equation in Definition 7, then

$$j(E) = 1728 \frac{g_2^3}{g_2^3 - 27g_3^2} = 1728 \frac{g_2^3}{\Delta} \in F$$

is called the j -invariant of E .

In Definition 11, $\Delta \neq 0$ and $1728 = 2^6 \cdot 3^3$. Since we are not in characteristic 2 or 3, $j(E)$ is well-defined. If $F = \mathbb{C}$, then when E is the elliptic curve defined by the lattice $L \subseteq \mathbb{C}$,

$$j(L) = j(E). \quad (13)$$

Lastly, we need to know how to reduce points on elliptic curves.

Definition 12 [Reduction of Rationals on Elliptic Curves]. Let $n \in \mathbb{N}$ and $x_1, x_2 \in \mathbb{Q}$ with denominators prime to n . Then $x_1 \equiv x_2 \pmod{n}$ means $x_1 - x_2 = a/b$, where $\gcd(a, b) = 1$, $a, b \in \mathbb{Z}$, and $n|a$. For any $x = c/d \in \mathbb{Q}$ with $\gcd(d, n) = 1 = \gcd(c, d)$, there exists a unique $r \in \mathbb{Z}$, with $0 \leq r \leq n-1$, such that $x \equiv r \pmod{n}$, denoted by

$$r = \bar{x} \pmod{n}.$$

Note that we may take $r \equiv \overline{cd^{-1}} \pmod{n}$, where d^{-1} is the unique multiplicative inverse of d modulo n . Hence, if $P = (x, y)$ is a point on an elliptic curve $E = E(\mathbb{Q})$ over \mathbb{Q} , with denominators of x and y prime to n , then

$$\bar{P} \pmod{n} \text{ means } (\bar{x} \pmod{n}, \bar{y} \pmod{n}).$$

Also, $\bar{E} \pmod{n}$ denotes the curve reduced modulo n , namely the curve defined by $y^2 = x^3 + \bar{a} \pmod{n} x + \bar{b} \pmod{n}$, with $x = \bar{x} \pmod{n}$, and $y = \bar{y} \pmod{n}$. The cardinality of the set $\bar{E} \pmod{n}$ is denoted by $|\bar{E} \pmod{n}|$.

Now, we use the above to paint the picture that will bring the STW conjecture into focus.

3. STW and FLT

In general, an elliptic curve E defined over a field F may be given by the *global* Weierstrass equation

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6, \quad (14)$$

where $a_j \in F$ for $1 \leq j \leq 6$. Then when F has characteristic different from 2, we may complete the square, replacing y by $(y - a_1x - a_3)/2$ to get the more familiar Weierstrass equation

$$y^2 = 4x^3 + b_2x^2 + 2b_4x + b_6 \quad (15)$$

with

$$b_2 = a_1^2 + 4a_2,$$

$$b_4 = 2a_4 + a_1a_3,$$

and

$$b_6 = a_3^2 + 4a_6.$$

In this case, the discriminant $\Delta(E) = \Delta$ is given by

$$\Delta(E) = -b_2^2b_8 - 8b_4^3 - 27b_6^2 + 9b_2b_4b_6, \quad (16)$$

where

$$b_8 = a_1^2a_6 + 4a_2a_6 - a_1a_3a_4 + a_2a_3^2 - a_4^2.$$

Also, the j -invariant is given by

$$j(E) = c_4^3/\Delta(E), \quad (17)$$

where

$$c_4 = b_2^2 - 24b_4 \quad (18)$$

and

$$j(E) = 1728 + c_6^2/\Delta, \quad (19)$$

where

$$c_6 = -b_2^3 + 36b_2b_4 - 216b_6. \quad (20)$$

We may further simplify equation (15) by replacing (x, y) with $((x - 3b_2)/36, y/108)$ to achieve

$$y^2 = x^3 - 27c_4x + 54c_6. \quad (21)$$

It can be shown that

$$\Delta(E) = \frac{c_4^3 - c_6^2}{1728}. \quad (22)$$

Remark 6. Note, however, that if we begin with equation (21), then the discriminant is

$$\Delta(E) = 2^6 \cdot 3^9(c_4^3 - c_6^2),$$

which differs from (22) by a factor of $2^{12} \cdot 3^{12}$, and this is explained by the scaling introduced in change of variables in going from (14) to (15), then to (21).

Remark 6 shows that a change of variables may “inflate” a discriminant with new factors. Thus, for our development, we need to find a “minimal discriminant”. In order to proceed with this in mind, we need the following concept.

Definition 13 [Admissible Change of Variables]. If $E = E(\mathbb{Q})$ is an elliptic curve over \mathbb{Q} , given by (14) where we may assume that $a_j \in \mathbb{Z}$ for $j = 1, 2, 3, 4, 6$, then an *admissible* change of variables is one of the form

$$x = u^2X + r \quad \text{and} \quad y = u^3Y + su^2X + t,$$

where $u, r, s, t \in \mathbb{Q}$ and $u \neq 0$ with resulting equation

$$Y^2 + a'_1XY + a'_3Y = X^3 + a'_2X^2 + a'_4X + a'_6, \quad (23)$$

where

$$a'_1 = \frac{a_1 + 2s}{u}, \quad a'_2 = \frac{a_2 - sa_1 + 3r - s^2}{u^2},$$

$$a'_3 = \frac{a_3 + ra_1 + 2t}{u^3}, \quad a'_4 = \frac{a_4 - sa_3 + 2ra_2 - (t + rs)a_1 + 3r^2 - 2st}{u^4},$$

and

$$a'_6 = \frac{a_6 + ra_4 + r^2a_2 + r^3 - ta_3 - t^2 - rta_1}{u^6}.$$

Remark 7. In the special case where $r = s = t = 0$, the admissible change of variables multiplies the a_i by u^{-i} for $i = 1, 2, 3, 4, 6$. In this case, we say a_i has *weight* i . It can be shown that two elliptic curves over \mathbb{Q} are related by an admissible change of variables if and only if they have the same j -invariant. There is another term in the literature used to describe this phenomenon as well. *Two elliptic curves over \mathbb{Q} having the same j -invariant are said to be twists of one another.*

Since the discriminant Δ is given by (22) in terms of c_4 and c_6 , Δ is unaffected by r, s, t in an admissible change of variables given that the new variables for (23) are related by $c'_4 = c_4/u^4$ and $c'_6 = c_6/u^6$. Hence, the triple (Δ, c_4, c_6) is a detector for curves that are equivalent under an admissible change of variables. In fact, by the above discussion, two elliptic curves E_1 and E_2 with discriminants Δ_1 and Δ_2 , respectively, related by an admissible change of variables, must satisfy $\Delta_1/\Delta_2 = u^{\pm 12}$. This now sets the stage for looking at elliptic curves with minimal discriminants.

For the ensuing development, the notation of Definition 13 remains in force.

Definition 14 [Minimal Equations for Elliptic Curves]. If $E = E(\mathbb{Q})$ is an elliptic curve over \mathbb{Q} , given by (14) where $a_j \in \mathbb{Z}$ for $j = 1, 2, 3, 4, 6$ with discriminant Δ , then (14) is called *minimal* at the prime p if the power of p dividing Δ cannot be decreased by making an admissible change of variables with the property that the new coefficients $a'_j \in \mathcal{O}_p$, the p -adic integers. If (14) is minimal for all primes p with $a_j \in \mathbb{Z}$ for $j = 1, 2, 3, 4, 6$, then it is called a *global minimal Weierstrass equation*.

Remark 8. Since an equation for $E(\mathbb{Q})$ given in Definition 14, can be assumed, without loss of generality, to have integral coefficients, $|\Delta|_p \leq 1$, where $|\cdot|_p$ is the p -adic absolute value. Hence, in only finitely many steps $|\Delta|_p$ can be increased and still maintain $|\Delta|_p \leq 1$. Hence, it follows that in finitely many admissible changes of variables, we can get an equation minimal for E at p . In other words, there always exists a global minimal Weierstrass equation for $E(\mathbb{Q})$.

For the following, we define the following, where χ is a quadratic Dirichlet

character, meaning that $\chi(y) = -1, 0, 1$ according as y is a quadratic nonresidue, 0, or a quadratic residue, respectively, for $y \in \mathbb{F}_p$.

$$N_p = p + 1 + \sum_{x \in \mathbb{F}_p} \chi(x^3 + ax + b), \quad (24)$$

being the number of points on the elliptic curve $E(\mathbb{F}_p)$, including the point at infinity, over a field of p elements for a prime p .

Definition 15 [The Reduction Index for Elliptic Curves]. Suppose that E is an elliptic curve over \mathbb{Q} given by a minimal Weierstrass equation. If the $\bar{E}(\text{mod } p) \neq 0$ for a prime p , then p is said to be a prime of *good reduction* for E . Furthermore, if N_p for a prime p is given by (24), then let

$$a_p(E) = p + 1 - N_p.$$

If p is a prime of good reduction, then $a_p(E)$ is called the *good reduction index* for E at p , and the sequence $\{a_p(E)\}_p$ indexed over the primes of good reduction is called the *good reduction sequence* for E . Primes that are *not* of good reduction are called primes of *bad reduction* for E , and $a_p(E)$ is called the *bad reduction index* for E .

Note that there are only finitely many primes of bad reduction since these are the primes dividing Δ .

Example 5. Consider the elliptic curve given by $y^2 + y = x^3 - x^2$. Via the formulas in (14)-(22), we have $a_1 = 0$, $a_3 = 1$, $a_2 = -1$, $a_4 = 0 = a_6$, $b_2 = -4$, $b_4 = 0$, $b_6 = 1$, and $b_8 = -1$. Therefore,

$$\begin{aligned} \Delta(E) &= -b_2^2 b_8 - 8b_4^3 - 27b_6^2 + 9b_2 b_4 b_6 \\ &= -(-4)^2(-1) - 8(0)^3 - 27 \cdot 1^2 + 9(-4)(0)(1) = -11, \end{aligned}$$

so E has good reduction at all primes $p \neq 11$. Now we compute the good reduction index for this curve at various primes $p \neq 11$, which we call a *good reduction table* for E .

p	2	3	5	7	11	13	17	19	23	29	31	37	41
N_p	5	5	5	10	11	10	20	20	25	30	25	35	50
$a_p(E)$	-2	-1	1	-2	1	4	-2	0	-1	0	7	3	-8

Remark 9. To say that p is a prime of good reduction for E is to say that E is nonsingular over \mathbb{F}_p , meaning that $\Delta(\overline{E}(\text{mod } p))$ is not divisible by p . We now explain this in detail. A point $P = (x_0, y_0)$ on an elliptic $E(F) = E$ curve over a field F is called a *singular point* if P satisfies the equation, defining E , given by

$$f(x, y) = y^2 + a_1xy + a_3y - x^3 - a_2x^2 - a_4x - a_6 = 0 \tag{25}$$

with the partial derivatives satisfying

$$\partial f / \partial x(P) = \partial f / \partial y(P) = 0.$$

Thus, to say that P is a singular point of E is to say that E is a singular curve at P . To say that E is nonsingular over F is to say that the curve has *no singular points*. It can be shown that E is nonsingular if and only if $\Delta(E) \neq 0$. Note that E never has a singular point at infinity.

Remark 10. The good reduction index is a mechanism for representing arithmetic data about E that is captured in patterns of the good reduction sequence $\{a_p(E)\}_p$. How it does this is contained in the subtext of the Shimura-Taniyama-Weil conjecture. The pattern involves the normalized modular cusp forms of weight 2 and level $n \in \mathbb{N}$ that we introduced in Definition 5.

Definition 16 [Modular Elliptic Curves]. Let $E(\mathbb{Q})$ be an elliptic curve over \mathbb{Q} with good reduction sequence $\{a_p(E)\}_p$. If there exist an $n \in \mathbb{N}$ and a normalized weight 2 cusp form of level n ,

$$f(z) = q + \sum_{j=2}^{\infty} a_j(f)q^j, \text{ where } q = \exp(2\pi iz),$$

such that

$$a_p(E) = a_p(f),$$

then E is called a *modular elliptic curve*.

Now we may state the celebrated conjecture.

Conjecture 1 [The Shimura-Taniyama-Weil (STW) Conjecture]. If E is an elliptic curve over \mathbb{Q} , then E is modular.

Example 6. By Example 2, the function given in (6) spans $S_2(\Gamma_0(11))$ and is explicitly given by

$$\begin{aligned} f(z) &= \eta(z)^2 \eta(11z)^2 = \sum_{n=1}^{\infty} c_n q^n = q \prod_{n=1}^{\infty} (1 - q^n)^2 \cdot (1 - q^{11n})^2 \\ &= q - \mathbf{2q}^2 - \mathbf{q}^3 + 2q^4 + \mathbf{q}^5 + 2q^6 - \mathbf{2q}^7 - 2q^9 - 2q^{10} + \mathbf{q}^{11} - 2q^{12} + \mathbf{4q}^{13} \\ &\quad + 4q^{14} - q^{15} - 4q^{16} - \mathbf{2q}^{17} + 4q^{18} + 2q^{20} + 2q^{21} - 2q^{22} - \mathbf{q}^{23} - 4q^{25} \\ &\quad - 8q^{26} + 5q^{27} - 4q^{28} + 2q^{30} + \mathbf{7q}^{31} + \dots + \mathbf{3q}^{37} + \dots - \mathbf{8q}^{41} + \dots \end{aligned}$$

We have highlighted the prime powers of q and their coefficients to show that these coefficients are exactly the nonzero values of the good reduction index $a_p(E)$ in Example 5, thereby illustrating that E is a modular function.

Remark 11. The notion of a *conductor* of an elliptic curve must now come into play for our discussion. The technical definition involves a cohomological description that we do not have the tools to describe. However, we can talk about it in reference to the discriminant and related prime divisors in order to understand what it means. Given an elliptic curve $E(\mathbb{Q}) = E$ with global minimal Weierstrass equation and discriminant $\Delta(E) = \Delta$, the conductor n divides Δ and has the same prime factors as Δ . The power to which a given prime appears in n is determined as follows. The power of a prime p dividing n is 1 if and only if $E(\mathbb{F}_p)$ has a *node*, which is characterized by having two candidate tangents at the point, which in turn, means that (23) has a double root. If $p > 3$, then the power of p dividing n is 2 if and only if $E(\mathbb{F}_p)$ has a cusp. In the case where $p = 2$ or $p = 3$, which we selectively have ignored for the sake of simplicity of presentation, the conductor can be computed using Tate's algorithm, which is uncomplicated, although the process of using it can be somewhat protracted, see [8]. For $p \neq 2, 3$, the power of p dividing the conductor n is at most 2, so for our purposes, the above discussion suffices.

From the above, we conclude that the conductor of E is not divisible by any primes of good reduction, also called *stable* reduction. In other words, only primes of bad reduction divide the conductor. Moreover, a prime p to the first power exactly divides the conductor precisely when $E(\mathbb{F}_p)$ has a node, in which case E is said to have *multiplicative* or *semi-stable* reduction at p . Hence, E has reduction at *all* primes, in which case E is called semi-stable, precisely when the conductor n is squarefree. For instance, the curve in Example 5 has conductor 11, an instance of a semi-stable elliptic curve. The conductor of E is exactly divisible by p^2 precisely when $E(\mathbb{F}_p)$ has a cusp, in which case we say that E has *additive* or *unstable* reduction.

It may be shown that the conductor is an *isogeny invariant*, which means the following. An *isogeny* between two elliptic curves E_1 and E_2 is an analytic map $h : E_1 \mapsto E_2$, where the identity gets mapped to the identity. Two curves are *isogenous* if there is a nonconstant isogeny h between them. Hence, for the conductor to be an isogeny invariant means that the conductor of isogenous curves remains the same.

The STW conjecture implies that we have the conductor n equal to the level n in $\Gamma_0(n)$ of weight 2 cusp forms, see the reformulation of STW in terms of L -functions below.

Now we illustrate the modularity theorem in different terms that will bring more of the structure and interconnections to light. To do this, we concentrate upon the example $n = 11$, which will be a template for the general theory.

Example 7. From Example 2, for $n = 11$, the group $\Gamma_0(11)$ can be shown to be generated by

$$T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 8 & 1 \\ -33 & -4 \end{pmatrix}, \quad V = \begin{pmatrix} 9 & 1 \\ -55 & -6 \end{pmatrix},$$

and if $\gamma \in S_2(\Gamma_0(11))$, then we map $\Gamma_0(11)$ to \mathbb{C} , additively via $\phi_\gamma(U) = \omega_1$, $\phi_\gamma(V) = \omega_2$, and $\phi_\gamma(T) = 0$. Hence, $L = [\omega_1, \omega_2]$ is a lattice in \mathbb{C} . It can be shown that \mathbb{C}/L , called a *complex torus*, is analytically isomorphic to an elliptic curve $E(\mathbb{C})$, where L is determined by E up to what is known as *homothety*, which

means that if L_1 is another lattice determining E , then $L = \lambda L_1$ for some $\lambda \in \mathbb{C}$. For our purposes the “analytic isomorphism”



$$\mathbb{C}/L \mapsto E(\mathbb{C})$$

is explicitly given by

$$z \mapsto \begin{cases} (\wp(z), \wp'(z), 1) & \text{if } z \notin L, \\ (0, 1, 0) & \text{if } z \in L, \end{cases}$$

see Remark 7. This is a holomorphic map carrying \mathbb{C}/L one-to-one onto the elliptic curve $E = E(\mathbb{C})$, where E is given by the form

$$y^2 = 4x^3 - g_2x - g_3,$$


with g_2 and g_3 given in Definition 9.  Together, we get a holomorphic map from $X_0(11)$ onto \mathbb{C}/L , then onto $E(\mathbb{C})$.  Thus, it can be shown that this provides a holomorphic surjection

$$X_0(11) = \frac{\Gamma_0(11)}{\mathfrak{H}^*} \mapsto E(\mathbb{C}), \text{ where } \mathfrak{H}^* = \mathfrak{h} \cup \mathbb{Q} \cup \{\infty\},$$

where $X_0(11)$ is called a *compact Riemann surface*, which is a complex one-dimensional manifold. \mathbb{C}/L is also a complex manifold and the principal feature of such surfaces is that holomorphic maps can be defined between them as we have done above, see [7] for more details.

One may actually calculate the j -invariant via (9) to get

$$j(L) = -\frac{(2^4 \cdot 31)^3}{11^5}, \quad (26)$$

which demonstrates that E is defined over \mathbb{Q} and gives more meaning to the above mapping involving $X_0(11)$ and E over \mathbb{Q} . However, from (1)  we have

$$j = \frac{c_4^3}{\Delta} = 1728 + \frac{c_6^2}{\Delta}. \quad (27)$$

It can be shown that there is an integer $k \neq 0$ such that

$$c_4 = 2^4 \cdot 31k^2, \quad c_6 = 2^3 \cdot 2501k^3, \quad \text{and } \Delta = -11^5 k^6. \quad (28)$$

It follows that (28) yields a global minimal Weierstrass equation exactly when k has no odd square factor, and

$$k \equiv r \pmod{16}, \text{ where } r \in \{1, 2, 5, 6, 9, 10, 12, 13, 14\}. \quad (29)$$

We call the association of $X_0(11)$ and $E = E(\mathbb{Q})$ given by (29), with global minimal Weierstrass equation provided by (28), a \mathbb{Q} -structure of E . The simplest \mathbb{Q} -structure occurs when $k = 1$ in which case we get the global minimal equation is given by

$$E(\mathbb{C}) : y^2 + y = x^3 - x^2 - 10x - 20. \quad (30)$$

What we have accomplished is a mapping of $X_0(11)$ onto $E(\mathbb{C})$.

Now, if we define

$$\begin{pmatrix} \omega'_2 \\ \omega'_1 \end{pmatrix} = \begin{pmatrix} 1 & 3 \\ 0 & 5 \end{pmatrix} \begin{pmatrix} \omega_2 \\ \omega_1 \end{pmatrix}$$

and we let $L' = [\omega'_1, \omega'_2]$, it can be shown that $j(L') = -16^3/11$, so a corresponding elliptic curve E' can be defined over \mathbb{Q} , and this curve is given by

$$E' : y^2 + y = x^3 - x^2, \quad (31)$$

which is the curve in Example 5, with discriminant -11 , and as we saw above the discriminant of (30) is -11^5 . In Remark 11, we saw that the conductor is an isogeny invariant, in this case $n = 11$.

We may reformulate the STW conjecture now in terms of the above, which we have illustrated for the case $n = 11$.

◆ STW Conjecture in Terms of Modular Parametrizations

Given an elliptic curve E over \mathbb{Q} , there exists an $n \in \mathbb{N}$ for which there is a nonconstant surjective holomorphic map $F : X_0(n) \mapsto E$, defined over \mathbb{Q} , in which case E is said to have a modular parametrization modulo n , and E is called a *Weil curve*.

Remark 12. We have illustrated the above for the case $n = 11$ in Example 7, but the theory, called *Eichler-Shimura theory*, holds for any of the compact Riemann

surfaces $X_0(n)$, where n is the level of the weight 2 cusp forms, so given the aforementioned proof of STW, the above is a statement of the *modularity theorem*.

The phrase “defined over \mathbb{Q} ” in the above interpretation of the STW conjecture is important in that we may have holomorphic surjections without the rationality property but for which the L -functions of the curves and the cusp forms do not agree. Now we must explain this comment by introducing the notions of L -functions for elliptic curves and forms. Note that the construction of the map from $X_0(11)$ to $E(\mathbb{C})$ in Example 7 is indeed defined over \mathbb{Q} . In the literature, such maps are rational maps defined at every point, called *morphisms*, see [7].

We turn our attention to L -functions. Elliptic curves that are isogenous over \mathbb{Q} have the same L -functions which we now define and discuss.

Let $E(\mathbb{Q})$ be an elliptic curve over \mathbb{Q} given by a global minimal Weierstrass equation, which is no loss of generality by Remark 8. Then the L -function for E , having discriminant Δ is given by

$$L(E, s) = \prod_{p|\Delta} [(1 - a_p(E)p^{-s})^{-1}] \prod_{p \nmid \Delta} [(1 - a_p(E)p^{-s} + p^{1-2s})^{-1}].$$

It can be shown that $L(E, s)$ converges for $\Re(s) > 2$, and is given by an absolutely convergent Dirichlet series. Thus, we may write

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

Now by Definition 5, a normalized cusp form $f \in S_2(\Gamma_0(n))$ of weight 2 and level n satisfies

$$f(z) = q + \sum_{n=2}^{\infty} a_n(f)q^n.$$

Thus, we may define the L -function of f by

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

Now the STW conjecture may be reformulated in terms of L functions:

◆ **STW Conjecture in Terms of L -functions**

For every elliptic curve E defined over \mathbb{Q} , there exists a normalized cusp form of weight 2 and level n , $f \in S_2(\Gamma_0(n))$, such that

$$L(f, s) = L(E, s),$$

and n is the conductor of E .

We have concentrated upon $X_0(11)$ in Example 7 since it is the simplest case, namely having what is called *genus one* with corresponding $S_2(\Gamma_0(11))$ having dimension one as we have seen above. In general, the dimension of $S_2(\Gamma_0(n))$ is called the *genus* of $X_0(n)$. To see the intimate connection with FLT, we return to the discussion of Frey curves (1). Suppose that

$$a^p + b^p = c^p \tag{32}$$

is a counterexample to FLT for a prime $p \geq 5$. The Frey curve is given by

$$E : y^2 = x(x - a^p)(x - c^p), \tag{33}$$

for which

$$\Delta = 16a^{2p}b^{2p}c^{2p}, \tag{34}$$

and

$$c_4 = 16(a^{2p} - a^p c^p + c^{2p}). \tag{35}$$

Then when a, b, c are pairwise relatively prime, it can be shown that the conductor of E is the product of all primes dividing abc , which tells us, by Remark 11, that E is semi-stable.

Now we are in a position to return to a discussion of the STW conjecture and FLT. In 1995, Taylor and Wiles published papers [9] and [10], which proved that every semi-stable elliptic curve is modular. In 1999, Conrad et al. [2] proved the STW conjecture for all elliptic curves with conductor not divisible by 27. Then in 2001, Breuil et al. published a proof of the full STW conjecture, which we now call the *modularity theorem* [1]. However, in 1990, Ribet proved the following, which via the affirmative verification of the STW conjecture, allowed a proof of FLT as follows.

Theorem 2 [Ribet's Theorem]. *Suppose that E is an elliptic curve over \mathbb{Q} given by a global minimal Weierstrass equation and having discriminant $\Delta = \prod_{p|\Delta} p^{f_p}$ and conductor $n = \prod_{p|\Delta} p^{g_p}$, both canonical prime factorizations. Furthermore, if E has a modular parametrization of level n with $f \in S_2(\Gamma_0(n))$ having normalized expansion*

$$f(z) = q + \sum_{n=2}^{\infty} a_j(f)q^n,$$

and for a fixed prime p_0 , set

$$n' = \frac{n}{\prod_{\substack{p \\ p_0 \nmid f_p \\ g_p=1}} p}. \quad (36)$$

Then there exists an $f' \in S_2(\Gamma_0(n'))$ such that $f' = \sum_{n=1}^{\infty} b_j(f')q^n$ with $b_j(f') \in \mathbb{Z}$ satisfying $a_j(f) \equiv b_j(f') \pmod{p_0}$ for all $n \in \mathbb{N}$.

Proof. See [6]. □

Now we may state our target result, which follows [4, Corollary 12.13, p. 399], where it is cited as a Frey-Serre-Ribet result.

Theorem 3 [Proof of Fermat's Last Theorem]. *The STW conjecture implies FLT.*

Proof. Assume that FLT is false. Then by Theorem 2, the Frey curve given in (33) has conductor $n = \prod_{p|abc} p$, which when compared to the coefficients in (36) yields $n' = 2$. However, by Example 2, $S_2(\Gamma_0(2))$ is the zero space, so $b_j(f') = 0$ for all $n \in \mathbb{N}$. Yet, $b_j(f') \equiv a_j(f) \pmod{p_0}$ for all $n \in \mathbb{N}$. In particular, $0 = b_1(f') \equiv a_1(f) = 1 \pmod{p_0}$, a contradiction. □

References

- [1] C. Breuil, B. Conrad, F. Diamond and R. Taylor, On the modularity of elliptic curves over \mathbf{Q} : wild 3-adic exercises, J. Amer. Math. Soc. 14 (2001), 843-939.

- [2] B. Conrad, F. Diamond and R. Taylor, Modularity of certain potentially Barsotti-Tate Galois representations, *J. Amer. Math. Soc.* 12 (1999), 521-567.
- [3] G. Frey, Links between stable elliptic curves and certain Diophantine equations, *Ann. Univ. Sarav. Ser. Math.* 1 (1986), 1-40.
- [4] A. W. Knap, *Elliptic curves*, Math. Notes 40, Princeton University Press, Princeton, N.J., 1992.
- [5] R. A. Mollin, *Advanced Number Theory with Applications*, CRC, Taylor and Francis Group, Boca Raton, London, New York, 2009.
- [6] K. A. Ribet, On modular representations of $\text{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ arising from modular forms, *Invent. Math.* 100 (1990), 431-476.
- [7] J. H. Silverman, *The Arithmetic of Elliptic Curves*, Springer, New York, Berlin, Heidelberg, 1985.
- [8] J. Tate, Algorithm for determining the type of a singular fiber in an elliptic pencil, *Modular Functions of One Variable, IV*, Lecture Notes in Math., Vol. 476, Springer, Berlin, 1975, pp. 33-52.
- [9] R. Taylor and A. Wiles, Ring-theoretic properties of certain Hecke algebras, *Ann. of Math. (2)* 141 (1995), 553-572.
- [10] A. Wiles, Modular elliptic curves and Fermat's last theorem, *Ann. of Math. (2)* 141 (1995), 443-551.