R. van Dobben de Bruyn

The Modularity Theorem

Bachelor's thesis, June 21, 2011

Supervisors: Dr R.M. van Luijk, Dr C. Salgado

Mathematisch Instituut, Universiteit Leiden

Contents

Introduction

One of the longest standing open problems in mathematics was Fermat's Last Theorem, asserting that the equation $a^n + b^n = c^n$ does not have any nontrivial (i.e. with $abc \neq 0$) integral solutions when n is larger than 2. The proof, which was completed in 1995 by Wiles and Taylor, relied heavily on the Modularity Theorem, relating elliptic curves over $\mathbb{Q}$ to modular forms.

The Modularity Theorem has many different forms, some of which are stated in an analytic way using Riemann surfaces, while others are stated in a more algebraic way, using for instance L-series or Galois representations. This text will present an elegant, elementary formulation of the theorem, using nothing more than some basic vocabulary of both elliptic curves and modular forms.

For elliptic curves E over $\mathbb{Q}$, we will examine the reduction $\tilde{E}$ of E modulo any prime p, thus introducing the quantity

$$a_p(E) = p + 1 - \#\tilde{E}(\mathbb{F}_p).$$

We will also give an almost complete description of the conductor N_E associated to an elliptic curve E, and compute it for the curve used in the proof of Fermat's Last Theorem.

As for modular forms, we will mostly cover the basic definitions and examples, thus introducing the Fourier series

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

of a given modular form f. Furthermore, we will examine modular forms with respect to certain groups called congruence subgroups, and we take a closer look at a group that is denoted $\Gamma_0(N)$.

Having all the vocabulary in place, we arrive at the Modularity Theorem, which asserts that given an elliptic curve E with conductor N_E, there exists a modular form f with respect to $\Gamma_0(N_E)$, such that

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

holds for all primes p.

We cannot prove the Modularity Theorem in this text, but we will give a short sketch of how Fermat's Last Theorem follows from it.

1 Elliptic Curves

In this chapter, K will denote a perfect field, with an algebraic closure $\bar{K}$, and Galois group $G = G_{\bar{K}/K}$. The reader who is not familiar with algebraic geometry is invited to read for instance the first chapter of [6] or the first two chapters of [9]. For a more accessible (but also longer) introduction, one might read [4].

Throughout this text, a curve means an irreducible affine or projective variety of dimension 1 over the algebraically closed field $\bar{K}$. We say that the curve C is *defined over* K if its (homogeneous) ideal can be generated by (homogeneous) elements over K. Furthermore, if we write $\mathbb{A}^n$ or $\mathbb{P}^n$, this is understood to be the affine or projective space over $\bar{K}$. Of course, $\mathbb{A}^n$ and $\mathbb{P}^n$ are defined over K.

1.1 Definitions and Examples

We will firstly give an abstract definition of an elliptic curve, followed by the more comprehensible definition of a Weierstrass curve. It turns out that every elliptic curve is isomorphic to a Weierstrass curve.

Definition 1.1.1. An *elliptic curve* over $\bar{K}$ is a pair (E, O), where E is a nonsingular projective curve of genus 1 over $\bar{K}$, and O is a point on E. If E is defined over K and O is a K-rational point, then (E, O) is *defined over* K.

We will mostly just write E for the elliptic curve (E, O).

Definition 1.1.2. A *Weierstrass polynomial* over K is a polynomial of the form

$$Y^2Z + a_1XYZ + a_3YZ^2 - X^3 - a_2X^2Z - a_4XZ^2 - a_6Z^3,$$

with $a_1, a_2, a_3, a_4, a_6 \in K$. The associated curve in $\mathbb{P}^2$ is called a *Weierstrass curve*. Note that such a curve is defined over K.

Remark 1.1.3. The polynomial

$$F = Y^2Z + a_1XYZ + a_3YZ^2 - X^3 - a_2X^2Z - a_4XZ^2 - a_6Z^3$$

is indeed irreducible, which is necessary for the associated variety to be a curve. This can for instance be seen by viewing it as a polynomial in $K(Y, Z)[X]$:

$$F = -X^3 - a_2ZX^2 + (a_1Y - a_4Z)ZX + (Y^2 + a_3YZ - a_6Z^2)Z.$$

This polynomial is Eisenstein by Z, hence irreducible.

Remark 1.1.4. We will usually write the dehomogenized equation

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6,$$

and understand that we will always wish to consider the projective curve given by the homogeneous polynomial

$$Y^2Z + a_1XYZ + a_3YZ^2 - X^3 - a_2X^2Z - a_4XZ^2 - a_6Z^3.$$

At the end of this section, we will find a criterion for a Weierstrass curve to be nonsingular. For now, we make the following observations.

Proposition 1.1.5. *Let E be a nonsingular Weierstrass curve. Then the point $O = [0 : 1 : 0]$ lies on E, and the pair (E, O) is an elliptic curve.*

Proof. By computation we see that $O \in E$. We only need to show that the genus of E is equal to 1. This is due to the fact that any projective curve in $\mathbb{P}^2$ given by some irreducible homogeneous polynomial of degree n has genus equal to $\frac{(n-1)(n-2)}{2}$ (see Exercise 8.6.6 of [4]). We use the case $n = 3$. □

Proposition 1.1.6. *Let E be an elliptic curve over K. Then there exists a Weierstrass curve C over K and an isomorphism $\phi\colon E \to C$ satisfying $\phi(O) = [0 : 1 : 0]$.*

Proof. See [9, Prop. III.3.1]. □

We will, by abuse of language, call a nonsingular Weierstrass curve an elliptic curve. Henceforth we will mostly work with Weierstrass curves (not necessarily smooth).

Definition 1.1.7. Let C be the (possibly singular) Weierstrass curve given by the equation

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6.$$

Then we define the following quantities associated to C:

$$\begin{aligned}
b_2 &= 4a_2 + a_1^2,\\
b_4 &= 2a_4 + a_1a_3,\\
b_6 &= 4a_6 + a_3^2,\\
b_8 &= a_1a_6 + 4a_2a_6 - a_1a_3a_4 + a_2a_3^2 - a_4^2,\\
c_4 &= b_2^2 - 24b_4,\\
c_6 &= -b_2^3 + 36b_2b_4 - 216b_6.
\end{aligned}$$

Remark 1.1.8. If $\mathrm{char}(K) \neq 2$, then the map $\mathbb{P}^2 \to \mathbb{P}^2$ induced by

$$(x, y) \mapsto (x, 2y + a_1x + a_3)$$

maps C isomorphically to the curve given by the equation

$$y^2 = 4x^3 + b_2x^2 + 2b_4x + b_6.$$

If furthermore $\mathrm{char}(K) \neq 3$, then the map induced by

$$(x, y) \mapsto (36x + 3b_2, 108y)$$

maps this last curve isomorphically to the curve given by the short Weierstrass equation:

$$y^2 = x^3 - 27c_4x - 54c_6.$$

The associated Weierstrass polynomial

$$Y^2Z - X^3 + 27c_4XZ^2 + 54c_6Z^3$$

is called a *short Weierstrass polynomial.*

Definition 1.1.9. Let C be a Weierstrass curve over K.

(a) The *discriminant* is given by

$$\Delta = -b_2^2 b_8 - 8b_4^3 - 27b_6^2 + 9b_2 b_4 b_6.$$

(b) The *j-invariant* is given by

$$j = \frac{c_4^3}{\Delta}.$$

Observe that $1728\Delta = 2^6 \cdot 3^3 \Delta = c_4^3 - c_6^2$. Furthermore, the discriminant is closely linked to the question whether or not the curve is nonsingular. We have the following proposition.

Proposition 1.1.10. *Let C be a Weierstrass curve. Then C is nonsingular if and only if $\Delta \neq 0$.*

Proof. Observe that

$$\frac{\partial F}{\partial Z}([0:1:0]) \neq 0,$$

where $F = Y^2Z + a_1XYZ + a_3YZ^2 - X^3 - a_2X^2Z - a_4XZ^2 - a_6Z^3$. Hence, C is smooth at infinity.

We will only finish the proof for $\mathrm{char}(K) \neq 2$. For the remaining case, see Appendix A of [9].

If $\mathrm{char}(K) \neq 2$, we observe that C is isomorphic to the curve C' in $\mathbb{P}^2$ given by

$$C' : \ y^2 = 4x^3 + b_2x^2 + 2b_4x + b_6. \tag{1}$$

Since isomorphisms of curves map singular points to singular points, we find that C is nonsingular if and only if C' is.

Now write $f(x)$ for the right-hand side of (1), and observe that for C' to be singular at (x, y) we need $2y = f'(x) = 0$, which occurs exactly when $y = 0$ and $\Delta(f) = 0$. The result follows since $\Delta(f) = 16\Delta$. □

Finally, we will distinguish some different kinds of singularities. In order to do so, we must firstly define the multiplicity of a point on a curve. We will do this only for curves in $\mathbb{P}^2$.

Definition 1.1.11. Let $C \subseteq \mathbb{A}^2$ be the curve given by the equation $f = 0$, for some irreducible $f \in \bar{K}[x, y]$ of degree d. Write

$$f = f_0 + \ldots + f_d,$$

where f_i is a homogeneous polynomial of degree i for all $i \in \{0, \ldots, d\}$. Then the *multiplicity of* $(0, 0)$ *on* C is the quantity

$$\mu_{(0,0)}(C) := \inf\{i \in \{0, \ldots, d\} : f_i \neq 0\}.$$

Remark 1.1.12. Observe that $P = (0,0)$ is on C if and only if $\mu_P(C) > 0$, and P is a singular point on C if and only if $\mu_P(C) > 1$.

Example 1.1.13. Let C be the singular Weierstrass curve given by $y^2 = x^3 + x^2$. Then:

$$f = x^3 + x^2 - y^2;$$
$$f_0 = 0,\ f_1 = 0,\ f_2 = x^2 - y^2,\ f_3 = x^3,$$

so that $\mu_{(0,0)}(C) = 2$.

Definition 1.1.14. Let $C \subseteq \mathbb{P}^2$ be a curve, and $P \in \mathbb{P}^2$ a point. We define the *multiplicity of P on C* as follows: make a linear change of coordinates such that P becomes $[0:0:1]$, and let $C' \subseteq \mathbb{A}^2$ the curve given by $C \cap U_2$, where

$$U_2 = \{[x:y:z] \in \mathbb{P}^2 : z \neq 0\} \cong \mathbb{A}^2.$$

Then we put

$$\mu_P(C) = \mu_{(0,0)}(C').$$

Remark 1.1.15. If C is a Weierstrass curve, then we know from the proof of Proposition 1.1.10 that the point $[0:1:0]$ at infinity is nonsingular. Hence any singular point is found on the affine part of C. The map given by

$$(x, y) \mapsto (x - a, y - b)$$

sends (a, b) to $(0, 0)$ and maps C isomorphically to the curve given by the equation

$$(y+b)^2 + a_1(x+a)(y+b) + a_3(y+b) = (x+a)^3 + a_2(x+a)^2 + a_4(x+a) + a_6.$$

In particular, there will always be a term in y^2, so that the homogeneous part f_2 of degree 2 will always be nonzero. Hence,

$$\mu_{(a,b)}(C) \leq 2,$$

so that the 'worst' singularity that can occur is a double point. In particular, any singular point on C will be a double point.

Remark 1.1.16. In fact, there can be at most one singular point, for if P, Q are two singular points, the unique line in $\mathbb{P}^2$ through P and Q will intersect C with multiplicity at least 4. This is impossible by Bézout's Theorem.

Next, we will see which types of double points can occur on a curve. We will once again start with the point $(0, 0)$, and we let the reader make the necessary modifications to apply this definition for arbitrary points.

Definition 1.1.17. Let C be the curve in $\mathbb{A}^2$ given by an irreducible polynomial $f \in \bar{K}[x, y]$. Let f_i be the homogeneous part of degree i, and assume that f has a double point in $(0, 0)$, i.e. that $f_0 = f_1 = 0 \neq f_2$. Then we say that:

- f has a *node* at $(0, 0)$ if f_2 has two distinct linear factors;
- f has a *cusp* at $(0, 0)$ if f_2 does not, i.e. if it is a square in $\bar{K}[x, y]$.

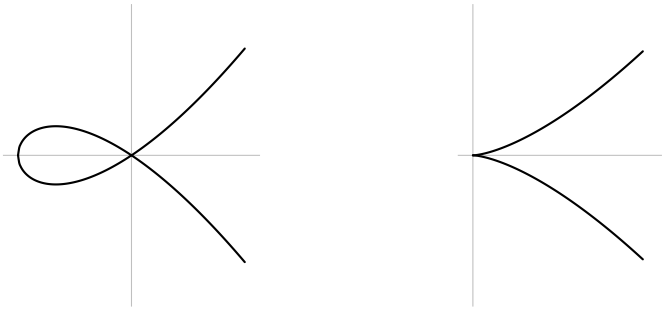
Example 1.1.18. Let C be the curve $y^2 = x^3 + x^2$ from Example 1.1.13. Then we have

$$f_2 = x^2 - y^2 = (x - y)(x + y),$$

so that C has a node at the origin if $\mathrm{char}(K) \neq 2$, and a cusp if $\mathrm{char}(K) = 2$.

Example 1.1.19. Let C be the curve given by $y^2 = x^3$. Then $f_2 = -y^2 = (iy)^2$, where $i^2 = -1$. Hence, C has a cusp at the origin.

Remark 1.1.20. Observe that C has a node at a point P of multiplicity 2 if and only if there are two tangent directions at P, and otherwise C has a cusp at P. Visually, the previous two examples look like this:



$y^2 = x^3 + x^2$ $\qquad$ $y^2 = x^3$

Figure 1.1: The two singular Weierstrass curves.

Finally, we connect the two types of singularities to the quantities c_4, c_6 and Δ.

Proposition 1.1.21. *Let C be a singular Weierstrass curve. Then C has a node if and only if $c_4 \neq 0$. Otherwise, C has a cusp.*

Proof. We will only consider the case $\mathrm{char}(K) \neq 2, 3$. For the remaining cases, see Appendix A of [9].

If $\mathrm{char}(K) \neq 2, 3$, we have an equation of the form

$$y^2 = x^3 - 27c_4x - 54c_6.$$

We know that $0 = 1728\Delta = c_4^3 - c_6^2$, so $c_6 = 0$ if and only if $c_4 = 0$. Hence, if either one is zero, we have $f_2 = -y^2$, so C has a cusp.

On the other hand, if neither is zero, then the singular point of C is the point $(x_0, 0)$ for some x_0 which is a double root of $x^3 - 27c_4x - 54c_6$. Since $c_6 \neq 0$, we have $x_0 \neq 0$, and the map given by $(x, y) \mapsto (x - x_0, y)$ maps C isomorphically to the curve given by

$$y^2 = (x + x_0)^3 - 27c_4(x + x_0) - 54c_6,$$

and $(x_0, 0)$ is mapped to $(0, 0)$. The degree 2 part of the last equation equals

$$y^2 - 3x_0x^2 = (y - \sqrt{3x_0}x)(y + \sqrt{3x_0}x),$$

where $\sqrt{3x_0}$ is some square root of $3x_0$. Since $3x_0 \neq 0$ and $\mathrm{char}(K) \neq 2$, the two factors are distinct, so C has a node. $\square$

1.2 Minimal Weierstrass Form

In this section, let $K = \mathbb{Q}$ be the field of rational numbers. Let C be a Weierstrass curve over $\mathbb{Q}$, with equation

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6.$$

If $u \in \mathbb{Q}^*$ and $r, s, t \in \mathbb{Q}$ are given, we can consider the map given by

$$(x, y) \mapsto \left(u^{-2}(x - r), u^{-3}(y - s(x - r) - t)\right), \tag{2}$$

of which the inverse is given by $(x, y) \mapsto (u^2x + r, u^3y + u^2sx + t)$.

Doing so, we obtain an isomorphic Weierstrass curve with equation given by

$$y^2 + a_1'xy + a_3'y = x^3 + a_2'x^2 + a_4'x + a_6'.$$

The coefficients of the latter can be computed via

$$\begin{aligned}
ua_1' &= a_1 + 2s \\
u^2a_2' &= a_2 - sa_1 + 3r - s^2 \\
u^3a_3' &= a_3 + ra_1 + 2t \\
u^4a_4' &= a_4 - sa_3 + 2ra_2 - (t + rs)a_1 + 3r^2 - 2st \\
u^6a_6' &= a_6 + ra_4 + r^2a_2 + r^3 - ta_3 - t^2 - rta_1.
\end{aligned}$$

Also, the associated quantities b_i' are related to the original b_i by

$$\begin{aligned}
u^2b_2' &= b_2 + 12r \\
u^4b_4' &= b_4 + rb_2 + 6r^2 \\
u^6b_6' &= b_6 + 2rb_4 + r^2b_2 + 4r^3 \\
u^8b_8' &= b_8 + 3rb_6 + 3r^2b_4 + r^3b_2 + 3r^4.
\end{aligned}$$

Finally, the quantities c_4, c_6, Δ and j satisfy:

$$\begin{aligned}
u^4c_4' &= c_4 \\
u^6c_6' &= c_6 \\
u^{12}\Delta' &= \Delta \\
j' &= j.
\end{aligned}$$

This last identity also explains the name of the j-invariant.

Lemma 1.2.1. *Let C be a Weierstrass curve. The only isomorphism onto another Weierstrass curve fixing $[0 : 1 : 0]$ is a map*

$$(x, y) \mapsto \left(u^{-2}(x - r), u^{-3}(y - s(x - r) - t)\right)$$

as above, with $u \in \mathbb{Q}^*$ and $r, s, t \in \mathbb{Q}$.

Proof. See Proposition III.3.1(b) in [9]. □

Remark 1.2.2. If E is an elliptic curve over $\mathbb{Q}$, then we can do a change of variables as in (2) such that all the coefficients become integers, as follows:

We choose r, s and t to be zero, and set $u = \frac{1}{k}$, where k is the least common multiple of the denominators of the a_i. Then after applying our map

$$(x, y) \mapsto \left(u^{-2}(x - r), u^{-3}(y - s(x - r) - t)\right),$$

we get the Weierstrass curve with coefficients $a_i' = k^i a_i$. By the choice of k, all coefficients are now integers.

Definition 1.2.3. Let E be an elliptic curve over $\mathbb{Q}$. An *integral Weierstrass form* of E is a Weierstrass curve E' that is isomorphic to E, such that all coefficients of E' are integers.

Definition 1.2.4. Let E be an elliptic curve over $\mathbb{Q}$. A *minimal Weierstrass form* of E is an integral Weierstrass form E' minimizing the absolute value of the discriminant. The corresponding polynomial is called a *minimal Weierstrass polynomial*.

It is clear that every Weierstrass curve has a minimal Weierstrass form. Furthermore, it is sometimes easy to check that a given Weierstrass curve is in minimal form.

Proposition 1.2.5. *Let E be an integral Weierstrass curve. Suppose that for every prime p one of the following properties holds:*

- $v_p(\Delta) < 12$;
- $v_p(c_4) < 4$;
- $v_p(c_6) < 6$.

Then E is in minimal form.

Proof. It is easy to see that any substitution making $|\Delta|$ smaller will come from setting $|u| > 1$ in (2), so that $v_p(u) > 0$ for some prime p. But then $v_p(\Delta') \leq v_p(\Delta) - 12$, and similarly for c_4 and c_6. But one of these is not possible by assumption. □

Remark 1.2.6. For primes $p \neq 2, 3$, the converse holds as well, as we will see in Proposition 1.2.12.

Remark 1.2.7. Let E be an elliptic curve and let p be a prime. The previous proposition suggests that we might want to consider an integral Weierstrass form E' minimizing $v_p(\Delta)$, instead of $|\Delta|$. In fact, we can weaken the integrality condition, in order to include curves for which the a_i' might be nonintegral, but at least satisfy $v_p(a_i') \geq 0$.

Definition 1.2.8. A Weierstrass form E' of E satisfying $v_p(a_i') \geq 0$ for all $i \in \{1, 2, 3, 4, 6\}$ is called *p-integral*.

Definition 1.2.9. A p-integral Weierstrass form E_p of E is said to be *p-minimal* (or *minimal at p*) if $v_p(\Delta)$ is minimal.

Proposition 1.2.10. *Let E be an elliptic curve, and suppose we have a q-minimal Weierstrass form E_q of E for every prime q. Then any minimal Weierstrass form E' of E has discriminant*

$$\Delta(E') = \prod_{q \text{ prime}} q^{v_q(\Delta(E_q))}.$$

Proof. By Proposition VIII.8.2 of [9], there exists an integral Weierstrass form C of E having the desired discriminant. By q-minimality of E_q, there exists no integral Weierstrass form C_q of E with $v_q(\Delta(C_q)) < v_q(\Delta(E_q)) = v_q(\Delta(C))$, for any given prime q. Since this holds for every prime q, we see that C is in minimal Weierstrass form, hence has the same discriminant as E'. □

Corollary 1.2.11. *Let E be an elliptic curve in minimal Weierstrass form and let q be a prime. Then every integral Weierstrass form E' of E satisfies*

$$v_q(\Delta(E')) \geq v_q(\Delta(E)).$$

Proof. It even holds for every q-integral Weierstrass form, since the q-minimal Weierstrass form E_q of E satisfies $v_q(\Delta(E_q)) = v_q(\Delta(E))$. □

Proposition 1.2.12. *For primes $p \neq 2, 3$, the converse of Proposition 1.2.5 holds as well.*

Proof. Let E be an elliptic curve; suppose that E is in minimal Weierstrass form, but that $v_p(c_4), c_p(c_6)$ and $v_p(\Delta)$ are at least $4, 6$ and 12 respectively. Then the integral Weierstrass curve

$$C : y^2 = x^3 - 27c_4x - 54c_6$$

is isomorphic to E, and $\Delta(C) = 6^{12}\Delta(E)$ has the same valuation at p as $\Delta(E)$. Furthermore, our assumptions on c_4, c_6 and Δ imply that the curve

$$y^2 = x^3 - 27p^{-4}c_4 - 54p^{-6}c_6$$

obtained by the map $(x, y) \mapsto (p^{-2}x, p^{-3}y)$ is integral, and its discriminant has valuation $v_p(\Delta(E)) - 12$. This is impossible by the previous corollary. □

The minimal Weierstrass form is not unique. However, it is almost unique:

Proposition 1.2.13. *The minimal Weierstrass form of an elliptic curve E over $\mathbb{Q}$ is unique up to a change of coordinates*

$$(x, y) \mapsto (u^2x + r, u^3y + u^2sx + t),$$

with $u \in \{\pm 1\}$ and $r, s, t \in \mathbb{Z}$.

Proof. Since any two minimal forms have the same discriminant, we have $u^{12} = 1$. The transformation formulas for b_6 and b_8 show that $2r$ and $3r$ are the zeroes of monic polynomials with integer coefficients, i.e. they are integral over $\mathbb{Z}$. Since they are in $\mathbb{Q}$, this shows that both $2r$ and $3r$ are integers, hence r is as well. Similarly, the formulas for a_2 and a_6 show that s and t are integers. □

1.3 Reduction Modulo Primes

Just like in the previous section, the field K will always be $\mathbb{Q}$ throughout this section.

Definition 1.3.1. Let $p \in \mathbb{Z}$ be a prime, E an elliptic curve over $\mathbb{Q}$ in minimal Weierstrass form. Then the *reduction of E modulo p* is the (possibly singular) Weierstrass curve over $\bar{\mathbb{F}}_p$ given by

$$\tilde{E}(\bar{\mathbb{F}}_p) : y^2 + \bar{a}_1 xy + \bar{a}_3 y = x^3 + \bar{a}_2 x^2 + \bar{a}_4 x + \bar{a}_6.$$

It is defined over $\mathbb{F}_p$, and we denote its set of $\mathbb{F}_p$-rational points by $\tilde{E}(\mathbb{F}_p)$.

Remark 1.3.2. If E is not in minimal Weierstrass form, then we simply choose some E' isomorphic to E that is in minimal Weierstrass form. The reduction of E' will simply be called the reduction of E. This is independent of the choice of E', since by the previous proposition every change in coordinates over $\mathbb{Q}$ keeping E' in minimal form can also be carried out over $\mathbb{F}_p$, simply by reducing u, r, s and t modulo p.

Now we can define the reduction type of an elliptic curve E at a prime p.

Definition 1.3.3. Let $p \in \mathbb{Z}$ prime, E an elliptic curve over $\mathbb{Q}$ in minimal Weierstrass form. Then E is said to have *good* (or *stable*) *reduction at p* if $\tilde{E}$ is nonsingular. If not, then E has *bad reduction at p*.

Definition 1.3.4. Let E be an elliptic curve over $\mathbb{Q}$ in minimal Weierstrass form, and let p be a prime at which E has bad reduction. If $\tilde{E}(\bar{\mathbb{F}}_p)$ has a node, then E has *semistable reduction at p*. If $\tilde{E}(\bar{\mathbb{F}}_p)$ has a cusp, then E is said to have *unstable reduction at p*.

Remark 1.3.5. The words stable, semistable and unstable are used because of the behavior of the reduction types when the field K is enlarged. Since all our curves are defined over $\mathbb{Q}$, they are also defined over any number field, and one could define the reduction at a prime ideal in a number ring. The stable curves will always remain stable over larger ground fields, but semistable curves can become stable, and unstable curves can become stable or semistable.

The explanation for this seemingly wild behavior lies in the observation that minimal polynomials over $\mathbb{Q}$ do not necessarily have to be minimal over finite extensions of $\mathbb{Q}$. See Proposition VII.5.4 in [9].

Remark 1.3.6. If E has bad reduction at p, some authors use the term *multiplicative* (*additive*, respectively) reduction in stead of semistable (unstable, respectively) reduction. The reason for this is that the set of nonsingular points on $\tilde{E}(\mathbb{F}_p)$ has a group structure isomorphic to a multiplicative group, either $\mathbb{F}_p^*$ or the kernel of the norm map $\mathbb{F}_{p^2}^* \to \mathbb{F}_p^*$ (isomorphic to the additive group $\mathbb{F}_p$, respectively). See Proposition III.2.5 and Exercise 3.5 of [9].

Given an elliptic curve E, it is easy to see what the reduction type is at any given prime p.

Lemma 1.3.7. *Let E be an elliptic curve in minimal Weierstrass form as above with discriminant Δ, and let $p \in \mathbb{Z}$ be a prime. Then:*

- *E has good reduction at p if and only if $p \nmid \Delta$;*
- *E has semistable reduction at p if and only if $p \mid \Delta$, and $p \nmid c_4$;*
- *E has unstable reduction at p if and only if $p \mid \Delta, c_4$.*

Proof. Clear from Proposition 1.1.10 and Proposition 1.1.21. □

Definition 1.3.8. Let E be an elliptic curve in minimal Weierstrass form, $p \in \mathbb{Z}$ a prime. We define

$$a_p(E) = p + 1 - \#\tilde{E}(\mathbb{F}_p).$$

Beware that $a_p(E)$ is not directly related to the a_i in the equation for E.

Fact 1.3.9 (Hasse–Weil inequality)**.** We have the following bound:

$$|a_p(E)| \leq 2\sqrt{p}.$$

For a proof, see section 14.4 of [4].

Finally, we will introduce the *conductor* of an elliptic curve E. It is divisible by the same primes as the discriminant. Giving a precise definition requires more than we can discuss here, but the conductor is given by

$$N_E = \prod_{p \mid \Delta} p^{f_p},$$

where

$$f_p = \begin{cases} 1 & \text{if } E \text{ has semistable reduction at } p, \\ 2 & \text{if } E \text{ has unstable reduction at } p \text{ and } p \notin \{2,3\}, \\ 2 + \delta_p & \text{if } E \text{ has unstable reduction at } p \text{ and } p \in \{2,3\}. \end{cases}$$

Here, δ_2 and δ_3 are nonnegative integers depending on E. They satisfy $\delta_2 \leq 6$ and $\delta_3 \leq 3$, and they can be computed via Tate's algorithm, which can be found in [10].

Observe that we have not specified what f_p is when E has good reduction. We do not have to, because those p do not divide the discriminant.

1.4 The Frey Curve and Fermat's Last Theorem

Once again, let $K = \mathbb{Q}$ be the field of rational numbers. We will discuss a part of the proof of Fermat's Last Theorem (FLT):

Theorem 1.4.1 (FLT)**.** *The projective curve over $\bar{\mathbb{Q}}$ given by $x^n + y^n = z^n$ has only trivial points over $\mathbb{Q}$ (namely $[0:1:1], [1:0:1]$ and $[1:-1:0]$ if n is odd, and $[0:1:\pm 1], [1:0:\pm 1]$ if n is even).*

Remark 1.4.2. If the curve $x^n + y^n = z^n$ does not have any nontrivial points, then also the curve $x^{kn} + y^{kn} = z^{kn}$ does not, for every $k \in \mathbb{Z}_{>0}$. Since every integer larger than 2 is divisible by either 4 or an odd prime, it suffices to prove FLT when n is either 4 or an odd prime.

Remark 1.4.3. Elementary proofs are known for $n = 3$ and $n = 4$, so we will restrict ourselves to the case where n is an odd prime larger than 3.

Now assume that $q > 3$ is a prime, and a, b and c are pairwise coprime nonzero integers satisfying

$$a^q + b^q = c^q.$$

Since not all three of a, b and c can be odd, we can assume without loss of generality that b is even. This automatically implies that a and c are odd. Furthermore, we can assume that $a \equiv -1 \mod 4$, by replacing (a, b, c) with $(-a, -b, -c)$, if necessary.

Definition 1.4.4. We define the *Frey curve* E associated to a, b, c and q to be the Weierstrass curve

$$E : y^2 = x(x - a^q)(x + b^q).$$

That is, we have the following values.

$a_1 = 0$	$b_2 = 4(b^q - a^q)$	$c_4 = 16(c^{2q} - (ab)^q)$
$a_2 = b^q - a^q$	$b_4 = -2(ab)^q$	$c_6 = 64a^{3q} + 96a^{2q}b^q - 96a^q b^{2q} - 64b^{3q}$
$a_3 = 0$	$b_6 = 0$	$\Delta = 16(abc)^{2q}$
$a_4 = -(ab)^q$	$b_8 = -(ab)^{2q}$	
$a_6 = 0$		

Remark 1.4.5. If we write $f = x(x - a^q)(x + b^q)$, and put $\alpha_1, \alpha_2, \alpha_3$ for the roots of f, then we find that

$$\Delta(f) = \prod_{i<j} (\alpha_i - \alpha_j)^2 = (0 - a^q)^2(0 + b^q)^2(a^q + b^q)^2 = (abc)^{2q}.$$

This also shows that the discriminant equals

$$\Delta = 16\Delta(f) = 16(abc)^{2q}.$$

Remark 1.4.6. Observe that E is not necessarily in minimal form yet. However, if p is a prime dividing the discriminant, then p divides exactly one of a, b and c. Hence, unless $p = 2$, we see that p cannot divide c_4, so by Proposition 1.2.5, we see that E is minimal at p, meaning that we cannot obtain a curve of which the discriminant has fewer factors p.

Hence, all there is to do is make E minimal at $p = 2$. Note that c_4 has exactly 4 factors 2, so that any transformation making E minimal must have $u = 2$.

Lemma 1.4.7. *A minimal Weierstrass form of E is given by*

$$y^2 + xy = x^3 + \frac{b^q - a^q - 1}{4}x^2 - \frac{a^q b^q}{16}x.$$

Proof. We do the variable substitution as in (2) of the previous section, and we take $u = 2$, $r = t = 0$ and $s = 1$. That is, we consider the map

$$(x, y) \mapsto \left(2^{-2}x, 2^{-3}(y - sx)\right),$$

and we get the desired equation. The equation is integral since $a^q \equiv a \equiv -1 \mod 4$ and $16 \mid b^q$ (by the assumption $q > 3$). By the transformation formulas, we have

$$\Delta = 2^{-8}(abc)^{2q}, \qquad c_4 = c^{2q} - a^q b^q.$$

Since Δ and c_4 have no factors in common, the equation is minimal.

□

Corollary 1.4.8. *The minimal discriminant of E is:*

$$\Delta = 2^{-8}(abc)^{2q}.$$

Corollary 1.4.9. *The conductor of E equals*

$$N_E = \operatorname{rad}(abc),$$

where the radical of an integer is the product of its prime divisors.

Proof. Since Δ and c_4 of the minimal equation are coprime, E has semistable reduction at every prime dividing Δ. Hence,

$$N_E = \prod_{p \mid \Delta} p = \operatorname{rad}(\Delta),$$

and this is equal to $\operatorname{rad}(abc)$ since $2 \mid b$ and $2q > 8$. □

Remark 1.4.10. We could have computed the minimal form and the corresponding conductor using Tate's algorithm, as in [10]. For primes $p > 2$, the algorithm tells us after *Step* 2 that $f_p = 1$, and for $p = 2$, we have to run all the way to *Step* 11 to make E minimal. Then we have to start again from the beginning, where *Step* 2 tells us that $f_p = 1$.

Remark 1.4.11. If we had assumed that $a \equiv 1 \mod 4$ instead of $a \equiv -1 \mod 4$, then E was already in minimal form. In particular, it would have had unstable reduction at $p = 2$, since both Δ and c_4 are divisible by 2. Then Tate's algorithm could be used to compute f_2, but this is considerably more work than what we have done here.

We will come back to the Frey curve in Chapter 3.

2 Modular Forms

2.1 Definitions

Before we give the definition of a modular form, we will firstly introduce some related notions.

Definition 2.1.1. The *modular group* $\mathrm{SL}_2(\mathbb{Z})$ is the multiplicative group given by

$$\mathrm{SL}_2(\mathbb{Z}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbb{Z}, ad - bc = 1 \right\}.$$

Fact 2.1.2. The modular group is generated by $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$.

Proof. See for instance Theorem VII.2 of [7] or Exercise 1.1.1 of [3]. □

Definition 2.1.3. The *upper half plane* $\mathcal{H}$ is the set $\{\tau \in \mathbb{C} : \mathrm{Im}(\tau) > 0\}$.

Definition 2.1.4. For all $\tau \in \mathcal{H}$ and $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$, we define

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} (\tau) := \frac{a\tau + b}{c\tau + d}.$$

Remark 2.1.5. If we identify $\tau \in \mathcal{H}$ with $[\tau : 1] \in \mathbb{P}^1(\mathbb{C})$, the above definition coincides with the natural action of $\mathrm{SL}_2(\mathbb{C})$ on $\mathbb{P}^1(\mathbb{C})$ given by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} [x : y] = [ax + by : cx + dy].$$

Fact 2.1.6. For all $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$ and $\tau \in \mathcal{H}$, it holds that

- $\mathrm{Im}(\gamma(\tau)) = \frac{\mathrm{Im}(\tau)}{(c\tau+d)^2}$;
- $\frac{\mathrm{d}}{\mathrm{d}\,\tau}\gamma(\tau) = \frac{1}{(c\tau+d)^2}$.

Observe that the first property assures that $\gamma(\tau)$ will be in $\mathcal{H}$ for all $\tau \in \mathcal{H}$, so that the action of $\mathrm{SL}_2(\mathbb{C})$ on $\mathbb{P}^1(\mathbb{C})$ restricts to an action of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathcal{H}$.

Now we can come to the definition of a weakly modular function.

Definition 2.1.7. Let $k \in \mathbb{Z}$ be an integer and $f \colon \mathcal{H} \to \mathbb{C}$ a meromorphic function. We say that f is *weakly modular of weight* k if

$$f(\gamma(\tau)) = (c\tau + d)^k f(\tau)$$

for all $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(Z)$ and $\tau \in \mathcal{H}$.

Example 2.1.8. Weak modularity of weight 0 is nothing more than $\mathrm{SL}_2(\mathbb{Z})$-invariance. This shows that all constant functions are weakly modular of weight 0.

Example 2.1.9. Let k be odd, and f weakly modular of weight k. Then the matrix $\left(\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$ shows that $f(\tau) = (-1)^k f(\tau)$, so that f is identically zero.

Unfortunately, it is not easy to give nontrivial examples of weak modularity. We will construct some examples in the next section.

Remark 2.1.10. We have seen that $\mathrm{d}\,\gamma(\tau) = (c\tau + d)^{-2}\,\mathrm{d}\,\tau$, so that being weakly modular of weight 2 is the same as having $\mathrm{SL}_2(\mathbb{Z})$-invariant path integrals on $\mathcal{H}$:

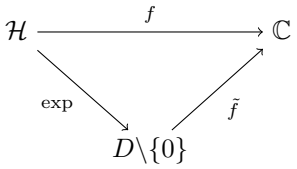
$$f(\gamma(\tau))\,\mathrm{d}(\gamma(\tau)) = (c\tau + d)^2 f(\tau)(c\tau + d)^{-2}\,\mathrm{d}\,\tau = f(\tau)\,\mathrm{d}\,\tau.$$

Definition 2.1.11. We will write $\exp\colon \mathcal{H} \to \mathbb{C}^*$ for the map $\tau \mapsto e^{2\pi i \tau}$. Since $\mathrm{Im}(\tau) > 0$, we have $|\exp(\tau)| < 1$. If we write $D = \{z \in \mathbb{C} : |z| < 1\}$, we see that exp is actually a map $\mathcal{H} \to D\backslash\{0\}$.

Remark 2.1.12. If $f\colon \mathcal{H} \to \mathbb{C}$ is weakly modular, we know in particular that

$$f(\tau + 1) = f\left(\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}(\tau)\right) = 1^k f(\tau) = f(\tau),$$

for all $\tau \in \mathcal{H}$. Hence, weakly modular functions are $\mathbb{Z}$-periodic. As a map, f factors as follows:



By the topological analogue of the fundamental homomorphism theorem, $\tilde{f}$ must be a continuous function, and in fact it is holomorphic.

Definition 2.1.13. The *Fourier expansion of* f is just the Laurent expansion of $\tilde{f}$ around 0. That is:

$$\tilde{f}(q) = \sum_{n=-\infty}^{\infty} a_n(f) q^n,$$

so that

$$f(\tau) = \sum_{n=-\infty}^{\infty} a_n(f) q^n, \qquad \text{where } q = \exp(\tau).$$

Remark 2.1.14. Since $\tilde{f}$ is holomorphic on $D\backslash\{0\}$, the Laurent series is convergent on the same set (see Theorem 4.3.2 of [5]). Hence, the Fourier series of f converges on $\mathcal{H}$. We can ask whether $\tilde{f}$ can be extended holomorphically to D.

Definition 2.1.15. If $\tilde{f}$ can be continued holomorphically to D, we say that f is *holomorphic at* ∞.

Lemma 2.1.16. *The following are equivalent:*

- *f is holomorphic at ∞;*
- *$f(\tau)$ is bounded as $\mathrm{Im}(\tau) \to \infty$;*
- *all coefficients $a_n(f)$ for $n < 0$ of the Fourier expansion are zero.*

Proof. We know that $\tilde{f}$ has a removable singularity at $q = 0$ if and only if $\lim_{q\to 0} \tilde{f}(q) < \infty$. This means exactly that $\tilde{f}(q)$ is bounded as $q \to 0$, or $a_n(f) = 0$ for all $n < 0$. See for instance the discussion at the end of section 4.3 in [5]. □

Definition 2.1.17. Let $k \in \mathbb{Z}$ be an integer and $f\colon \mathcal{H} \to \mathbb{C}$ a function. We say that f is a *modular form of weight k* if it satisfies the following conditions:

- f is weakly modular of weight k,
- f is holomorphic on $\mathcal{H}$,
- f is holomorphic at ∞.

The set of modular forms of weight k will be denoted by $\mathcal{M}_k(\mathrm{SL}_2(\mathbb{Z}))$.

Definition 2.1.18. If $f\colon \mathcal{H} \to \mathbb{C}$ is a modular form of weight k, then f is called a *cusp form* if it satisfies one of the following equivalent conditions:

- $\lim_{\mathrm{Im}(\tau)\to\infty} f(\tau) = 0$;
- $a_0(f) = 0$;
- $\tilde{f}$ has a zero in $q = 0$.

The set of cusp forms of weight k is denoted by $\mathcal{S}_k(\mathrm{SL}_2(\mathbb{Z}))$.

Remark 2.1.19. For all $k \in \mathbb{Z}$, the set $\mathcal{M}_k(\mathrm{SL}_2(\mathbb{Z}))$ comes with a natural $\mathbb{C}$-vectorspace structure: for $f \in \mathcal{M}_k(\mathrm{SL}_2(\mathbb{Z}))$ and $\lambda \in \mathbb{C}$, we find that λf is indeed weakly modular of weight k, as well as holomorphic on $\mathcal{H}$ and at ∞. Hence,

$$\lambda f \in \mathcal{M}_k(\mathrm{SL}_2(\mathbb{Z})).$$

Similarly, for $f, g \in \mathcal{M}_k(\mathrm{SL}_2(\mathbb{Z}))$, we have $f + g \in \mathcal{M}_k(\mathrm{SL}_2(\mathbb{Z}))$, so that we indeed have a $\mathbb{C}$-vectorspace.

Remark 2.1.20. The product of a modular form of weight k with a modular form of weight ℓ is a modular form of weight $k + \ell$, and as such, the direct sum

$$\mathcal{M}(\mathrm{SL}_2(\mathbb{Z})) = \bigoplus_{k\in\mathbb{Z}} \mathcal{M}_k(\mathrm{SL}_2(\mathbb{Z}))$$

becomes a graded $\mathbb{C}$-algebra.

Example 2.1.21. For odd k, we find $\mathcal{M}_k(\mathrm{SL}_2(\mathbb{Z})) = 0$, by Example 2.1.9.

Remark 2.1.22. Since the product of a cusp form of weight k and a modular form of weight ℓ is a cusp form of weight $k + \ell$, the subspace

$$\mathcal{S}(\mathrm{SL}_2(\mathbb{Z})) = \bigoplus_{k\in\mathbb{Z}} \mathcal{S}_k(\mathrm{SL}_2(\mathbb{Z}))$$

of $\mathcal{M}(\mathrm{SL}_2(\mathbb{Z}))$ is in fact an ideal. It is by definition homogeneous.

2.2 Eisenstein Series and the Discriminant

Now, as promised, we will provide some nontrivial examples of modular forms.

Definition 2.2.1. Let $k \geq 4$ be an even integer. Define the *Eisenstein series of weight k* as

$$G_k(\tau) = \sum_{(c,d)\neq(0,0)} \frac{1}{(c\tau + d)^k},$$

for all $\tau \in \mathcal{H}$, where the sum runs over all nonzero pairs of integers.

Lemma 2.2.2. *The Eisenstein series of weight $k \geq 4$ is absolutely convergent and converges uniformly on compact subsets of $\mathcal{H}$.*

Proof. We write $L = \mathbb{Z}^2 \backslash \{(0,0)\}$ and $L_n = \{(c,d) \in L : \sup\{|c|, |d|\} = n\}$ for all $n \in \mathbb{Z}_{>0}$. Then we have:

$$\sum_{(c,d)\in L_n} \frac{1}{\sup\{|c|, |d|\}^k} = \#L_n \cdot \frac{1}{n^k}.$$

A simple computation shows that $\#L_n = 8n$, so we find

$$\sum_{(c,d)\in L} \frac{1}{\sup\{|c|, |d|\}^k} = \sum_{n=1}^{\infty} \sum_{(c,d)\in L_n} \frac{1}{\sup\{|c|, |d|\}^k} = \sum_{n=1}^{\infty} \frac{8n}{n^k} = 8\zeta(k-1).$$

The idea is to estimate $G_k(\tau)$ in terms of the one we have just calculated.

Let $A, B \in \mathbb{R}_{>0}$ be given, and let $\Omega = \{\tau \in \mathcal{H} : |\operatorname{Re}(\tau)| \leq A, \operatorname{Im}(\tau) \geq B\}$. The set Ω is depicted in Figure 2.2.

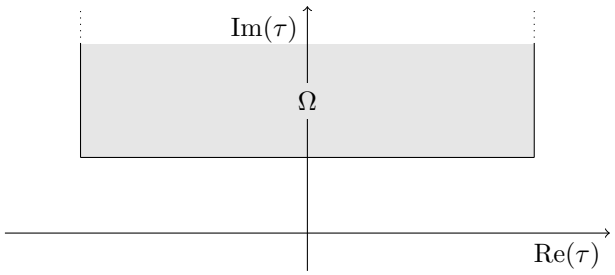


Figure 2.2: The set Ω.

Define $C = \inf\{\frac{B}{2}, \frac{B}{3A}, \frac{1}{3}\}$. Then one can show that

$$|\tau + \delta| > C \sup\{1, \delta\},$$

for all $\tau \in \Omega$ and $\delta \in \mathbb{R}$. Hence, for all $(c,d) \in L$, $\tau \in \Omega$, we have

$$|c\tau + d| = |c| \left|\tau + \frac{d}{c}\right| > |c| C \sup\left\{1, \left|\frac{d}{c}\right|\right\} = C \sup\{|c|, |d|\},$$

unless $c = 0$ in which case the result is obvious.

Hence, we have

$$\frac{1}{|c\tau+d|^k} < \frac{1}{C^k}\frac{1}{\sup\{|c|,|d|\}^k},$$

for all $\tau \in \Omega$. Summing over all k shows that the series converges absolutely and uniformly on Ω. □

Lemma 2.2.3. *The Eisenstein series of weight k is a modular form of weight k.*

Proof. We have already seen that it converges absolutely and uniformly on compact subsets of $\mathcal{H}$. Hence, $G_k(\tau)$ is holomorphic on $\mathcal{H}$ and its terms may be rearranged.

If we take any $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$, we find that

$$\begin{aligned} G_k(\gamma(\tau)) &= \sum_{(c',d')\neq(0,0)} \frac{1}{\left(c'\left(\frac{a\tau+b}{c\tau+d}\right)+d'\right)^k} \\ &= (c\tau+d)^k \sum_{(c',d')\neq(0,0)} \frac{1}{((c'a+d'c)\tau+(c'b+d'd))^k}. \end{aligned}$$

Since multiplication on the right by γ is a bijection from $\mathbb{Z}^2$ to itself fixing $(0,0)$, we see that

$$(c'a+d'c, c'b+d'd) = (c',d')\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

runs through $\mathbb{Z}^2\backslash\{(0,0)\}$ as (c',d') does. Hence, we have

$$G_k(\gamma(\tau)) = (c\tau+d)^k G_k(\tau),$$

so that G_k is weakly modular of weight k.

Finally, the computations in the proof of the preceding lemma show that $C \to \frac{1}{3}$ as $B \to \infty$, so

$$\sum_{(c,d)\neq(0,0)} \frac{1}{(c\tau+d)^k} < \frac{1}{C^k} \sum_{(c,d)\neq(0,0)} \frac{1}{\sup\{|c|,|d|\}^k} = C^{-k} 8\zeta(k-1) \to 3^k \cdot 8\zeta(k-1)$$

as $B \to \infty$, and G_k is bounded as $\tau \to i\cdot\infty$. □

Now we will also give an example of a cusp form. In order to do so, we must compute some Fourier coefficients.

Lemma 2.2.4. *Let $k \geq 4$ be an even integer. The Eisenstein series G_k has Fourier expansion*

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

where the function σ_{k-1} is given by

$$\sigma_{k-1}(n) = \sum_{d|n} d^{k-1}.$$

Proof. We will start from the well-known identities

$$\frac{1}{\tau} + \sum_{d=1}^{\infty} \left(\frac{1}{\tau - d} + \frac{1}{\tau + d} \right) = \pi \cot \pi\tau = \pi i - 2\pi i \sum_{m=1}^{\infty} e^{2\pi i m \tau}.$$

We will once again write $q = e^{2\pi i \tau} = \exp \tau$, and we get

$$\frac{1}{\tau} + \sum_{d=1}^{\infty} \left(\frac{1}{\tau - d} + \frac{1}{\tau + d} \right) = \pi i - 2\pi i \sum_{m=1}^{\infty} q^m.$$

Differentiating with respect to τ and changing the sign on both sides gives us

$$\frac{1}{\tau^2} + \sum_{d=1}^{\infty} \left(\frac{1}{(\tau - d)^2} + \frac{1}{(\tau + d)^2} \right) = 2\pi i \sum_{m=1}^{\infty} 2\pi i m q^m,$$

or equivalently

$$\sum_{d \in \mathbb{Z}} \frac{1}{(\tau + d)^2} = (2\pi i)^2 \sum_{m=1}^{\infty} m q^m.$$

Differentiating a further $k-2$ times (w.r.t. τ), we get

$$\sum_{d \in \mathbb{Z}} (k-1)! \frac{1}{(\tau + d)^k} = (2\pi i)^k \sum_{m=1}^{\infty} m^{k-1} q^m.$$

If we replace τ by $c\tau$, this gives us

$$\sum_{d \in \mathbb{Z}} (k-1)! \frac{1}{(c\tau + d)^k} = (2\pi i)^k \sum_{m=1}^{\infty} m^{k-1} q^{cm}.$$

Summing over all $c \in \mathbb{Z}_{>0}$, we get

$$\begin{aligned} G_k(\tau) &= \sum_{(c,d) \neq (0,0)} \frac{1}{(c\tau + d)^k} \\ &= 2\zeta(k) + 2 \sum_{c=1}^{\infty} \sum_{d \in \mathbb{Z}} \frac{1}{(c\tau + d)^k} \\ &= 2\zeta(k) + 2 \frac{(2\pi i)^k}{(k-1)!} \sum_{c=1}^{\infty} \sum_{m=1}^{\infty} m^{k-1} q^{cm}. \end{aligned}$$

If we take a closer look at the sum appearing on the right-hand side, we see that the contribution to q^n is exactly $\sigma_{k-1}(n)$ for all $n \in \mathbb{Z}_{>0}$, and the statement follows. □

Corollary 2.2.5. *The Eisenstein series of even weight $k \geq 4$ satisfies*

$$a_0(G_k) = 2\zeta(k) = -\frac{(2\pi i)^k}{k!} B_k,$$

where B_k is the k-th Bernoulli number, defined by the formal power series

$$\frac{t}{e^t - 1} = \sum_{k=0}^{\infty} B_k \frac{t^k}{k!}.$$

Proof. The first equality follows from the Lemma, and the second equality follows from Prop. VIII.7 of [7]. Beware that Serre uses different 'Bernoulli numbers', and writes $2k$ instead of k. □

Corollary 2.2.6. *The normalized Eisenstein series $E_k = G_k/(2\zeta(k))$ of even weight $k \geq 4$ has Fourier series*

$$E_k(\tau) = 1 - \frac{2k}{B_k} \sum_{n=1}^{\infty} \sigma_{k-1}(n) q^n.$$

In particular, it has rational coefficients with a common denominator. When multiplied by the numerator of B_k, we even get a modular form with integer coefficients.

Proof. We use the Fourier expansion of G_k and the formula of $2\zeta(k)$ to find

$$\begin{aligned} E_k(\tau) = \frac{G_k(\tau)}{2\zeta(k)} &= 1 - 2\frac{(2\pi i)^k}{(k-1)!}\frac{k!}{(2\pi i)^k B_k} \sum_{n=1}^{\infty} \sigma_{k-1}(n) q^n \\ &= 1 - 2\frac{k}{B_k} \sum_{n=1}^{\infty} \sigma_{k-1}(n) q^n. \end{aligned}$$

□

Corollary 2.2.7. *It holds that:*

$$a_0(G_4) = \frac{\pi^4}{45}, \qquad a_0(G_6) = \frac{2\pi^6}{945}.$$

Proof. This follows since $B_4 = -\frac{1}{30}$ and $B_6 = \frac{1}{42}$. □

Definition 2.2.8. Define $g_2 = 60G_4$, $g_3 = 140G_6$. Also, define

$$\Delta = g_2^3 - 27g_3^2 = (60 \cdot 2\zeta(4))^3 E_4^3 - 27(140 \cdot 2\zeta(6))^2 E_6^2.$$

This Δ is called the *discriminant*.

Lemma 2.2.9. *The discriminant is a cusp form of weight* 12.

Proof. Since both E_4^3 and E_6^2 are modular forms of weight 12, so is Δ. Furthermore, we have

$$\begin{aligned} 60 \cdot 2\zeta(4) &= 60\frac{\pi^4}{45} = \frac{4}{3}\pi^4, \\ 140 \cdot 2\zeta(6) &= 140\frac{2\pi^6}{945} = \frac{8}{27}\pi^6. \end{aligned}$$

Hence,

$$(60 \cdot 2\zeta(4))^3 = \frac{2^6}{3^3}\pi^{12} = 27\frac{2^6}{3^6}\pi^{12} = (140 \cdot 2\zeta(6))^2,$$

so that $\Delta = \frac{2^6}{3^3}\pi^{12}(E_4^3 - E_6^2)$. Since both E_4 and E_6 have constant coefficient 1, so do E_4^3 and E_6^2. Hence, their difference has constant coefficient 0, so that Δ is a cusp form. □

2.3 The Ring of Modular Forms

We will prove some facts about $\mathcal{M}(\mathrm{SL}_2(\mathbb{Z}))$ and the ideal $\mathcal{S}(\mathrm{SL}_2(\mathbb{Z}))$. In order to do so, we need some notation.

Definition 2.3.1. Let $f\colon \mathcal{H} \to \mathbb{C}$ be meromorphic, $p \in \mathcal{H}$ a point. Then $v_p(f)$ is the unique integer n such that $f \cdot (\tau - p)^{-n}$ is holomorphic and nonzero at p. It is called the *order* or *valuation* of f at p.

This is in fact a standard notation from complex function theory, and can be defined for any open $U \subseteq \mathbb{C}$ and meromorphic $f\colon U \to \mathbb{C}$. If f is weakly modular, we define $v_\infty(f)$ to be $v_0(\tilde{f})$, with $\tilde{f}$ as in 2.1.12.

Remark 2.3.2. Note that f is holomorphic at ∞ if and only if $v_\infty(f) \geq 0$, and a modular form is a cusp form if and only if $v_\infty > 0$.

We state some basic properties of the valuations at all points of a modular form.

Proposition 2.3.3. *Let f be weakly modular of weight k, let $p \in \mathcal{H}$ and $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$ be given. Then*

$$v_p(f) = v_{\gamma(p)}(f).$$

Proof. This follows from the identity

$$f(\gamma(\tau)) = (c\tau + d)^k f(\tau),$$

using that $c\tau + d$ has only real zeroes and that $\mathbb{R} \cap \mathcal{H} = \varnothing$. □

Hence, the valuation of f at a point P in $\mathcal{H}$ depends only on the $\mathrm{SL}_2(Z)$-orbit of P, and we can define the (by the previous proposition well-defined) map

$$\begin{aligned} v(f)\colon \mathrm{SL}_2(\mathbb{Z})\backslash\mathcal{H} &\to \mathbb{Z} \\ \mathrm{SL}_2(\mathbb{Z})x &\mapsto v_x(f), \end{aligned}$$

and we will denote the valuation of f at an element $x \in \mathrm{SL}_2(\mathbb{Z})\backslash\mathcal{H}$ simply by $v_x(f)$.

Lemma 2.3.4. *Let f be a weakly modular function of weight k that is not identically zero. Then the following formula holds:*

$$v_\infty(f) + \frac{1}{2}v_i(f) + \frac{1}{3}v_{\zeta_3}(f) + \sum_{\substack{x \in \mathrm{SL}_2(\mathbb{Z})\backslash\mathcal{H}, \\ x \neq i, \zeta_3}} v_p(f) = \frac{k}{12}. \tag{3}$$

Proof. This is proven by integrating $\frac{1}{2\pi i}\frac{\mathrm{d}f}{f}$ on the boundary of a fundamental domain for the $\mathrm{SL}_2(\mathbb{Z})$-action on $\mathcal{H}$. We will not give the details here, but they can be found in Theorem VII.3 of [7]. Beware that Serre uses $2k$ for what we call k. □

We will use the lemma to compute the dimension of $\mathcal{M}_k(\mathrm{SL}_2(\mathbb{Z}))$ for $k \in \mathbb{Z}$.

Lemma 2.3.5. *We have $\mathcal{M}_k(\mathrm{SL}_2(\mathbb{Z})) = 0$ for $k < 0$ and $k = 2$.*

Proof. If f is any nonzero modular form of weight $k < 0$, we have

$$v_\infty(f) + \frac{1}{2}v_i(f) + \frac{1}{3}v_{\zeta_3}(f) + \sum_{\substack{x \in \mathrm{SL}_2(\mathbb{Z})\backslash\mathcal{H}, \\ x \neq i, \zeta_3}} v_p(f) = \frac{k}{12} < 0.$$

On the other hand, $v_p(f) \geq 0$ for all $p \in \mathcal{H} \cup \{\infty\}$, since f is holomorphic on $\mathcal{H}$ and holomorphic at ∞. This is a problem for f, which it solves by ceasing to exist.

As for $k = 2$, we see that (3) has no solutions in nonnegative integers $v_x(f)$ ($x \in \mathcal{H} \cup \{\infty\}$). □

We want to use (3) for the cusp form Δ, but we can only do this if $\Delta \neq 0$.

Proposition 2.3.6. *The discriminant is not identically zero.*

Proof. We apply Lemma 2.3.4 to $f = G_2$ and $f = G_3$, which are nonzero since their constant terms equal $2\zeta(4)$ and $2\zeta(6)$ respectively. For G_2, the only possibility is $v_{\zeta_3}(G_2) = 1$ and $v_p(G_2) = 0$ for all other $p \in \mathrm{SL}_2(\mathbb{Z})\backslash\mathcal{H} \cup \{\infty\}$. For G_3, the only possibility is $v_i(G_3) = 1$ and $v_p(G_3) = 0$ for all other $p \in \mathrm{SL}_2(Z)\backslash\mathcal{H} \cup \{\infty\}$. This shows that Δ is not zero at i. □

Now equation (3) implies that Δ is nowhere zero on $\mathcal{H}$, since $v_\infty(\Delta) = 1$. Hence, on $\mathcal{H} \cup \{\infty\}$, the discriminant only has a zero at ∞, which is of order 1.

Lemma 2.3.7. *Let f be a cusp form of weight k. Then $g = \frac{f}{\Delta}$ is a modular form of weight $k - 12$.*

Proof. Clearly, g is weakly modular of weight $k - 12$. It is holomorphic on $\mathcal{H}$ since Δ is nonvanishing on $\mathcal{H}$, and it is holomorphic at ∞ since $v_\infty(g) = v_\infty(f) - v_\infty(\Delta) = v_\infty(f) - 1 \geq 0$. □

Corollary 2.3.8. *Multiplication by Δ defines a linear isomorphism*

$$\mathcal{M}_{k-12}(\mathrm{SL}_2(\mathbb{Z})) \to \mathcal{S}_k(\mathrm{SL}_2(\mathbb{Z})).$$

Proof. It is clearly linear, and its inverse is given by $f \mapsto \frac{f}{\Delta}$. □

Corollary 2.3.9. *The ideal $\mathcal{S}(\mathrm{SL}_2(\mathbb{Z}))$ is the principal ideal generated by Δ.*

Proof. Apply Lemma 2.3.7 to each homogeneous component $\mathcal{S}_k(\mathrm{SL}_2(\mathbb{Z}))$. □

Now that we know what the ideal $\mathcal{S}(\mathrm{SL}_2(\mathbb{Z}))$ is, we can concern ourselves with the structure of the entire ring $\mathcal{M}(\mathrm{SL}_2(\mathbb{Z}))$.

Lemma 2.3.10. *If $f_k \in \mathcal{M}_k(\mathrm{SL}_2(\mathbb{Z}))$ has constant term $a_0(f_k) = 1$, then*

$$\mathcal{M}_k(\mathrm{SL}_2(\mathbb{Z})) = \mathcal{S}_k(\mathrm{SL}_2(\mathbb{Z})) \oplus \mathbb{C} f_k.$$

Proof. If f is any modular form of weight k, then $\lambda = a_0(f)$ is the unique $\lambda \in \mathbb{C}$ for which $f - \lambda f_k$ is a cusp form. Hence, any modular form of weight k can be uniquely written as a linear combination of f_k and a cusp form. □

Theorem 2.3.11. *For even $k \geq 0$ we have*

$$\dim \mathcal{M}_k(\mathrm{SL}_2(\mathbb{Z})) = \begin{cases} \lfloor \frac{k}{12} \rfloor + 1 & \text{if } k \not\equiv 2 \mod 12, \\ \lfloor \frac{k}{12} \rfloor & \text{if } k \equiv 2 \mod 12. \end{cases} \tag{4}$$

Proof. We will use induction, and distinguish two cases: $k \not\equiv 2 \mod 12$ and $k \equiv 2 \mod 12$ respectively. For $k < 0$ (and $k = 2$, respectively), we know that $\mathcal{M}(\mathrm{SL}_2(\mathbb{Z})) = 0$ by Lemma 2.3.5.

To establish a base case, note that for $k \in \{0, 4, 6, 8, 10, 14\}$ we have already found a modular form f_k with constant term 1 (namely $1, E_4, E_6, E_8, E_{10}$ and E_{14} respectively). Hence,

$$\mathcal{M}_k(\mathrm{SL}_2(\mathbb{Z})) = \mathcal{S}_k(\mathrm{SL}_2(\mathbb{Z})) \oplus \mathbb{C} f_k.$$

Since $\dim \mathcal{S}_k(\mathrm{SL}_2(\mathbb{Z})) = \dim \mathcal{M}_{k-12}(\mathrm{SL}_2(\mathbb{Z})) = 0$, we have $\mathcal{M}_k(\mathrm{SL}_2(\mathbb{Z})) = \mathbb{C} f_k$ for $k \in \{0, 4, 6, 8, 10, 14\}$. The general case follows inductively from Lemma 2.3.10, since both sides of (4) increase by 1 when k is increased by 12. □

Corollary 2.3.12. *Let $k \in \mathbb{Z}$ be given. The set*

$$\{E_4^a E_6^b : a, b \in \mathbb{Z}_{\geq 0}, 4a + 6b = k\}$$

is a basis of $\mathcal{M}_k(\mathrm{SL}_2(\mathbb{Z}))$.

Proof. For odd k, as well as for $k \leq 2$, this is obvious, so let $k \geq 4$ be an even integer. A counting argument shows that the number of pairs $(a, b) \in \mathbb{Z}_{\geq 0}^2$ such that $4a + 6b = k$ equals $\lfloor \frac{k}{12} \rfloor + 1$ if $k \not\equiv 2 \mod 12$, and $\lfloor \frac{k}{12} \rfloor$ if $k \equiv 2 \mod 12$.

Suppose the $E_4^a E_6^b$ satisfy some linear relation. Then the weakly modular function $\frac{E_4^3}{E_6^2}$ of weight 0 satisfies some algebraic relation over $\mathbb{C}$. Hence, it is must be constant. But $E_6(i) = 0 \neq E_4(i)$, and we get a contradiction.

Hence, the given modular forms are linearly independent, and since their number equals the dimension, we are done. □

Corollary 2.3.13. *The map $\varepsilon \colon \mathbb{C}[X, Y] \to \mathcal{M}(\mathrm{SL}_2(\mathbb{Z}))$ given by $X \mapsto E_4$, $Y \mapsto E_6$ is an isomorphism of $\mathbb{C}$-algebras. If we give X weight 4 and Y weight 6, this is an isomorphism of graded $\mathbb{C}$-algebras.*

Proof. This is just the previous corollary reformulated. □

To summarize the results from this section:

Theorem 2.3.14. *We have the following equalities:*

$$\mathcal{M}(\mathrm{SL}_2(\mathbb{Z})) = \mathbb{C}[E_4, E_6],$$
$$\mathcal{S}(\mathrm{SL}_2(\mathbb{Z})) = (\Delta),$$

where the notation $\mathbb{C}[E_4, E_6]$ should be read as a polynomial ring in the free variables E_4, E_6 and (Δ) is the homogeneous ideal generated by Δ. □

2.4 Congruence Subgroups

Now that we have completely determined the structure of $\mathcal{M}(\mathrm{SL}_2(\mathbb{Z}))$ and $\mathcal{S}(\mathrm{SL}_2(\mathbb{Z}))$, we can generalize the notion of modularity. This is done by replacing $\mathrm{SL}_2(\mathbb{Z})$ in the definition of a modular form by special types of subgroups.

Definition 2.4.1. Let $N \in \mathbb{Z}_{>0}$. The kernel of the homomorphism

$$\phi_N \colon \mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$$
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \begin{pmatrix} \bar{a} & \bar{b} \\ \bar{c} & \bar{d} \end{pmatrix}$$

is denoted $\Gamma(N)$. In fact, an easy calculation shows that ϕ_N is surjective, so

$$\mathrm{SL}_2(\mathbb{Z})/\Gamma(N) \cong \mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z}).$$

We also put

$$\Gamma_0(N) = \left\{ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z}) : c \equiv 0 \mod N \right\}$$

and

$$\Gamma_1(N) = \left\{ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z}) : c \equiv 0 \mod N,\ a \equiv d \equiv 1 \mod N \right\}.$$

Remark 2.4.2. We have the inclusions

$$\Gamma(N) \subseteq \Gamma_1(N) \subseteq \Gamma_0(N) \subseteq \mathrm{SL}_2(\mathbb{Z}).$$

Proposition 2.4.3. *Let $N \in \mathbb{Z}_{>0}$. Then the following statements hold:*

(a) $\Gamma_1(N) \triangleleft \Gamma_0(N)$, *and* $\Gamma_0(N)/\Gamma_1(N) \cong (\mathbb{Z}/N\mathbb{Z})^*$;
(b) $\Gamma(N) \triangleleft \Gamma_1(N)$, *and* $\Gamma_1(N)/\Gamma(N) \cong \mathbb{Z}/N\mathbb{Z}$;
(c) $[\mathrm{SL}_2(\mathbb{Z}) : \Gamma(N)] = N^3 \prod_{p|N} (1 - \frac{1}{p^2})$.

Proof. For (a), we use the homomorphism

$$\Gamma_0(N) \to (\mathbb{Z}/N\mathbb{Z})^*,$$
$$\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \mapsto \bar{d}.$$

It has kernel $\Gamma_1(N)$, and surjectivity follows since for all $d \in \mathbb{Z}$ with $\gcd(d, N) = 1$ there are $a, b \in \mathbb{Z}$ such that $\left(\begin{smallmatrix} a & b \\ N & d \end{smallmatrix}\right)$ has determinant 1.

For (b), we use the homomorphism

$$\begin{aligned}\Gamma_1(N) &\to \mathbb{Z}/N\mathbb{Z},\\ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) &\mapsto \bar{b}.\end{aligned}$$

It is has kernel $\Gamma(N)$, and surjectivity follows since $\left(\begin{smallmatrix} 1 & b \\ 0 & 1 \end{smallmatrix}\right) \in \Gamma_1(N)$ for all $b \in \mathbb{Z}$.

To prove (c), note that $[\mathrm{SL}_2(\mathbb{Z}) : \Gamma(N)] = \#\,\mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$. We will use induction on e to prove the formula when $N = p^e$.

For $e = 1$, the result is obvious. Let $e \geq 1$ be given, and consider the natural map $\phi_e \colon \mathrm{SL}_2(\mathbb{Z}/p^{e+1}\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/p^e\mathbb{Z})$. It is surjective since $\mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/p^e\mathbb{Z})$ is.

If we choose $a, b, c \in \mathbb{Z}/p^{e+1}\mathbb{Z}$ such that $a \equiv 1 \mod p^e$ and $b \equiv c \equiv 0 \mod p^e$, then bc is congruent to 0 modulo p^{2e}, so in particular modulo p^{e+1}. Hence, $ad - bc \equiv ad \mod p^{e+1}$ holds for all $d \in \mathbb{Z}/p^{e+1}\mathbb{Z}$. This shows that there is a unique $d \in \mathbb{Z}/p^{e+1}\mathbb{Z}$ such that $ad - bc \equiv 1 \mod p^{e+1}$. Observe that this d will automatically be congruent to 1 modulo p^e.

Hence, the number of elements $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathbb{Z}/p^{e+1}\mathbb{Z}$ that map to $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right) \in \mathbb{Z}/p^e\mathbb{Z}$ is equal to p^3, and the desired formula follows inductively for $N = p^e$.

Now proceed by the Chinese Remainder Theorem, using that for arbitrary rings R_1 and R_2, the map

$$\begin{aligned}\mathrm{SL}_2(R_1 \times R_2) &\to \mathrm{SL}_2(R_1) \times \mathrm{SL}_2(R_2),\\ \begin{pmatrix} (a_1,a_2) & (b_1,b_2) \\ (c_1,c_2) & (d_1,d_2) \end{pmatrix} &\mapsto \left(\begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix}, \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix} \right)\end{aligned}$$

is an isomorphism. □

Corollary 2.4.4. *We have*

$$\begin{aligned}[\mathrm{SL}_2(\mathbb{Z}) : \Gamma_0(N)] &= N \prod_{p \mid N} (1 + \tfrac{1}{p});\\ [\Gamma_0(N) : \Gamma_1(N)] &= N \prod_{p \mid N} (1 - \tfrac{1}{p});\\ [\Gamma_1(N) : \Gamma(N)] &= N.\end{aligned}$$

Proof. The second and third identity are clear form the proposition, and the first follows from the identity

$$[\mathrm{SL}_2(\mathbb{Z}) : \Gamma_0(N)] = \frac{[\mathrm{SL}_2(\mathbb{Z}) : \Gamma(N)]}{[\Gamma_0(N) : \Gamma_1(N)][\Gamma_1(N) : \Gamma(N)]}.$$

□

Now we can specify the subgroups of interest within $\mathrm{SL}_2(\mathbb{Z})$.

Definition 2.4.5. Let $N \in \mathbb{Z}_{>0}$. A subgroup $\Gamma \subseteq \mathrm{SL}_2(\mathbb{Z})$ is called a *congruence subgroup of level* N if $\Gamma(N) \subseteq \Gamma$.

Note that a congruence subgroup of level N is also a congruence subgroup of level kN for all $k \in \mathbb{Z}_{>0}$. Furthermore, any congruence subgroup has finite index in $\mathrm{SL}_2(\mathbb{Z})$, since $\Gamma(N)$ has (for every $N \in \mathbb{Z}_{>0}$).

Example 2.4.6. The groups $\Gamma(N), \Gamma_1(N)$ and $\Gamma_0(N)$ are congruence subgroups of level N. We have a chain of inclusions

$$\Gamma(N) \xrightarrow{\quad N \quad} \Gamma_1(N) \xrightarrow{N\prod(1-\frac{1}{p})} \Gamma_0(N) \xrightarrow{N\prod(1+\frac{1}{p})} \mathrm{SL}_2(\mathbb{Z}),$$

where the number on top of the arrow indicates the index.

Definition 2.4.7. Let $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$ be given and let $k \in \mathbb{Z}$ be an integer. Then define the operator $[\gamma]_k$ on meromorphic functions $f\colon \mathcal{H} \to \mathbb{C}$ via

$$f[\gamma]_k(\tau) = (c\tau + d)^{-k} f(\gamma(\tau)).$$

Remark 2.4.8. The notation $f[\gamma]_k$ instead of $[\gamma]_k f$ is used since the action of $\mathrm{SL}_2(\mathbb{Z})$ on the meromorphic functions $f\colon \mathcal{H} \to \mathbb{C}$ is a right action in the sense that

$$(f[\gamma_1]_k)[\gamma_2]_k = f[\gamma_1\gamma_2]_k$$

for all $\gamma_1, \gamma_2 \in \mathrm{SL}_2(\mathbb{Z})$ and $f\colon \mathcal{H} \to \mathbb{C}$ meromorphic. See Lemma 1.2.2 of [3].

Definition 2.4.9. Let $k \in \mathbb{Z}$ be an integer and $f\colon \mathcal{H} \to \mathbb{C}$ a meromorphic function. Let Γ be a congruence subgroup of some level N. Then f is said to be *weakly modular of weight k with respect to* Γ if

$$f[\gamma]_k = f$$

for all $\gamma \in \Gamma$.

Note that for $\Gamma = \mathrm{SL}_2(\mathbb{Z})$, this definition coincides with our previous definition of weak modularity.

In view of the Modularity Theorem, we will be mostly interested in $\Gamma_0(N)$. Note that $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right) \in \Gamma_0(N)$, so that weakly modular functions with respect to $\Gamma_0(N)$ are $\mathbb{Z}$-periodic, analogous to Remark 2.1.12. (In fact, even weakly modular functions with respect to $\Gamma_1(N)$ are $\mathbb{Z}$-periodic.)

In the general case, since $\Gamma(N) \subseteq \Gamma$, we have $\left(\begin{smallmatrix} 1 & N \\ 0 & 1 \end{smallmatrix}\right) \in \Gamma$, so that a weakly modular function f is at least $N\mathbb{Z}$-periodic. Hence, it has some Fourier series

$$f(\tau) = \sum_{n=-\infty}^{\infty} a_n q_N^n, \qquad \text{where } q_N = \exp\tfrac{\tau}{N} \tag{5}$$

It is tempting to call f holomorphic at ∞ if $a_n = 0$ for all $n < 0$, but it turns out we have to be more careful.

Proposition 2.4.10. *The orbit of ∞ under the $\mathrm{SL}_2(\mathbb{Z})$-action on $\mathbb{P}^1(\mathbb{C})$ is $\mathbb{P}^1(\mathbb{Q}) = \mathbb{Q} \cup \{\infty\}$.*

Proof. Let $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$ be given, then $\gamma(\infty) = \frac{a\infty+b}{c\infty+d} = \frac{a}{c}$, so that the orbit of ∞ is contained in $\mathbb{Q} \cup \{\infty\}$.

On the other hand, if $\frac{a}{c} \in \mathbb{Q}$ is given (with $a, c \in \mathbb{Z}$ coprime integers), then there exist $b, d \in \mathbb{Z}$ such that $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$. Hence, every element of $\mathbb{Q}$ is in the orbit of ∞. □

The previous proposition motivates the following definition.

Definition 2.4.11. Let $\Gamma \subseteq \mathrm{SL}_2(\mathbb{Z})$ be a congruence subgroup. A *cusp* for Γ is a Γ-equivalence class of points in $\mathbb{P}^1(\mathbb{Q})$.

Remark 2.4.12. Since every element of Γ fixes every cusp, the number of cusps for Γ is at most the number of cosets of Γ in $\mathrm{SL}_2(\mathbb{Z})$. Since Γ has finite index in $\mathrm{SL}_2(\mathbb{Z})$, there are only finitely many cusps for Γ.

Definition 2.4.13. Let $f\colon \mathcal{H} \to \mathbb{C}$ be weakly modular of weight k with respect to some congruence subgroup Γ, and let P be a cusp for Γ. Then f is *holomorphic at P* if $f[\alpha]_k$ is holomorphic at ∞, for all $\alpha \in \mathrm{SL}_2(\mathbb{Z})$ such that $\alpha(P) = \infty$.

Now we can come to the definition of a modular form with respect to a congruence subgroup.

Definition 2.4.14. Let $k \in \mathbb{Z}$ be an integer, $\Gamma \subseteq \mathrm{SL}_2(\mathbb{Z})$ a congruence subgroup and $f\colon \mathcal{H} \to \mathbb{C}$ a function. We say that f is a *modular form of weight k with respect to Γ* if it satisfies the following conditions:

- f is weakly modular of weight k with respect to Γ,
- f is holomorphic on $\mathcal{H}$,
- f is holomorphic at all cusps for Γ.

The set of modular forms of weight k with respect to Γ will be denoted by $\mathcal{M}_k(\Gamma)$.

Remark 2.4.15. The third condition means that $f[\alpha]_k$ is holomorphic at ∞ for all $\alpha \in \mathrm{SL}_2(\mathbb{Z})$. It suffices to check this for all α in some set of coset representatives.

Since the third condition in Definition 2.4.14 seems like a difficult one to prove, given a function f, we will state the following result.

Proposition 2.4.16. *Let Γ be a congruence subgroup of level N, and let f be a function satisfying the following conditions:*

- *f is weakly modular of weight k with respect to Γ,*
- *f is holomorphic on $\mathcal{H}$,*
- *f is holomorphic at ∞.*

Let the Fourier expansion (as in (5)) of f be given by

$$f(\tau) = \sum_{n=0}^{\infty} a_n q_N^n, \qquad \text{where } q_N = \exp \tfrac{\tau}{N}.$$

Then the following are equivalent:

- *f is a modular form with respect to Γ;*
- *$f[\alpha]_k$ is holomorphic at ∞ for all $\alpha \in \mathrm{SL}_2(\mathbb{Z})$;*
- *there exist constants $C, r \in \mathbb{R}_{>0}$ such that for all $n > 0$ the following inequality holds:*
$$|a_n| \leq Cn^r.$$

Proof. The first two are equivalent by definition. The third implies the second by Prop. 1.2.4 of [3], while the converse follows from section 5.9 of the same book. □

Definition 2.4.17. Let k, Γ and f be as above such that f is a modular form of weight k with respect to Γ. If furthermore $f[\alpha]_k$ has a zero at ∞ for all $\alpha \in \mathrm{SL}_2(\mathbb{Z})$, then f is called a *cusp form of weight k with respect to* Γ. We write $\mathcal{S}_k(\Gamma)$ for the set of all such forms.

Once again, we only have to check that $f[\alpha]_k$ has a zero at ∞ for all α in a set of coset representatives.

Remark 2.4.18. Note that, in contrary to the $\mathrm{SL}_2(\mathbb{Z})$ case, one might not have $\left(\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right) \in \Gamma$. This suggests that nonzero modular forms of odd weight k with respect to Γ may well exist.

It turns out that indeed, nonzero modular forms of odd weight k with respect to certain congruence subgroups Γ do exist. See section 3.6 of [3] for details.

Definition 2.4.19. We will once again write $\mathcal{M}(\Gamma)$ for the graded $\mathbb{C}$-algebra

$$\mathcal{M}(\Gamma) = \bigoplus_{k \in \mathbb{Z}} \mathcal{M}_k(\Gamma).$$

It has a homogeneous ideal given by

$$\mathcal{S}(\Gamma) = \bigoplus_{k \in \mathbb{Z}} \mathcal{S}_k(\Gamma).$$

Finally, we will produce a family of examples of modularity with respect to congruence subgroups.

Example 2.4.20. Let N be a positive integer. To provide an example of a modular form with respect to $\Gamma_0(N)$, consider the weight 2 Eisenstein series

$$G_2(\tau) = \sum_{c \in \mathbb{Z}} \sum_{\substack{d \in \mathbb{Z}: \\ (c,d) \neq (0,0)}} \frac{1}{(c\tau + d)^2}.$$

Since this series converges only conditionally, we must be careful in which order the summation is carried out. In fact, although G_2 is holomorphic on $\mathcal{H}$, conditional convergence keeps it from being modular, and we have

$$G_2[\gamma]_2(\tau) = G_2(\tau) - \frac{2\pi i c}{c\tau + d}$$

for all $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$.

However, there is still hope, since $G_2 \circ N \colon \tau \mapsto G_2(N\tau)$ can be shown to satisfy the very similar relation

$$(G_2 \circ N)[\gamma]_2(\tau) = (G_2 \circ N)(\tau) - \frac{2\pi i c}{N(c\tau + d)},$$

for all $\gamma \in \Gamma_0(N)$.

Definition 2.4.21. We will write $G_{2,N}(\tau) = G_2(\tau) - N(G_2 \circ N)(\tau)$.

Proposition 2.4.22. *The Fourier series of $G_{2,N}$ is*

$$G_{2,N}(\tau) = -\frac{\pi^2}{3}\Big((N-1) + 24\sum_{n=1}^{\infty}\Big(\sum_{\substack{d \mid n \\ N \nmid d}} d\Big)q^n\Big).$$

Proof. Firstly, by the computation of the Fourier series of G_k for $k \geq 4$, we find

$$G_2(\tau) = 2\zeta(2) - 8\pi^2\sum_{n=1}^{\infty}\sigma_1(n)q^n = \frac{\pi^2}{3} - 8\pi^2\sum_{n=1}^{\infty}\sigma_1(n)q^n.$$

If we subtract from this $N(G_2 \circ N)(\tau)$, we get

$$\begin{aligned} G_{2,N}(\tau) &= G_2(\tau) - N(G_2 \circ N)(\tau) \\ &= \frac{\pi^2}{3} - 8\pi^2\sum_{n=1}^{\infty}\sigma_1(n)q^n - N\frac{\pi^2}{3} + 8N\pi^2\sum_{n=1}^{\infty}\sigma_1(n)q^{nN} \\ &= \frac{\pi^2}{3}\Big((1-N) - 24\sum_{n=1}^{\infty}\sigma_1(n)q^n + 24\sum_{n=1}^{\infty}\sigma_1(n)Nq^{nN}\Big). \end{aligned}$$

Now observe that the contribution to q^n equals $\sigma_1(n)$ if $N \nmid n$, and it equals $\sigma_1(n) - N\sigma_1(\frac{n}{N})$ if $N \mid n$. This last expression is equal to

$$\begin{aligned} \sigma_1(n) - N\sigma_1(\tfrac{n}{N}) &= \sum_{d \mid n} d - \sum_{d \mid \frac{n}{N}} Nd \\ &= \sum_{d \mid n} d - \sum_{\substack{d \mid n \\ N \mid d}} d \\ &= \sum_{\substack{d \mid n \\ N \nmid d}} d. \end{aligned}$$

On the other hand, if $N \nmid n$, then we get

$$\sigma_1(n) = \sum_{d|n} d = \sum_{\substack{d|n \\ N\nmid d}} d.$$

Hence, we get

$$G_{2,N}(\tau) = \frac{\pi^2}{3}\Big((1-N) - 24\sum_{n=1}^{\infty}\Big(\sum_{\substack{d|n \\ N\nmid d}} d\Big)q^n\Big).$$

□

Proposition 2.4.23. *The function $G_{2,N}$ is a modular form of weight 2 with respect to $\Gamma_0(N)$.*

Proof. By definition, $G_{2,N}$ satisfies

$$G_{2,N}[\gamma]_2(\tau) = G_2(\tau) - \frac{2\pi i c}{c\tau + d} - N(G_2 \circ N)(\tau) + N\frac{2\pi i c}{N(c\tau+d)} = G_{2,N}(\tau),$$

for all $\gamma \in \Gamma_0(N)$. Hence, $G_{2,N}$ is weakly modular of weight 2 with respect to $\Gamma_0(N)$. It is holomorphic on $\mathcal{H}$ since G_2 is. Similarly, it is holomorphic at ∞.

From the previous proposition, we know that the Fourier coefficients a_n for $n > 0$ satisfy

$$|a_n| = \left|-8\pi^2 \sum_{\substack{d|n \\ N\nmid d}} d\right| \leq 8\pi^2 \sum_{d|n} d \leq 8\pi^2 \sum_{d|n} n \leq 8\pi^2 n^2,$$

so that $G_{2,N}$ is a modular form by Proposition 2.4.16. That is,

$$G_{2,N} \in \mathcal{M}_2(\Gamma_0(N)).$$

□

For $N > 1$, the Fourier series shows that $G_{2,N}$ is not a cusp form, so in particular it is nonzero. This immediately shows that even $\Gamma_0(2)$, which has index 3 in $\mathrm{SL}_2(\mathbb{Z})$, allows more modular forms than $\mathrm{SL}_2(\mathbb{Z})$, since the latter does not have any modular forms of weight 2.

Finally, we state the following fact:

Fact 2.4.24. There are no cusp forms of weight 2 with respect to $\Gamma_0(2)$.

Proof. See Exercise 3.1.4(e) and Theorem 3.5.1 of [3]. □

3 The Modularity Theorem

Now that we have defined both elliptic curves and modular forms, we can state the main theorem.

3.1 Statement of the Theorem

Recall that given an elliptic curve E over $\mathbb{Q}$ in minimal Weierstrass form, we can reduce it modulo any prime p to obtain a (possibly singular) Weierstrass curve $\tilde{E}$ over $\bar{\mathbb{F}}_p$. Associated to this is the quantity

$$a_p(E) = p + 1 - \#\tilde{E}(\mathbb{F}_p).$$

Also, given a modular form $f \in \mathcal{M}_k(\Gamma_0(N))$ of weight k with respect to the congruence subgroup $\Gamma_0(N)$, we can consider its Fourier series

$$f(\tau) = \sum_{n=0}^{\infty} a_n(f) q^n, \qquad \text{where } q = \exp \tau.$$

We see that the notation a_p is used here twice. This is no coincidence, as we come to our main theorem:

Theorem 3.1.1 (The Modularity Theorem)**.** *Let E be an elliptic curve over $\mathbb{Q}$ with conductor N_E. Then there exists a cusp form $f \in \mathcal{S}_2(\Gamma_0(N_E))$ such that $a_1(f) = 1$ and $a_p(f) = a_p(E)$ for all primes p.*

Actually, there are some other properties this f will satisfy. For instance, it is a *newform*, meaning that it does not 'come from' a modular form with respect to $\Gamma_0(M)$ for any divisor M of N_E. Furthermore, it is a common eigenvector of certain linear operators on $\mathcal{M}_2(\Gamma_0(N_E))$, called *Hecke operators*. Discussing any of these extra properties in detail will take us too far afield.

If E is an elliptic curve over $\mathbb{Q}$, we say that it is modular if a cusp form f as in the Modularity Theorem exists, and the theorem can be rephrased as: "all elliptic curves over $\mathbb{Q}$ are modular".

There are many other ways to state the Modularity Theorem, some of which can be found in [3]. Some of the formulations use analytic properties of modular curves, whereas others use more algebraic techniques like L-series or Galois representations.

The version that was finally proved (in 1995 for semistable curves by Wiles and Taylor, and in 2001 for all curves by Breuil, Conrad, Diamond and Taylor) is the one stated in Thm 9.6.2 of [3], using Galois representations.

3.2 Fermat's Last Theorem

Finally, this section will sketch the proof of Fermat's Last Theorem (FLT) from the Modularity Theorem.

Recall from section 1.4 that we had an elliptic curve E over $\mathbb{Q}$ in minimal Weierstrass form given by

$$y^2 + xy = x^3 + \frac{b^q - a^q - 1}{4}x^2 - \frac{a^q b^q}{16}x,$$

associated to the solution $a^q + b^q = c^q$ of the Fermat equation. We assumed that a, b and c are pairwise coprime, satisfying $b \equiv 0 \mod 2$ and $a \equiv -1 \mod 4$.

We computed that the discriminant of E equals

$$\Delta = 2^{-8}(abc)^{2q},$$

and the conductor is $N_E = \mathrm{rad}(abc)$.

The Modularity Theorem asserts that E is modular, i.e. there exists a cusp form f of weight 2 with respect to $\Gamma_0(N_E)$, such that $a_1(f) = 1$ and $a_p(f) = a_p(E)$, for all primes p.

We will use a simplified version of Ribet's Level Lowering Theorem. We need the following definition.

Definition 3.2.1. Let E be an elliptic curve over $\mathbb{Q}$ with minimal discriminant Δ and conductor N. Let $q > 3$ be a prime, and define

$$N_q = N \cdot \prod_{\substack{p \| N \\ q | v_p(\Delta)}} \frac{1}{p}.$$

Example 3.2.2. For the Frey curve, we compute

$$N_q = \mathrm{rad}(abc) \cdot \prod_{\substack{p \| \mathrm{rad}(abc) \\ q | v_p(2^{-8}(abc)^{2q})}} \frac{1}{p}.$$

We see that the primes dividing $\mathrm{rad}(abc)$ are exactly the primes dividing abc, and for all primes $p > 2$ we have

$$v_p(2^{-8}(abc)^{2q}) = v_p((abc)^{2q}) = 2q \cdot v_p(abc).$$

Hence, $q \mid v_p(2^{-8}(abc)^{2q})$ for all $p > 2$, so that

$$v_p(N_q) = v_p(\mathrm{rad}(abc)) - 1 = 0$$

for all primes $p > 2$. Furthermore, $v_2(N_q) = v_2(N_E) = 1$, showing that $N_q = 2$.

A condition to apply Ribet's Level Lowering Theorem is that E does not have any q-isogenies. We will not make precise what this means, but one can show that the Frey curve satisfies this condition. See [8] for details.

Definition 3.2.3. Let $N \in \mathbb{Z}_{>0}$ be an integer and $f \in \mathcal{S}_2(\Gamma_0(N))$ a cusp form with $a_1(f) = 1$. Then the *field generated by* f is the subfield of $\mathbb{C}$ generated by the $a_i(f)$.

Here is a fact that we will not prove. See section 2 of [8].

Fact 3.2.4. The field generated by f is a totally real finite extension of $\mathbb{Q}$.

Definition 3.2.5. Let E be an elliptic curve over $\mathbb{Q}$, and let $f \in \mathcal{S}_2(\Gamma_0(N))$ be as above, generating the number field K. Let q be a prime number. Then we say that E *arises modulo* q *from* f (denoted $E \sim_q f$) if there is some prime ideal $\mathfrak{q} \subseteq \mathcal{O}_K$ over q such that

$$a_p(E) \equiv a_p(f) \mod \mathfrak{q},$$

for almost all primes p.

The simplified version of Ribet's Level Lowering Theorem is the following.

Theorem 3.2.6. *Let E be a modular elliptic curve over $\mathbb{Q}$ with conductor $N = N_E$, and let q be a prime. Suppose that E does not have any q-isogenies. Then there exists a cusp form $f \in \mathcal{S}_2(\Gamma_0(N_q))$ such that $E \sim_q f$.*

Remark 3.2.7. The Level Lowering Theorem is also known as the Epsilon Conjecture (or ε-conjecture). This is because it is considered to be only a very small part of the proof of FLT, whereas the Modularity Theorem is really the hard part.

With the Level Lowering Theorem, we can finally complete the proof of Fermat's Last Theorem.

Proof of FLT. For the Frey curve, we have computed that $N_q = 2$. Ribet's Level Lowering Theorem and the Modularity Theorem now together assert that there exists a cusp form $f \in \mathcal{S}_2(\Gamma_0(2))$ such that $E \sim_q f$. But this is impossible, since there are no cusp forms of weight 2 with respect to $\Gamma_0(2)$ by Fact 2.4.24. □

References

[1] M.F. Atiyah, I.G. MacDonald, *Introduction to Commutative Algebra*. Addison-Wesley, 1969.

[2] J.E. Cremona, *Algorithms for Modular Elliptic Curves* (Second edition). Cambridge University Press, 1997.

[3] F. Diamond, J. Shurman, *A First Course in Modular Forms*. Springer, 2005.

[4] S.J. Edixhoven, L. Taelman, *Algebraic Geometry*. Leiden University, The Netherlands, Version 2011.

[5] R.E. Greene, S.G. Krantz, *Function Theory of One Complex Variable* (Third Edition). AMS, 2006

[6] R. Hartshorne, *Algebraic Geometry*. Springer, 1977.

[7] J.P. Serre, *A Course in Arithmetic*. Springer, 1973.

[8] S. Siksek, *The Modular Approach to Diophantine Equations*. Version Feb 15, 2007.

[9] J.H. Silverman, *The Arithmetic of Elliptic Curves* (Second edition). Springer, 1986.

[10] J.H. Silverman, *Advanced Topics in the Arithmetic of Elliptic Curves*. Springer, 1994.

[11] J.H. Silverman, J. Tate, *Rational Points on Elliptic Curves*. Springer, 1992.