An Introduction to Elliptic Curves, Modular Forms, and the Modularity Theorem

---

A Thesis
Presented to
The Division of Mathematical and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Blaze Okonogi-Neth

May 2023

Approved for the Division
(Mathematics)

_____

Robert Chang

Table of Contents

Abstract

This paper will develop the necessary material to state the Galois theoretic version of the modularity theorem. Chapter 1 focuses on introducing elliptic curves. We will give a basic definition, introduce the group structure, define the Tate module, and discuss reductions of curves over $\mathbb{Q}$. In Chapter 2, we motivate a certain connection between elliptic curves and modular forms by showing that all elliptic curves come with an associated lattice, and that modular forms naturally define functions on these lattices. We give a precise definition of modular forms in Chapter 3, with a discussion on the prototypical modular forms known as Eisenstein series. This allows us to explore why modular forms are of much arithmetic interest. Then, we briefly motivate the Hecke operators before defining them in Chapter 4. The Hecke operators provide a key tool in our attempt to leverage the algebraic structure to gain insight into the arithmetic of modular forms. Hecke operators also allow us to define newforms, a special type of modular forms necessary for our version of the modularity theorem. This machinery culminates in Chapter 5 with the introduction of Galois representations, a special kind of continuous homomorphism arising from the absolute Galois group. We attach a Galois representation to elliptic curves and newforms and state the Galois theoretic version of the modularity theorem, which describes when elliptic curves and newforms agree at the level of their Galois representations. We conclude by stepping through the history of the proof of Fermat's Last Theorem, demonstrating the importance of this central result.

Introduction

This thesis will attempt to explain a small piece of the striking connection between two pillars of mathematics: harmonic analysis and number theory. Harmonic analysis is the study of periodic functions—functions full of symmetry and repeating patterns—in terms of signals and waves. Number theory is the study of the integers, focused on uncovering arithmetic truths. These two parts of mathematics seem far removed from one another: The periodic functions found in harmonic analysis are continuous curves, whereas the integers are entirely discrete. However, in the last century, mathematicians have begun to uncover deep connections between the two areas. The project of understanding these connections is known as the Langlands program, and is one of the most ambitious research goals in modern mathematics. The connection we will be exploring is a small piece of the Langlands conjecture, but one that is of great historical and mathematical significance.

Starting from the side of harmonic analysis, in 1916, Srinivasa Ramanujan became interested in the function $\tau\colon \mathbb{N} \to \mathbb{Z}$ characterized by

$$\sum_{n=1}^{\infty} \tau(n) q^n = q \prod_{n=1}^{\infty} (1-q^n)^{24} =: \Delta(z) \qquad (q = e^{2\pi i z},\ \operatorname{Im} z > 0). \tag{0.1}$$

The function $\Delta(z)$ is a modular form. Modular forms are special types of function admitting a near endless amount of symmetry. These objects are so symmetric, and are so bizarre in their properties and applications, that it is surprising they exist at all. Modular forms are complex analytic objects, and what that means is that it is not enough to view the Ramanujan tau function over the real numbers. Rather, we have to input complex numbers into the function to see the incredible symmetry arise. Ramanujan was one of the first to wonder if these special functions might be connected to number theory. He did so by noticing a curious property of the τ function. Expanding the infinite product on the right-hand side of (0.1) and equating coefficients, we find

$$\sum_{n=1}^{\infty} \tau(n) q^n = q - 24q^2 + 252q^3 - 1472q^4 + 4830q^5 - 6048q^6 + \cdots . \tag{0.2}$$

Ramanujan noticed that the values of the τ function are completely determined by the values of $\tau(n)$ for prime n. For example, as 2 and 3 are prime numbers which multiply to 6, by knowing $\tau(2)$ and $\tau(3)$ we can find $\tau(6)$ by multiplication:

$$\tau(6) = \tau(2 \cdot 3) = \tau(2)\tau(3) = (-24)(252) = -6048,$$

which is indeed in agreement with (0.2). Ramanujan saw this pattern and conjectured that this multiplicative property held for all of the $\tau(n)$, but could not prove so. He made other

remarkable conjectures about this function, but they were all out of reach. It was not until 1974 that these conjectures would be proved, thank to mathematician Pierre Deligne [8] who used some of the initial ideas laid out in the Langlands program to bridge the problem from harmonic analysis to number theory.

There have been striking connections in the other direction as well, from number theory to harmonic analysis. To understand that direction we will need to introduce elliptic curves, which are a special type of polynomial function satisfying the equation $y^2 = x^3 + ax + b$. When graphed, these polynomials are smooth curves that are symmetric along the x-axis, looking something like this:

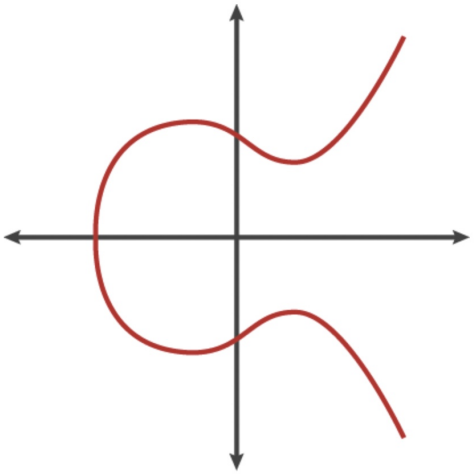


Figure 1: An Elliptic Curve

As number theorists, we are interested in the rational numbers, and so we can restrict to only look at rational points along our curve. We can use a tool from number theory to investigate these points known as modular arithmetic. Modular arithmetic is a way of viewing integers based on their remainder after division by some integer n. For example, 15 modulo 12 is 3, since after dividing 15 by 12 there are 3 left over. This is exactly how clock arithmetic works: you tell the time by working modulo 12. This is why 15:00 is 3:00 o'clock. This is important for us because some equations have no rational solutions unless they are viewed under a specific modulus. For instance, $x^2 = 5$ has no rational solutions, but $x^2 = 5 \pmod{20}$ does, since $5^2 = 25 = 5 \pmod{20}$. We call these modular solutions. We can count the number of modular solutions for our curve at each modulus n, and we denote this count by b_n. Then, we can create an infinite sequence, multiplying each b_n by a power of x and summing:

$$f(x) = \sum b_n x^n.$$

It was conjectured by Taniyama and Shimura that this equation would yield a modular form for any elliptic curve. This conjecture came to be known as the modularity theorem. This theorem is not only important to the theory, but it comes with great historical significance as well. The proof of Fermat's Last Theorem relied on a special case of the modularity theorem, which would ultimately be proved by Andrew Wiles [9].

Precisely stating a version of the modularity theorem, specifically the one proved by Wiles, will be the primary goal of this paper. Doing so will give us a glimpse of the surprising connection between two seemingly unrelated areas of math. We attempt to give a readable exposition of the main parts of the theory, providing ample examples and relegating long and technical proofs to the source material. The hope is that this paper will serve as a gentle introduction to a truly remarkable piece of mathematics.

Chapter 1

Elliptic Curves

The purpose of this chapter will be to introduce the necessary material on elliptic curves. We will give a straightforward definition and describe the group structure of this special kind of curve. Then, we will translate the curves to finite fields, which will allow us to discuss reduction of curves modulo primes. We will also look to certain well-behaved subgroups of our curves, the torsion subgroups, and use them to define the Tate module. This Tate module will be of central importance when we construct Galois representations in chapter 5. Comprehensive treatments of this material can be found in [3, 4].

1.1 Elliptic Curves

A *Diophantine equation* is a polynomial $f(x_1, ..., x_n) \in \mathbb{Z}[x_1, ..., x_n]$ with integer coefficients. Studying the solutions to these equations has been a classical point of interest in number theory. A natural question is to ask if a given Diophantine equation has integral or rational solutions. If so, the next step is to begin looking for said solutions, and ultimately determining all such solutions. For our purposes we will focus on discussing these questions with an eye towards cubic equations. Specifically, smooth cubics with at least one rational point, otherwise known as elliptic curves.

Formally, elliptic curves over a field are smooth projective algebraic curves of genus one. However, this definition is rather abstract, relying on a great deal of algebraic geometry. We will seek to simplify this definition as much as possible so that it is easier to work with.

Definition 1.1.1. A projective curve (see section A.3) $E \subseteq \mathbb{P}^2(\mathbb{Q})$ is said to be an *elliptic curve* provided that it is given by a cubic equation, contains no singularities, and has at least one rational point $O \in E(\mathbb{Q})$, which we call the *origin* of E.

In other words, an elliptic curve in the projective plane is given by a cubic polynomial with rational coefficients:

$$F(X, Y, Z) = aX^3+bX^2Y+cXY^2+dY^3+eX^2Z+fXYZ+gY^2Z+hXZ^2+jYZ^2+kZ^3 = 0$$

We can transfer this projective curve to Euclidean space by looking at the points $[X, Y, 1]$, which gives rise to the following affine curve,

$$aX^3 + bX^2Y + cXY^2 + dY^3 + eX^2 + fXY + gY^2 + hX + jY + k = 0 \qquad (1.1)$$

We must note that moving from projective space to Euclidean space comes with the risk of losing some points along the curve, specifically the points at infinity (see **??**. The next proposition gives a simplification through a change of of variables.

Proposition 1.1.2. *Let E be an elliptic curve given by (1.1) over a field K with* $\mathrm{char}(K) \neq 2, 3$. *Then there exists a curve E' of the form*

$$zy^2 = x^3 + Axz^2 + Bz^3 \quad \textit{with} \quad A, B \in K,\ 4A^3 + 27B^2 \neq 0.$$

Note the change of variables

$$\psi([X, Y, Z]) = \left[\frac{f_1(X,Y,Z)}{g_1(X,Y,Z)}, \frac{f_2(X,Y,Z)}{g_2(X,Y,Z)}, \frac{f_3(X,Y,Z)}{g_3(X,Y,Z)}\right]$$

where f_i, g_i are polynomials over K and $\psi(O) = [0, 1, 0]$, a point at infinity (see section A.3).

Note that in affine coordinates this is equivalent to

$$y^2 = x^3 + Ax + B$$

In this form, we have what is known as a Weierstrass equation. Note that we require the condition that $4A^3 + 27B^2 \neq 0$ for our curve to be smooth.

There is also a more general form of the Weierstrass equation

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

Note again that the above is an affine embedding of the curve E into $\mathbb{P}^2$. There is another form of the Weierstrass equation given by

$$y^2Z + a_1xyZ + a_3yZ^2 = x^3 + a_2x^2Z + a_4xZ^2 + a_6Z^3,$$

but we have chosen to split apart the affine part and the points at infinity by setting $Z = 1$. We can freely move between the affine and projective interpretations by multiplying through by the appropriate powers of Z. The affine case is often easier to compute with.

We say that two elliptic curves, E and E' are isomorphic if there exists an invertible change of variables $\psi\colon E \to E'$ given by rational functions over $\mathbb{Q}$ and which maps the origin of E to the origin of E'.

Example 1.1.3. Take

$$E : y^2 = x^3 + \frac{x}{2} + \frac{5}{3}.$$

Under the change of variables $x = X/6^2$ and $y = Y/6^3$, we obtain a new curve

$$E' : Y^2 = X^3 + 648X + 77760.$$

that is isomorphic to E. Note that our new curve, E', has integral coefficients.

Due to Siegel [10] we have the following result about integral points on elliptic curves.

Theorem 1.1.4 (Siegel)**.** *Let $E/\mathbb{Q}$ be an elliptic curve of the form $y^2 = x^3 + Ax + B$, with $A, B \in \mathbb{Z}$. Then E has finitely many integral points.*

While interesting, our focus will be on finding all rational points of a given curve $E : y^2 = x^3 + Ax + B$.

1.2 The Group Structure of $E(\mathbb{Q})$

We begin by defining the rational points of an elliptic curve.

Definition 1.2.1. Let the following denote the rational points of E:

$$E(\mathbb{Q}) = \{(x, y) \in E : x, y \in \mathbb{Q}\} \cup \{O\}.$$

Recall that $O = [0, 1, 0]$, the point at infinity.

One of the most crucial aspects of the theory of elliptic curves is that they can be endowed with a group structure. To do so, we must define an addition operation.

Consider a curve E given by

$$y^2 = x^3 + Ax + B$$

where $A, B \in \mathbb{Q}$. Consider two rational points P, Q along E. We can define L to be the line that intersect both P and Q. Note that we allow $P = Q$, in which case L is the tangent line at P. There must be a third rational point, R, that our line passes through because E is a cubic curve, whence

$$L \cap E = \{P, Q, R\}$$

We can then take this point R and reflect it across the X-axis thanks to the symmetry about the x-axis. We define $P + Q$ to be the point we find after reflecting R.

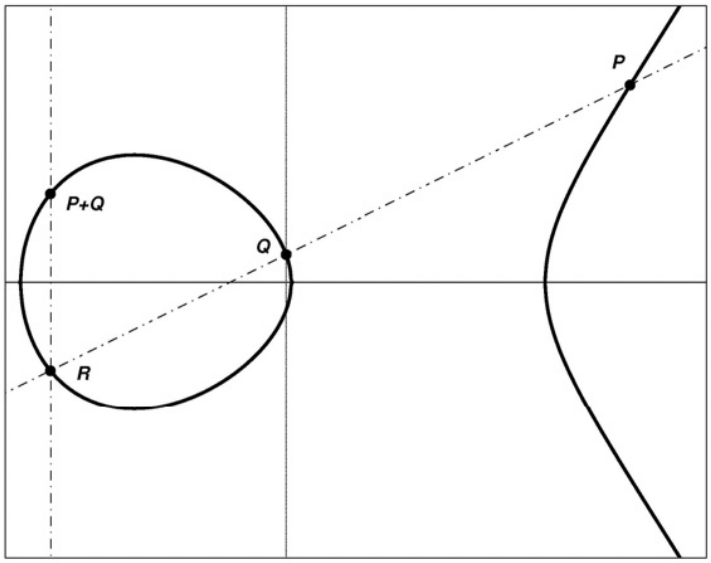


Figure 1.1: The Addition Law of Elliptic Curves

Example 1.2.2. Let E be the curve given by

$$y^2 = x^3 - 25x.$$

Recall this curve from our earlier discussion of congruent numbers. We take two points along E, $P = (5, 0)$ and $Q = (-4, 6)$, and compute $P + Q$.

We begin by finding the line L that intersects P and Q. We compute the slope

$$m = \frac{0-6}{5-(-4)} = \frac{-6}{9} = \frac{-2}{3},$$

thereby arriving at the equation

$$y = \frac{-2}{3}(x-5)$$

for L. The third point R of intersection $L \cap E$ is then obtained by solving the following system of equations

$$y = \frac{-2}{3}(x-5) \quad \text{and} \quad y^2 = x^3 - 25x.$$

Through substitution we arrive at

$$0 = x^3 - \frac{4}{9}x^2 - \frac{185}{9}x - \frac{100}{9} = (x-5)(x+4)(9x+5).$$

This third factor tells us the x-coordinate of our third point of intersection: $x = \frac{-5}{9}$. Then we plug this into our equation for E and see that $y = \frac{100}{27}$. This means that $R = \left(\frac{-5}{9}, \frac{100}{27}\right)$. Recall that we defined $P + Q$ to be the reflection of R across the x-axis, and so the final result is

$$P + Q = \left(\frac{-5}{9}, \frac{-100}{27}\right)$$

Note that we can define the addition law more generally for smooth projective cubic curves by defining the line L' that intersects R and O. Then, we define $P + Q$ as the third point of intersection of L' with E. In the case where E is given by a Weierstrass equation L' is always a vertical line, which means $P + Q$ is the reflection of R across the x-axis, which agrees with our more specific definition.

Proving that this additional law of elliptic curves defines a group operation is a fairly lengthy exercise, so we omit it. In fact, we have the following crucial theorem [11] on the structure of $E(\mathbb{Q})$.

Theorem 1.2.3 (Mordell-Weil)**.** *$E(\mathbb{Q})$ is a finitely generated abelian group. Any point $Q \in E(\mathbb{Q})$ can be written as a linear combination of*

$$Q = a_1 P_1 + ... + a_n P_n,$$

where $a_i \in \mathbb{Z}$ and $P_i \in E(\mathbb{Q})$.

This is equivalent to saying that there exists some $n \in \mathbb{N}$ and $P_1, P_2, \ldots, P_n \in E(\mathbb{Q})$ such that for all $Q \in E(\mathbb{Q})$, one can uniquely write $Q = a_1 P_1 + ... + a_n P_n$ with $a_i \in \mathbb{Z}$.

Proving this result would take us too far afield. We move to describing Elliptic curves in a different setting, which will prove to be pertinent later on.

1.3 Elliptic Curves Over Finite Fields

We can examine elliptic curves in settings other than $\mathbb{Q}$. In this section we will look at elliptic curves over finite fields.

Definition 1.3.1. Let $p \in \mathbb{Z}$ be prime. We define the finite field with $p \geq 2$ elements as follows,

$$\mathbb{F}_p = \{a(p) : a = 0, 1, 2, ..., p-1\}.$$

Curves over the field $\mathbb{F}_p$ must satisfy the same conditions as curves over $\mathbb{Q}$: they must be smooth and cubic.

Example 1.3.2. Let $p = 5$ and consider the curve

$$E : y^2 = x^3 + 1 \mod 5$$

defined over $\mathbb{F}_5$. Over the projective plane $\mathbb{P}^2(\mathbb{F}_5)$ this equation becomes

$$zy^2 = x^3 + z^3 \mod 5.$$

This is a cubic equation but we must check if it is smooth. Recall that a curve is smooth when its partial derivatives are never simultaneously 0. Direct computation shows

$$\frac{\partial F}{\partial x} = -3x^2, \quad \frac{\partial F}{\partial y} = 2yz, \quad \frac{\partial F}{\partial z} = y^2 - 3z^2 \mod 5.$$

If these partial derivatives are all 0, then $x = y = z = 0$. However, $[0, 0, 0]$ is not a point in the projective plane, and so there are no singularities and E is smooth.

Note that the same formula fails to define an elliptic curve over $\mathbb{F}_3$ because a singularity arise at $[2, 0, 1]$:

$$\frac{\partial F}{\partial x} = -3 \cdot 2^2 = 0, \quad \frac{\partial F}{\partial y} = 2 \cdot 0 \cdot 1 = 0, \quad \frac{\partial F}{\partial z} = 0^2 - 3 \cdot 1^2 = 0 \mod 3.$$

In general, by reducing the coefficients $A, B \in \mathbb{Z}$ modulo p, an elliptic curve $E/\mathbb{Q}$ with Weierstrass equation $y^2 = x^3 + Ax + B$ becomes a cubic curve E' defined over $\mathbb{F}_p$. The curve E' is not guaranteed to be smooth. But, if E' is smooth, then it is an elliptic curve. We will notice that whether or not E' is smooth depends on the choice of model for E.

Example 1.3.3. Consider the rational curve $E : y^2 = x^3 + 15625$. If we reduce our integer coefficients mod 5 we get $E' : y^2 = x^3$. We see that E' fails to be smooth since $(0, 0)$ is a singularity.

Now consider the change of variables for E given by $(x, y) \to (x^2X, 5^3Y)$. The result is a new model for the curve: $Y^2 = X^3 + 1$. Reducing this curve mod 5 does result in E' being smooth, as we showed in the previous example. Our mistake was not choosing the minimal model for our curve, a concept we will now introduce.

Definition 1.3.4. Let E be an elliptic curve given by $y^2 = x^3 + Ax + B$ with $A, B \in \mathbb{Q}$.

(1) The discriminant of E is defined by

$$\Delta_E = -16(4A^3 + 27B^2).$$

(2) Let S be the set of curves E' isomorphic to the curve $E/\mathbb{Q}$. Additionally, require $\Delta_{E'}$ to be an integer. Then, the minimal discriminant of E is the smallest, in absolute value, $\Delta_{E'}$ over the set S. We call E' the minimal model for E when it gives rise to the minimal discriminant of E.

Example 1.3.5. The curve $E : y^2 = x^3 + 5^6$ has discriminant $\Delta_E = -2^4 3^3 5^{12}$ and the curve $E' : y^2 = x^3 + 1$ has discriminant $\Delta_{E'} = -2^4 3^3$. As we saw in the last example, these curves are isomorphic through a change in variables. Since $\Delta_E > \Delta_{E'}$, we know that Δ_E cannot be minimal. In fact, $\Delta_{E'}$ is the minimal discriminant, but showing that fact is a bit involved.

To determine rather or not a given discriminant is minimal one must check that the valuation of the discriminant at each prime is less than 12. This is because the various discriminants can only differ by multiples of 12. This is discussed in chapter 7 of [4].

1.4 The Torsion Subgroup

To better understand the group structure $E(\mathbb{Q})$, a natural place to start is to look for well-behaved subgroups. In our case, these are the torsion subgroups of E.

Definition 1.4.1. A torsion point of an elliptic curve $E(\mathbb{Q})$ is any point $P \in E(\mathbb{Q})$ with finite order. That is, $nP = O$ for some $n \in \mathbb{N}$. Denote by

$$E(\mathbb{Q})_{\text{torsion}} = \{P \in E(\mathbb{Q}) : \text{ there is } n \in N \text{ such that } nP = O\}$$

the set of torsion points, which forms a finite group.

Theorem 1.4.2. *Let E be a non-singular cubic curve.*

$$E : Y^2 = f(x) = x^3 + ax^2 + bx + c, (a, b, c \in \mathbb{Q})$$

The following hold:

(1) A point $P = (x, y) \neq O$ on E is torsion of order 2 if and only if $y = 0$.

(2) The curve E has either 0, 1, or 3 points of order 2.

Proof. To prove (1) note that a point being of order 2 is to say that $2P = O$, or equivalently $P = -P$. In coordinate form, $(x, y) = (x, -y)$, which only holds when $y = 0$. For the other direction, if we assume that y is 0, then $P = -P$, and we are done.

For (2), notice that (1) tells us that there are as many points of order 2 as there are rational roots of $f(x) = x^3 + ax^2 + bx + c$, a degree 3 polynomial. If there are 2 roots, $\alpha, \beta \in \mathbb{Q}$, then the third root γ must also be in $\mathbb{Q}$, since $\alpha + \beta\gamma = -a$. □

Example 1.4.3. Consider the curve $E_n : y^2 = x^3 - n^2x = x(x-n)(x+n)$. Now, E_n has three rational points, $P = (0,0)$, $Q = (-n,0)$, and $T = (n,0)$. These points are all of order 2. That is, $2P = 2Q = 2T = O$, and $P + Q = T$. Since each of these points add to the identity the torsion subgroup is isomorphic to Klein-4.

$$E_n(\mathbb{Q})_{\text{torsion}} = \{O, P, Q, T\} \cong \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$$

Since the group $E(\mathbb{Q})$ is finitely generated and abelian it follows that $E(\mathbb{Q})_{\text{torsion}}$ is finite. Mazur's theorem [12] classifies the possible isomorphism classes of $E(\mathbb{Q})_{\text{torsion}}$,

Theorem 1.4.4 (Mazur's Theorem)**.** *Let $E/\mathbb{Q}$ be an elliptic curve. Then, $E(\mathbb{Q})_{torsion}$ is isomorphic to exactly one of the following groups:*

$$\mathbb{Z}_N \text{ with } N \in \{1, 2, 3, \ldots, 10, 12\}; \text{ or}$$
$$\mathbb{Z}_2 \oplus \mathbb{Z}_{2M} \text{ with } M \in \{1, 2, 3, 4\}.$$

There are infinitely many elliptic curves with torsion subgroups isomorphic to each of the groups in Mazur's theorem. Naturally, the next step is to begin computing torsion subgroups. Mazur's theorem tells us that if a P is a rational point along the elliptic curve E, and P has order greater than or equal to 12, $E(\mathbb{Q})$ is infinite. The following theorem [13] gives us more criteria for computing $E(\mathbb{Q})_{\text{torsion}}$.

Theorem 1.4.5 (Nagell-Lutz)**.** *Let $E/\mathbb{Q}$ be an elliptic curve with Weierstrass equation*

$$y^2 = x^3 + Ax + B, \ A, B \in \mathbb{Z}.$$

Then, every torsion point $P \neq O$ of E satisfies:

(1) The coordinates of P are integers;

(2) If P is of order $n \geq 3$ then $4A^3 + 27B^2$ is divisible by $y(P)^2$;

(3) If P is of order 2 then $y(P) = 0$ and $x(P)^3 + Ax(P) + B = 0$.

Example 1.4.6. Consider the following curve,

$$E : y^2 = x^3 - 2$$

Here, $A = 0$ and $B = -2$. We will use Nagell-Lutz to find the points of finite order and determine the torsion subgroup $E(\mathbb{Q})_{\text{torsion}}$.

By theorem 1.4.2 we know that the points of order 2 are exactly the zeroes of $x^3 - 2$. Our only candidate then is $(\sqrt[3]{2}, 0)$, but $\sqrt[3]{2}$ is not an integer, and so by Nagell–Lutz is not a torsion point. We move to check for points of order $n \geq 3$.

We begin by compiling a list of rational points $P = (x, y)$ whose square divides $4A^3 + 27B^2 = 108$. Through brute force calculation we can find the divisors of 108,

$$y \in \{1, 2, 3, 4, 6, 9, 12, 18, 27, 36, 54, 108\}$$

Narrowing this list down to the squares gives us the possible values of y,

$$y \in \{1, 2, 3, 6\}$$

We can now calculate the x coordinate for each possible value of y^2, remembering that x must be an integer.

$$\begin{aligned} 1 = x^3 - 2 &\iff x\ \sqrt[3]{3} \notin \mathbb{Z} \\ 4 = x^3 - 2 &\iff x\ \sqrt[3]{6} \notin \mathbb{Z} \\ 9 = x^3 - 2 &\iff x\ \sqrt[3]{11} \notin \mathbb{Z} \\ 36 = x^3 - 2 &\iff x\ \sqrt[3]{36} \notin \mathbb{Z} \end{aligned}$$

There are no valid pairs, and so the curve E has no non-trivial points of finite order, and so $E(\mathbb{Q})_{\text{torsion}} = \{O\}$

With these examples under our belt we would like to give a more general treatment of the torsion subgroups. Ultimately, this will allow us to define the Tate module of an elliptic curve, which will be crucial when constructing representations in our final chapter. We define the following map,

Definition 1.4.7. Consider some elliptic curve E and let P be a point along E. We can define the following multiplication-by-m map,

$$\begin{aligned} [m] &: E \to E \\ [m]P &= P + ... + P \end{aligned}$$

Note that the group operation is repeated m times.

Then the $m-$torsion subgroups of a curve E is simply the kernel of the multiplication-by-m map. We have the following definition,

Definition 1.4.8. Let E be an elliptic curve and let K be a field.

$$E[m] := \ker [m] = \{P = (x, y) \in \overline{K}^2 : P \in E, [m]P = O\}$$

Note the specialization of our coordinates to the algebraic closure of K. This is because the structure of $E[m]$ becomes much simpler when we do so. We have the following theorem on the structure of $E[m]$.

Theorem 1.4.9. *Let E be in elliptic curve and let $m \in \mathbb{Z}$. If $m \neq 0$ in the field K, which is to say that either K is of characteristic 0 or char$(k) > 0$ and char(K) does not divide m, then*

$$E[m] = \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/m\mathbb{Z}$$

Proving this theorem in generality requires more background than we are able to develop. However, the $m = 2$ case is straightforward, and relies on no more than the ideas developed in the proof of theorem 7. We give one more brief example.

Example 1.4.10. Consider the elliptic curve over F_3 given by $E = y^2 = x^3 + x$. Now, $E(F_3) \cong \mathbb{Z}/4\mathbb{Z}$, and this can be computed much the same as in our previous examples. This implies the existence of only one rational 2-torsion point, $(0, 0)$. To uncover the other 2-torsion points we have to lift up to $\overline{F_3}$. Examining our polynomial,

$$x^3 + x = x(x^2 + 1),$$

we see that the roots are 0 and $\sqrt{-1}$. Of course -1 is not a square working mod 3, and so we take an element $i \in \overline{F_3}$ such that $i^2 = -1$. This gives us four torsion points of order 2: $O, (0, 0), (i, 0), (-i, 0)$. Of course, this means that $E[2] \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$.

We would like to be able to study all of the m-torsion subgroups simultaneously. To do so we will construct what is known as the Tate module. One can reference example A.2.4 in the appendix for a more detailed account of a similar construction.

Definition 1.4.11. Let E be an elliptic curve and let $l \in \mathbb{Z}$ be a prime. The $(l-adic)$ Tate module of E is the group

$$Ta_l(E) = \varprojlim_n E[l^n]$$

where the inverse limit is being taken with respect to the natural maps

$$E[l^{n+1}] \to E[l^n]$$

This is possible because the l^n torsion groups form an inverse system with maps given by multiplication by some power of l. Each $E[l^n]$ is a $\mathbb{Z}/l^n\mathbb{Z}$-module and so the Tate module has the structure of a Z_l-module. We will use the Tate module to construct representations attached to elliptic curves in the final chapter of this paper.

1.5 The Reduction of Ellipitic Curves over $\mathbb{Q}$

We will now characterize the types of singularities a curve can have. This will ultimately let us classify the various types of reductions of a curve. Let E' be a cubic curve over some field K with Weierstrass equation $f(x, y) = 0$. That is,

$$f(x, y) = y^2 + a_1xy + a_3y - x^3 - a_2x^2 - a_4x - a_6$$

Suppose that $P = (x_0, y_0)$ is a singular point of E'. We can compute the taylor expansion of $f(x, y)$ centered at (x_0, y_0).

$$\begin{aligned} f(x, y) - f(x_0, y_0) &= \lambda_1(x - x_0)^2 + \lambda_2(x - x_0)(y - y_0) + \lambda_3(y - y_0)^2 - (x - x_0)^3 \\ &= ((y - y_0) - \alpha(x - x_0)) \cdot ((y - y_0) - \beta(x - x_0)) - (x - x_0)^3 \end{aligned}$$

where $\lambda_i, \alpha, \beta \in \overline{K}$.

Definition 1.5.1. A node is a singular point P such that $\alpha \neq \beta$. This implies the existence of two separate tangent lines,

$$\begin{aligned} y - y_0 &= \alpha(x - x_0) \\ y - y_0 &= \beta(x - x_0) \end{aligned}$$

A cusp is a singular point P such that $\alpha = \beta$. This implies the existence of a unique tangent line at P.

Definition 1.5.2. Let $E/\mathbb{Q}$ be an elliptic curve given by the minimal model. Consider the reduction E' of E at each prime $p \geq 2$. If E' defines an ellipitic curve over $\mathbb{F}_p$ then we say E has a **good reduction** mod p. If E' contains a singular point $P \in \mathbb{F}_p$ then E has a **bad reduction** mod p. If P is a cusp of E' then E has **additive (unstable) reduction**. On the other hand, if P is a node of E' then E has **multiplicative (semi-stable) reduction**. In this case, if both tangent lines at P have slopes in $\mathbb{F}_p$, the reduction is split multiplicative, and non-split otherwise.

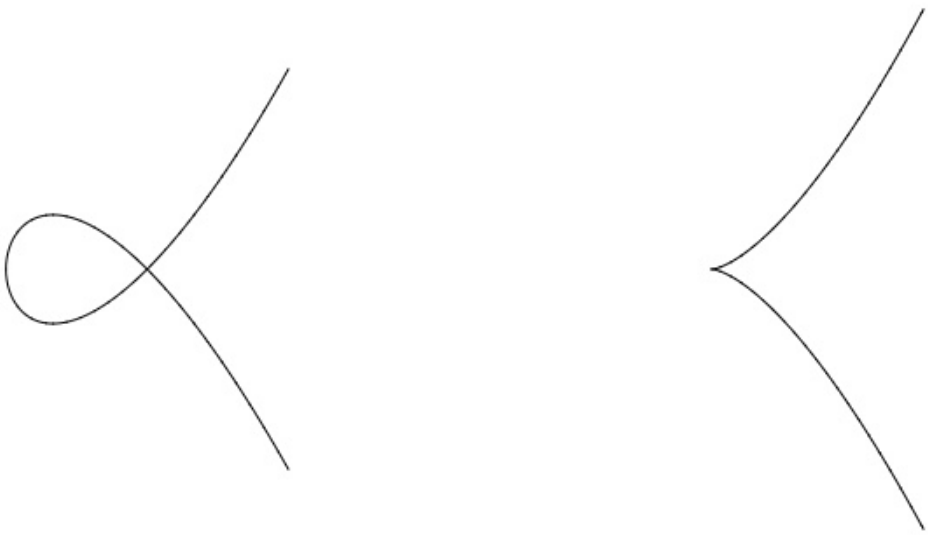


Figure 1.2: Example of a node and a cusp, adapted from [1]

The following proposition gives a condition for curves over an arbitrary field to be non-singular.

Proposition 1.5.3. *Consider the cubic curve $E/K : y^2 = f(x)$ where K is a field and $f(x) \in K[x]$ is monic and cubic. Let $\alpha, \beta, \gamma \in \overline{K}$ and suppose that $f(x) = (x-\alpha)(x-\beta)(x-\gamma)$. Then, consider the discriminant of $f(x)$,*

$$D = (\alpha-\beta)^2(\alpha-\gamma)^2(\beta-\gamma)^2$$

Then, E is nonsingular if and only if $D \neq 0$

Proof. To prove, begin by supposing that E is singular at the point $P = (x, y)$. We will show that this implies that Δ is 0. Following the definition,

$$\begin{aligned} f(x,y) &= y^2 - (x^3 + ax + b) = 0 \\ \frac{\partial f}{\partial x}(x,y) &= -(3x^2 + a) = 0 \\ \frac{\partial f}{\partial y}(x,y) &= 2y = 0 \end{aligned}$$

The above partial derivatives force $a = -3x^2$ and $y = 0$. Plugging these into $f(x, y) = y^2 - (x^3 + ax + b)$ tells us that $b = 2x^3$. Now, calculating the discriminant,

$$\begin{aligned}\Delta &= 4a^3 + 27b^2 \\ &= 4(-3x^2)^3 + 27(2x^3)^2 \\ &= 0\end{aligned}$$

For the other direction, suppose that Δ is 0. We will show that this forces E to be singular. Note that $\Delta = 0$ if and only if the $x^3 + ax + b$ has a double root at x. A double root occurs at x if and only if its derivative is $3x^2 + a$. That is,

$$\begin{aligned}f(x, 0) = 0^2 - (x^3 + ax + b) = 0 \\ \frac{\partial f}{\partial x}(x, y) = -(3x^2 + a) = 0 \\ \frac{\partial f}{\partial y} = 2(0) = 0\end{aligned}$$

Which shows $(x, 0)$ to be a singular points. This finishes the proof: E is nonsingular if and only if $\Delta \neq 0$.

□

Note that Δ_E is a multiple of D. Specifically, $\Delta_E = 16D$. This along with the above proposition gives us the following result,

Corollary 1.5.4. *Let $E/\mathbb{Q}$ be an elliptic curve with integer coefficients. Let p be any prime. If E has a bad reduction at p then $p|\Delta_E$. If we take a minimal model of E then $p|\Delta_E$ if and only if E has a bad reduction at p.*

We would like to be able to say something globally about a curve's reduction at a given prime. To do so, we can package all of the local reduction into a single invariant.

Definition 1.5.5. For all primes $p \in \mathbb{Z}$ define the following quantity

$$e_p = \begin{cases} 0 & \text{if } E \text{ has a good reduction at } p, \\ 1 & \text{if } E \text{ has multiplicative reduction at } p, \\ 2 & \text{if } E \text{ has additive reduction at } p \text{ and } p \neq 2, 3, \\ 2 + \delta_p & \text{if } E \text{ has additive reduction at } p \text{ and } p = 2 \text{ or } 3. \end{cases}$$

The conductor N of $E/\mathbb{Q}$ is

$$N = \prod_p p^{e_p},$$

where the product ranges over all primes p.

The conductor will appear as the level of newforms in the modularity theorem. In the case we're interested in, the modularity theorem for semi-stable elliptic curves, the conductor is square-free.

We give one more example before concluding our section on elliptic curves. This will illustrate a method of counting points along a curve. Let E' be an elliptic curve over $\mathbb{F}_{p^r}$. Of course, $E'(\mathbb{F}_{p^r}) \subseteq \mathbb{P}^2(\mathbb{F}_{p^r})$. The projective plane over $\mathbb{F}_{p^r}$ has $(p^r)^2 + p^r + 1$ many points, and so the number of points along E' over $\mathbb{F}_{p^r}$, $|E'(\mathbb{F}_{p^r})|$, is finite. The following theorem [3] gives a bound for this quantity,

Theorem 1.5.6. *(Hasse) Let E' be an elliptic curve over F_{p^r}, then*

$$p^r + 1 - 2\sqrt{p^r} < |E'(\mathbb{F}_{p^r})| < p^r + 1 + 2\sqrt{p^r}.$$

Example 1.5.7. Consider the curve $E : y^2 = x^3 + 3$. The minimal discriminant of E is $\Delta_E = -3888 = -2^4 \cdot 3^5$. It follows that the only primes of bad reduction are 2 and 3. It must be that $E'/\mathbb{F}_p$ is smooth for all $p \geq 5$. We will compute the points on $E(\mathbb{F}_5)$ taking the naive approach. The possible values of x are 0, 1, 2, 3, and 4. We can calculate the quadratic residues of $x \mod 5$ for each values. This gives the following set of residues,

$$0^2 = 0(5), \quad 1^2 = 1(5), \quad 2^2 = 4(5), \quad 3^2 = 4(5), \quad 4^2 = 1(5).$$

Then, plugging each value of x into $x^3 + 3$ to find the necessary values of y^2 yields

$$0^3 + 3 = 3(5), \quad 1^3 + 3 = 4(5), \quad 2^3 + 3 = 1(5), \quad 3^3 + 3 = 0(5), \quad 4^3 + 3 = 2(5).$$

We can then look up, using the quadratic residues, which values of y achieve these values of y^2 working mod 5. This gives the following set of points,

$$\{O, (1,2), (1,3), (2,1), (2,4), (3,0)\}.$$

And so, $|E'(\mathbb{F}_{p^r})| = 6$. We can plug this into Hasse's bound and see that matches the given range,

$$1.5278 = 5 + 1 - 2\sqrt{5} < 6 < 5 + 1 + 2\sqrt{5} = 10.4721.$$

Our next goal will be to explore elliptic curves defined over the complex numbers. This will motivate the theory of modular curves, which will ultimately serve as a bridge to modular forms.

Chapter 2

Modular Curves

This chapter will seek to connect elliptic curves to modular curves, which will ultimately lead to modular forms. We will view elliptic curves over $\mathbb{C}$ and note that they emit a lattice structure. After examining these lattices we will describe the complex upper half plane under the action of $\mathrm{SL}_2(\mathbb{Z})$ by fractional linear transformation, which will give us the modular group. Then, we will define modular curves and certain congruence subgroups. This material can be found in more detail in [1] and [3].

2.1 Lattices

We begin with periodic functions. A function $f\colon \mathbb{C} \to \mathbb{C}$ is said to be *periodic with respect to* ω provided there exists a nonzero $\omega \in \mathbb{C}$ such that

$$f(z+\omega) = f(z) \quad \text{for all } z \in \mathbb{C}.$$

Note that if ω is a period, then $n\omega$ is a period for all $n \in \mathbb{Z}$. Moreover, if ω_1 and ω_2 are periods, then $n\omega_1 + m\omega_2$ is a period for all $n, m \in \mathbb{Z}$.

Definition 2.1.1. A function f is said to be *doubly periodic* if it has two periods, ω_1 and ω_2, whose ratio is not real.

Remark. Suppose f is double periodic and the ratio ω_1/ω_2 of the two periods is real. There are two cases: Either the ratio is rational, or it is irrational. In the former case, both ω_1 and ω_2 are integer multiples of some other period ω. In the latter case, one can show that f has arbitrarily small periods, thereby implying that f is constant on all of the open connected sets on which it is analytic.

Definition 2.1.2. Let f be a double periodic function with periods ω_1 and ω_2. The pair (ω_1, ω_2) is said to be the *fundamental pair* or *lattice basis* if every period of f can be written as a linear combination of ω_1 and ω_2 with integer coefficients.

A fundamental pair (ω_1, ω_2) generates the *lattice*

$$L = \langle \omega_1, \omega_2 \rangle := \{n\omega_1 + m\omega_2 \in \mathbb{C} : n, m \in \mathbb{Z}\}.$$

For simplicity, we henceforth require that ω_1/ω_2 have positive imaginary part, that is, the ratio belongs to the upper half complex plane

$$H = \{a + bi \in \mathbb{C} : b > 0\}.$$

Two complex numbers z_1 and z_2 are said to be *equivalent modulo* L provided there exists $\omega \in L$ such that $z_1 - z_2 = \omega$. The quotient group $\mathbb{C}/L$ is the set of equivalence classes of $\mathbb{C}$ modulo L.

Definition 2.1.3. Let $L = \langle \omega_1, \omega_2 \rangle$ be a lattice. The *fundamental domain* of $\mathbb{C}/L$ is the parallelogram

$$F = \{\lambda\omega_1 + \mu\omega_2 : 0 \leq \lambda, \mu < 1\}.$$

Example 2.1.4. The Gaussian integers $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$ is a lattice generated by $w_1 = i$ and $w_2 = 1$. The fundamental domain of $\mathbb{C}/\mathbb{Z}[i]$ is the square

$$F = \{\lambda i + \mu : 0 \leq \lambda, \mu < 1\}.$$

Since, when working modulo $\mathbb{Z}[i]$, we have $\lambda i = \lambda i + 1$ and $\mu = \mu + i$ for all $\lambda, \mu \in \mathbb{R}$, the two opposing sides of the square F are identified to make F a torus.

Definition 2.1.5. Two fundamental pairs (ω_1, ω_2) and (ω_1', ω_2') are said to be *equivalent* exactly when they generate the same lattice, that is, when $\langle \omega_1, \omega_2 \rangle = \langle \omega_1', \omega_2' \rangle$.

Proposition 2.1.6. *Two fundamental pairs (ω_1, ω_2) and (ω_1', ω_2') are equivalent if and only if there is a 2×2 matrix with integer entries and with determinant equal to ± 1 such that*

$$\begin{pmatrix} \omega_2' \\ \omega_1' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \omega_2 \\ \omega_1 \end{pmatrix}$$

This is equivalent to the conditions

$$\begin{aligned} \omega_2' &= a\omega_2 + b\omega_1, \\ \omega_1' &= c\omega_2 + d\omega_1. \end{aligned}$$

Proof. Let A and B be our two lattice basis and assume that our two pairs generate the same lattice, $\langle \omega_1, \omega_2 \rangle = \langle \omega_1', \omega_2' \rangle$. By definition there exists integer square matrices, V and W, such that,

$$\begin{aligned} A &= BV \\ B &= AW \end{aligned}$$

We can rewrite this equation as,

$$\begin{aligned} A &= AWV \\ A(I - VW) &= 0 \end{aligned}$$

Of course, B is a lattice basis, and thus linear independent. This implies that $VW = I$. Calculating the determinant and noting that since V and W are integer matrices we have $\det(V), \det(W) \in \mathbb{Z}$ we get,

$$\det(V) \cdot \det(W) = \det(V \cdot W) = \det(I) = 1$$
$$\det(V) = \det(W) = \pm 1$$

For the other direction, assume that $A = BU$ where A, B are lattice basis and U is a unimodular, integer matrix. Then, U^{-1} exists and is also a unimodular integer matrix. It follows that,

$$A = BU$$
$$B = AU^{-1}$$

and thus that $\langle \omega_1, \omega_2 \rangle \subseteq \langle \omega_1', \omega_2' \rangle$ and $\langle \omega_1', \omega_2' \rangle \subseteq \langle \omega_1, \omega_2 \rangle$. It follows that $\langle \omega_1, \omega_2 \rangle \sim \langle \omega_1', \omega_2' \rangle$. □

The following tells us when quotients of $\mathbb{C}$ by different lattices are isomorphic to one another.

Proposition 2.1.7. *Let L and L' be lattices and let $z \in \mathbb{C}$. The quotients $\mathbb{C}/L$ is isomorphic to $\mathbb{C}/L'$ if and only if $L = zL'$.*

These propositions combine to give the following result,

Corollary 2.1.8. *If L and L' are lattices and there exists an analytic isomorphism such that $\mathbb{C}/L \cong \mathbb{C}/L'$ then there is some $\alpha \in \mathbb{C}^\times$ and $M \in \mathrm{SL}_2(\mathbb{Z})$ such that*

$$\begin{pmatrix} \omega_1' \\ \omega_2' \end{pmatrix} = \alpha M \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}.$$

Note that for any lattice $L = \langle \omega_1, \omega_2 \rangle$ with an oriented basis (that is, $\omega_1/\omega_2 \in H$) there is always another lattice $L' = \langle z, 1 \rangle$ such that $\mathbb{C}/L \cong \mathbb{C}/L'$. This is because we can always set $z = \omega_1/\omega_2$ and apply Proposition 2.1.6.

We can consider the case of $\mathbb{C}/\langle z, 1 \rangle \cong \mathbb{C}/\langle z', 1 \rangle$. By Corollary 2.1.8 we know there must be a matrix $M \in \mathrm{SL}_2(\mathbb{Z})$ and some $\alpha \in \mathbb{C}^\times$ such that

$$\begin{pmatrix} z' \\ 1 \end{pmatrix} = \alpha \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} z \\ 1 \end{pmatrix} = \begin{pmatrix} \alpha(az+b) \\ \alpha(cz+d) \end{pmatrix}.$$

Direct calculations yield $1 = \alpha(cz + d)$ and $\alpha = (cz + d)^{-1}$, which imply

$$z' = \frac{az+b}{cz+d} \qquad (ad - bc = 1).$$

These conclusions lead to the following proposition.

Proposition 2.1.9.

(1) *There is some $z \in H$ such that $\mathbb{C}/L \cong \mathbb{C}/\langle z, 1\rangle$*

(2) *Let $z, z' \in H$. Then, $\mathbb{C}/\langle z, 1\rangle \cong \mathbb{C}/\langle z', 1\rangle$ if and only if there exists $M \in \mathrm{SL}_2(\mathbb{Z})$ such that $z' = Mz = (az + b)/(cz + d)$.*

2.2 Elliptic Functions

Now we define elliptic functions on $\mathbb{C}/L$.

Definition 2.2.1. A function f is said to be *elliptic* provided

(1) f is doubly periodic;

(2) f is meromorphic.

Recall that a function f is meromorphic when its only singularities in the finite plane are poles. Constant functions are trivial examples of elliptic functions but the most important example is the Weierstrass $\wp$-function. Details of the construction are omitted as they are quite involved.

Definition 2.2.2. Let L be a lattice. The *Weierstrass $\wp$-function* relative to L is given by

$$\wp(z, L) = \frac{1}{z^2} + \sum_{0 \neq w \in L} \left(\frac{1}{(z - w)^2} - \frac{1}{w^2} \right).$$

The Laurent series of this function is also quite important, and for this we need the Eisenstein series.

Definition 2.2.3. Let L be a lattice and let $k \geq 2$. The *weight $2k$ Eisenstein series* with respect to L is given by

$$G_{2k}(L) = \sum_{0 \neq w \in L} \frac{1}{w^{2k}}.$$

We can now express the Laurent series for the Weierstrass $\wp$-function in terms of Eisenstein series.

Theorem 2.2.4. *Let L be a lattice.*

(1) *The Laurent series for the Weierstrass $\wp$-function is given by*

$$\wp(z, L) = \frac{1}{z^2} + \sum_{k=1}^{\infty} (2k + 1) G_{2k+2}(L) z^{2k}.$$

(2) *For all $z \in \mathbb{C}$ and $z \notin L$, the derivative $\wp'$ of $\wp$ is given by*

$$\left(\frac{\wp'(z, L)}{2} \right)^2 = \wp(z, L)^3 - 15 G_4(L) \wp(z, L) - 35 G_6(L).$$

It follows that $(\wp(z, L), \frac{1}{2}\wp'(z, L))$ is a point on the elliptic curve

$$E_L(\mathbb{C}) : y^2 = x^3 - 15 G_4(L) x - 35 G_6(L). \tag{2.1}$$

Thanks to the theorem, with $E_L(\mathbb{C})$ as above we have the map

$$\Phi\colon \mathbb{C}/L \to E_L(\mathbb{C}), \quad z \bmod L \mapsto (\wp(z,L), \tfrac{1}{2}\wp'(z,L)). \tag{2.2}$$

The following theorem asserts important properties about this map.

Theorem 2.2.5 (Uniformization Theorem)**.**

(1) *Let L be a lattice and let $E_L(\mathbb{C})$ be the elliptic curve as in (2.1). Then Φ as in (2.2) is a complex analytic isomorphism of abelian groups.*

(2) *Consider an elliptic curve $E/\mathbb{Q} : y^2 = x^3 + Ax + B$. There exists a lattice L over $\mathbb{C}$ such that $A = -15G_4(L)$, $B = -35G_6(L)$, and Φ an isomorphism between $\mathbb{C}/L$ and $E(\mathbb{C})$.*

These results take a substantial amount of work to prove, and so we direct the interested readers to section 1.4 of [1]. We settle for giving an overview of the main ideas. Theorem 2.2.5 states that every lattice L determines an elliptic curve $E_L/\mathbb{C}$ and every elliptic curve $E/\mathbb{C}$ has an associated lattice such that $E(\mathbb{C}) \cong \mathbb{C}/L$. We established earlier that every $z \in H$ determined a lattice, and thus $E(\mathbb{C}) \cong \mathbb{C}/\langle z, 1\rangle$. However, this z is only unique up to the actions of $M \in \mathrm{SL}_2(\mathbb{Z})$. This motivates an equivalence relation for elements of H modulo $\mathrm{SL}_2(\mathbb{Z})$.

2.3 The Modular Group

We will now define the modular group. Let H refer to the upper half plane of $\mathbb{C}$. Let $\mathrm{SL}_2(\mathbb{R})$ be the group of matrices with real entries and determinant equal to 1. Take $g \in \mathrm{SL}_2(\mathbb{R})$ and $z \in \mathbb{C}$. Define the group action of $\mathrm{SL}_2(\mathbb{R})$ on $\mathbb{C}$ by

$$gz = \frac{az+b}{cz+d}.$$

We now show that H is stable under this action by computing $\mathrm{Im}(gz)$ directly:

$$\begin{aligned}
\mathrm{Im}(gz) &= \frac{1}{2i}\left(\frac{az+b}{cz+d} - \frac{a\overline{z}+b}{c\overline{z}+d}\right) \\
&= \frac{1}{2i}\frac{(az+b)(c\overline{z}+d) - (a\overline{z}+b)(cz+d)}{|cz+d|^2} \\
&= \frac{(ad-bc)\mathrm{Im}(z)}{|cz+d|^2} \\
&= \frac{\mathrm{Im}(z)}{|cz+d|^2}.
\end{aligned}$$

It is simple to see that

$$-1 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R})$$

acts trivially on H. Thus, H is acted on by the *projective special linear group* $\mathrm{PSL}(2,\mathbb{R}) = \mathrm{SL}_2(\mathbb{R})/\{\pm 1\}$.

Definition 2.3.1. The group $\Gamma(1) = \mathrm{SL}_2(\mathbb{Z})/\{\pm 1\}$ is known as the modular group. Note that the modular group is the image of $\mathrm{SL}_2(\mathbb{Z})$ in $\mathrm{PSL}(2,\mathbb{R})$.

The modular group induces an equivalence relation on elements of the upper half-plane. We denote the set of equivalence classes by

$$Y(1) := H/\Gamma(1) = \frac{\{z \in H\}}{\{z \sim z' \text{ if and only if } z' = Mz \text{ for some } M \in \mathrm{SL}_2(\mathbb{Z})\}}.$$

Topologically, $Y(1)$ is homeomorphic to a punctured sphere. We can compactify it by "adding a point at infinity" to obtain $Y(1) \cup \{\infty\}$; see [1, Chapter 3] for more details.

Let $S, T \in \Gamma(1)$ be given by

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

It is easy to check for $z \in H$ that

$$Sz = \frac{-1}{z} \quad \text{and} \quad Tz = z + 1.$$

Next, define the set

$$D = \{z \in H : |\mathrm{Re}\, z| \leq \tfrac{1}{2},\ |z| \geq 1\}.$$

Theorem 2.3.2.
1. The modular group, $\Gamma(1)$ is generated by S and T.
2. The fundamental domain of $\Gamma(1)$ is given by D. That is, the canonical map $D \to H/G$ is surjective and its restriction to the interior of D is injective.

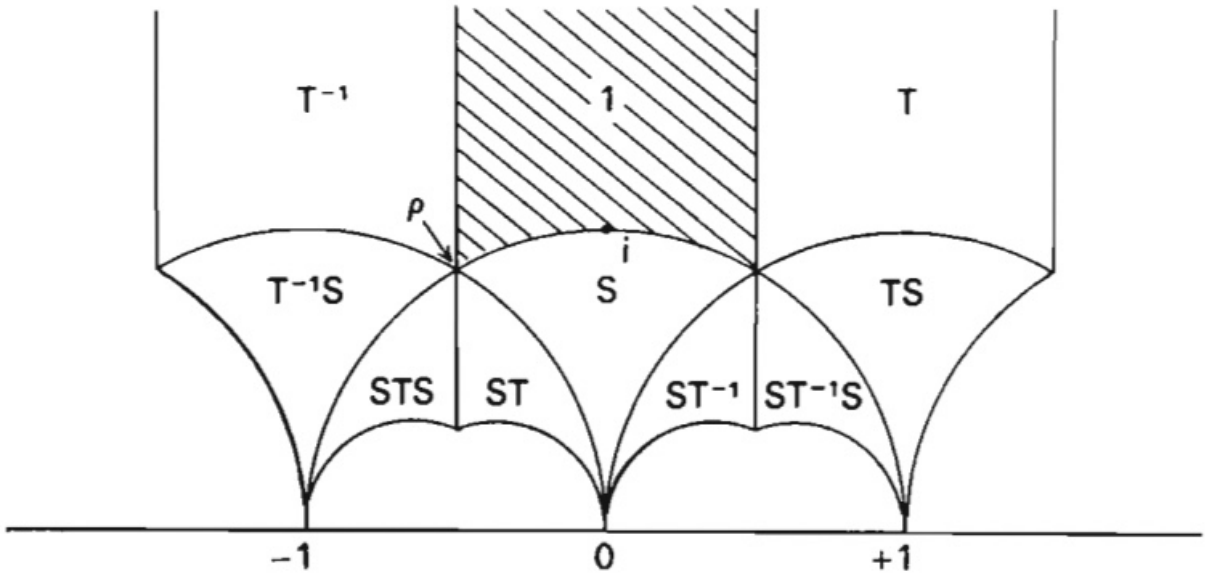


Figure 2.1: The Fundamental Domain of the Modular Group, adapted from [2]

The proof of 2.3.2 can be found in chapter 7 of [2].

2.4 Modular Curves and Congruence Subgroups

It is desirable to work with an expanded version of H. We extend the upper half plane by adjoining the projective line over $\mathbb{Q}$ to obtain

$$H^* := H \cup \mathbb{P}^1(\mathbb{Q}).$$

Recall that $\mathbb{P}^1(\mathbb{Q})$ is made up of points $[s, 1]$, with $s \in \mathbb{Q}$ the rational points along the real axis of the complex plane, and the point at infinity, $\infty := [1, 0]$.

We furthermore extend the action of $\Gamma(1)$ to H^*. Given $M \in \mathrm{SL}_2(\mathbb{Z})$ and $[s, t] \in P^1(\mathbb{Q})$, define

$$M[s, t] = -M[s, t] := [as + bt, cs + dt] \in P^1(\mathbb{Q}).$$

We can now define $X(1)$, the equivalence classes of $H^*/\Gamma(1)$ under this extended action,

$$X(1) = H^*/\Gamma(1) = \frac{\{z \in H\} \cup \{s \in \mathbb{Q}\} \cup \{\infty\}}{\{z \sim z' \text{ if and only if } z' = Mz \text{ for some } M \in \mathrm{SL}_2(\mathbb{Z})\}} \tag{2.3}$$

It can be shown that given any $s, s' \in \mathbb{P}^1(\mathbb{Q})$, there exists a matrix $M \in SL(2, \mathbb{Z})$ such that $s' = Ms$. Therefore, all points in $\mathbb{P}^1(\mathbb{Q})$ are equivalent to the point ∞ in $H^*/\Gamma(1)$, and so $X(1)$ contains exactly one more point than $Y(1)$. We call the point at infinity a *cusp*.

To generalize the notion of $X(1)$, we now introduce certain subgroups of $\mathrm{SL}_2(\mathbb{Z})$.

Definition 2.4.1. Let $N \geq 1$ be an integer. We define the following subgroups of $\mathrm{SL}_2(\mathbb{Z})$:

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{Z}) : c = 0 \bmod N \right\},$$
$$\Gamma_1(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) : c = 0,\ a = d = 1 \bmod N \right\},$$
$$\Gamma(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) : b = c = 0,\ a = d = 1 \bmod N \right\}.$$

A subgroup G of $\mathrm{SL}_2(\mathbb{Z})$ is a congruence subgroup when G contains $\Gamma(N)$ for some N. We call N the *level* of a given congruence subgroup.

The definitions immediately imply $\Gamma(N) \subset \Gamma_1(N) \subset \Gamma_0(N)$. Also, note that $\Gamma(1) = \mathrm{SL}_2(\mathbb{Z})$, which matches our previous notation.

It remains to generalize $X(1)$ to obtain what are known as *modular curves*, denoted X. This involves the straightforward replacement of $\Gamma(1)$ with an arbitrary congruence subgroup Γ in (2.3) to obtain

$$X = H/\Gamma = \frac{\{z \in H\} \cup \{s \in \mathbb{Q}\} \cup \{\infty\}}{\{z \sim z' \text{ if and only if } z' = Mz \text{ for some } M \in \Gamma\}}. \tag{2.4}$$

Now, the cusps in X are the elements that have a representative in $\mathbb{P}^1(\mathbb{Q})$. Alternatively, we have the following definition.

Definition 2.4.2. Let Γ be a congruence subgroup and consider $\mathbb{Q}^* = \coprod_i \Gamma(s_i)$ where each s_i is a distinct element of $\mathbb{Q}^*$. The s_i are the cusps of Γ.

While $X(1)$ had only a single cusp, it is possible for modular curves to have multiple. However, a congruence subgroup only ever has finitely many cusps. We can prove this using the following proposition.

Proposition 2.4.3. *A congruence subgroup of level N has a finite index in $SL(2,\mathbb{Z})$.*

Proof. Consider the following reduction mod N map,

$$\phi : SL(2,\mathbb{Z}) \to SL(2,\mathbb{Z}/N\mathbb{Z}).$$

The kernal of this map is $\Gamma(N)$ and so,

$$SL(2,\mathbb{Z})/\Gamma(N) \cong Im(\phi) < SL(2,\mathbb{Z}/N\mathbb{Z})$$

Since $SL(2,\mathbb{Z}/N\mathbb{Z})$ is a finite group $\Gamma(N)$ has finite index in $SL(2,\mathbb{Z})$. We know that $\Gamma(N) < \Gamma$, and the result follows. □

From the above lemma and definition 2.4.3 we know that $SL(2,\mathbb{Z})$ can be written as $\coprod_{i=1}^{n} \Gamma s_i$ where $n, s_i \in SL(2,\mathbb{Z})$. Then,

$$\mathbb{Q}^* = SL(2,\mathbb{Z})(\infty) = \bigcup_{i=1}^{n} \Gamma s_i(\infty).$$

It follows that Γ has at most n cusps, and so we have the result.

Chapter 3

Modular Forms

In this chapter we will give our first definition—one that will arise naturally from a connection to certain lattice functions—of modular forms. Then, we will give the prototypical example of a modular form, the Eisenstein series. The Eisenstein series lend themselves to explicit computation, and so we will briefly derive their q-expansions. This will uncover some of the arithmetic hidden in modular forms and motivate their connection to number theory. We will conclude with an informal, historically motivated discussion of the Hecke operators. More details on the subjects of this chapter can be found in [1, 5, 3].

3.1 Modular Forms

Using what we have developed for lattices we would like to motivate a certain class of functions for the modular group. Let F be a complex-valued function on the set of lattices over $\mathbb{C}$. We will impose a simple transformation law under multiplication on F.

Definition 3.1.1. Let $k \in \mathbb{Z}$. A complex-valued function F is said to be *of weight* $2k$ provided

$$F(\lambda\tau) = \lambda^{-2k} F(\tau)$$

holds for all lattices τ and all $\lambda \in \mathbb{C}^\times$.

For some normalized (ω_1, ω_2) we let $F(\omega_1, \omega_2)$ denote the value of F on the lattice $\tau\langle\omega_1, \omega_2\rangle$. Now, we can rewrite the expression from Definition 3.1.1 as a lattice function

$$F(\tau\omega_1, \tau\omega_2) = \lambda^{-2k} F(\omega_1, \omega_2). \tag{3.1}$$

Let $\lambda = \omega_2$ and recall from the previous section that any lattice $\langle\omega_1, \omega_2\rangle$ is equivalent up to homothety to a lattice $\langle z, 1\rangle$, where $z = \frac{\omega_1}{\omega_2} \in H$. Equation (3.1) shows us that the product $\omega_2^{2k} F(\omega_1, \omega_2)$ only depends on the choice of $z = \omega_1/\omega_2$. We then know that there exists a function f on H such that

$$F(\omega_1, \omega_2) = \omega_2^{-2k} f\left(\frac{\omega_1}{\omega_2}\right).$$

Taking F to be invariant under $\mathrm{SL}_2(\mathbb{Z})$ gives us

$$f(z) = (cz+d)^{-2k} f\left(\frac{az+b}{cz+d}\right) \text{ for all } z \in H \text{ and } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}),$$

which gives us what are known as *modular functions*. The relationship holds in the other direction as well, that is, we have a correspondence between modular functions of weight $2k$ and some lattice functions of weight $2k$. Now we can give our first definition of modular forms.

Definition 3.1.2. Let $k \in \mathbb{Z}$. A *modular form of weight k for* $\mathrm{SL}_2(\mathbb{Z})$ is a function $f\colon H \to \mathbb{C}$ satisfying three conditions.

(1) f is holomorphic;

(2) The modularity condition: $f(\frac{az+b}{cz+d}) = (cz+d)^k f(z)$ for all $z \in H$.

(3) $f(z)$ remains bounded as $\mathrm{Im}(z) \to +\infty$.

The third condition is important because it guarantees holomorphy at infinity. We care about the behavior at infinity since, as we mentioned in the previous chapter, our fundamental domain is not compact until we add in the point at infinity. This growth condition ensures that our modular forms behaves well under compactification. The second of these conditions is known as the modularity condition, and we see it arise from this connection to lattice functions. We require that it holds for an infinite set of matrices. We can see a couple of examples of what this looks like. Let us look at the following elements of $\mathrm{SL}_2(\mathbb{Z})$:

$$T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad U = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

For the matrix T, our condition means that

$$f(z+1) = f(z).$$

Having $c = 0$ and $d = 1$ means the factor in the condition goes away.

The matrix S gives us

$$f\left(\frac{-1}{z}\right) = z^k f(z).$$

Here, $\frac{-1}{z}$ is the transformation that flips the inside and outside of the unit semi circle; the negative sign ensures we stay in the upper half plane.

The matrix U has entries $b = c = 0$ and $a = d$, which leads to

$$f(z) = (-1)^k f(z).$$

Now, if the the weight k is odd, the function must be the zero function. With our definition, there are no modular forms of odd weight aside from the zero function. However, is possible to define modular forms over other groups, and if the group does not contain the negative of the identity matrix then you can have odd weight modular forms.

It turns out that the modularity condition is preserved under multiplication of matrices that satisfy it. This means that since we noted generators of the modular group in the last section, checking that a function satisfies the modularity condition boils down to checking the action of only two matrices.

3.2 Eisenstein Series

Our first example of a modular form will be the Eisenstein series. Recall that we have already defined the Eisenstein series as functions of lattices in Definition 2.2.3. We will take this initial definition of G_{2k} and evaluate it at the lattice $\langle z, 1\rangle$ with $z \in H$.

Definition 3.2.1. Let $k \geq 2$. The *weight* $2k$ *Eisenstein series* is

$$G_{2k}(z) := G_{2k}(\langle z, 1\rangle) = \sum_{\omega \in \langle z,1\rangle, \omega \neq 0} \frac{1}{\omega^{2k}} = \sum_{(m,n)\in\mathbb{Z}^2,(m,n)\neq 0} \frac{1}{(mz+n)^k}.$$

Note that going forward we will consider weight k Eisenstein series and simply take k to be even. To check that this is a modular form we would need to show that it is holomorphic on the upper half plane, that it satisfies

$$G_k(z+1) = G_k(z) = (-1)^k G_k(z),$$

and that it is bounded as $z \to +\infty$. We will not prove all of these facts, but we will comment on a small bit of the analysis that is going on in the background. Our main purpose will be to use Eisenstein series to show why modular forms are of number theoretic interest.

The first thing to note is that the Eisenstein series converges absolutely on compact subsets of the upper half plane, and by the Weierstrass theorem that is enough to imply holomorphy. Absolute convergence together with the change of variables $(m, n) \mapsto (m, m + 2n)$ also proves the first modularity condition:

$$G_k(z+1) = \sum_{(m,n)\neq(0,0)\in\mathbb{Z}^2} \frac{1}{(mz+n+m)^k} = \sum_{(m,n)\neq(0,0)\in\mathbb{Z}^2} \frac{1}{(mz+n)^k} = G_k(z).$$

Now, speaking more generally, if $f(z + 1) = f(z)$ holds, then f is periodic with period 1. This means we can express f as a function of $q = e^{2\pi iz}$. Call this representation $\tilde{f}$. If it is given that f is holomorphic on H it follows that $\tilde{f}$ is meromorphic on the punctured unit disk, $|q| < 1$. Consider the conformal mapping, $z \to e^{2\pi iz}$ that takes f to $\tilde{f}$. If we consider the mapping of the point $i\infty$ we see that it goes to 0. This is why if $\tilde{f}$ extends to a meromorphic function at the origin we say that f is meromorphic at infinity. The upshot is that in such a case $\tilde{f}$ gives rise to a Laurent expansion at some neighborhood around the origin,

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

Every modular form can be written as a power series in this way. This is known as the q-expansion of a modular form and is one of the primary ways of pulling arithmetic information out of modular forms. We will look at the q-expansion of Eisenstein series and see how much richer the representation is.

We take our original definition of Eisenstein series, split off the terms where $n = 0$, and do some algebraic manipulation to arrive at,

$$G_k(z) = \sum_{(m,n)\neq(0,0)\in\mathbb{Z}^2} \frac{1}{(mz+n)^k} = 2\sum_{n\geq 1} \frac{1}{n^k} + 2\sum_{n\geq 1}\left(\sum_{m\in\mathbb{Z}} \frac{1}{(mz+n)^k}\right) \tag{3.2}$$

We would like to get from here to the q-expansion. Isolating the terms that do not involve z gives us the constant term

$$a_0 = 2\sum_{n\geq 1}\frac{1}{n^k} = 2\zeta(k),$$

where ζ is the Riemann zeta function. The higher order coefficients are not values of the zeta function but are related to more number theoretic types of divisor sums.

To continue with the q-expansion computation we want to write out the power series in $e^{2\pi iz}$. The idea will be to rewrite the inner sum as

$$\sum_{n\in\mathbb{Z}}\frac{1}{(w+n)^k}$$

note that mz is fixed and so we write it as $w \in H$. This function is a periodic function with respect to w. It follows from a Fourier series expansion (details omitted) that for every point $w \in H$ and $k \geq 3$, we have

$$\left(\sum_{n\in\mathbb{Z}}\frac{1}{(w+n)^k}\right) = \frac{(-2\pi i)^k}{(k-n)!}\sum_{n\geq 1} n^{k-1}e^{2\pi inw}$$

This gives us an expansion of the periodic function in w as a function of $e^{2\pi inw}$.

We take our original equation (3.2) for the Eisenstein series and replace the inner sum with our new formula. We set $w = m\tau$, an integer multiple of a point in the upper half plane to obtain

$$\begin{aligned} G_k(z) &= 2\zeta(k) + \frac{2(2\pi i)^k}{(k-n)!}\sum_{m,n\geq 1} n^{k-1}e^{2\pi i(nm)z} \\ &= 2\zeta(k) + \frac{2(2\pi i)^k}{(k-n)!}\sum_{r\geq 1}\left(\sum_{d|r} d^{k-1}\right)e^{2\pi i(r)z}. \end{aligned}$$

Set

$$\sigma_{k-1}(n) = \sum_{d|r} d^{k-1} \quad \text{and} \quad q^n = e^{2\pi i(r)z},$$

then the expression above takes on the familiar form

$$G_k(z) = 2\zeta(k) + \frac{2(2\pi i)^k}{(k-n)!}\sum_{n\geq 1}\sigma_{k-1}(n)q^n. \tag{3.3}$$

This is the q-expansion of the Eisenstein series. We see values of the zeta function and divisor sums appear in this expanded form. This is one of the reasons that modular forms are of interest in number theory. The coefficients of the q-expansions for some examples turn out to be of number theoretic interest. On a slightly higher level, this is one of the reasons modular forms are so powerful. They are analytic objects that contain a striking amount of arithmetic information. Tools in analysis are powerful, where as tools in arithmetic are much less so. Modular forms give a vehicle for powerful analytic techniques to be applied to arithmetic questions. We can continue to push this example by using one of Euler's formulas for values of the zeta function.

Theorem 3.2.2 (Euler)**.** *For $k \geq 2$ even, we have*

$$\begin{aligned}\zeta(k) &= \frac{(-1)^{\frac{k}{2}+1} 2(\pi)^k B_k}{k!2} \\ , &= \frac{-(2\pi i)^k B_k}{k!2}\end{aligned}$$

where B_k is the kth Bernoulli number and

$$\frac{x}{e^x - 1} = \sum_{k \geq 0} B_k \frac{x^k}{k!} = 1 - \frac{x}{2} + \frac{x^2}{12} - \frac{x^4}{720} + \dots.$$

We continue the manipulation of (3.3) to obtain

$$G_k(z) = 2\zeta(k) - \frac{4k\zeta(k)}{B_k} \sum_{n \geq 1} \sigma_{k-1}(n) q^n$$

A modular form is still a modular form if you multiply or divide by a nonzero constant. We might normalize a modular form by dividing by the leading constant term so that it is normalized to 1. One way to do this here is to divide through by $2\zeta(k)$. This gives us what is called the *normalized weight k Eisenstein series*:

$$E_k(z) = \frac{G_k(z)}{2\zeta(k)} = 1 - \frac{2k}{B_k} \sum_{n \geq 1} \sigma_{k-1}(n) q^n.$$

Now, it is not obvious from this representation that we have a modular form. Writing modular forms in this expanded way obscures some of that information. We can check that this is holomorphic with respect to z, which is hiding in q, and we can see that it is periodic with respect to z, but there is no obvious way to check the second transformation law. However, this is just the cost of all the other information we have already gained.

Here are the q-expansions for some of the low weight Eisenstein series:

$$\begin{aligned}E_4 &= 1 + 240q + 2160q^2 \cdots = 1 + 240 \sum_{n \geq 1} \sigma_3(n) q^n, \\ E_6 &= 1 - 504q + 16632q^2 \cdots = 1 - 504 \sum_{n \geq 1} \sigma_5(n) q^n, \\ E_8 &= 1 + 480q + 61920q^2 \cdots = 1 + 480 \sum_{n \geq 1} \sigma_7(n) q^n.\end{aligned}$$

More generally, every E_k can be expressed as a polynomial in E_2 and E_3. The main takeaway of this work with Eisenstein series is the amount of rich arithmetic information contained inside of modular forms.

3.3 The Space of Modular Forms

Perhaps the most crucial aspect of this theory is that the space of weight k modular forms gives rise to a finite dimensional vector space. In fact, this is another reason we require precisely those three conditions in our definition. For instance, it is the boundedness condition

that keeps our vector space finite dimensional. This structure and the tools of linear algebra will allow us to continue pulling out arithmetic information from these analytic objects. In the next chapter we will define specific operators on the space of modular forms, known as Hecke operators, to gain some amount of control over it. First, we will introduce the context which led to the initial development of these operators, which will serve as some amount of motivation.

We will start by establishing two useful pieces of notation.

Definition 3.3.1. The set of modular forms with respect to Γ is denoted by $M_k(\Gamma)$ and the set of cusp forms with respect to Γ is denoted $S_k(\Gamma)$.

In this section we will see another important property of Eisenstein series: They are often helpful in constructing other functions.

Definition 3.3.2. Consider the weight k Eisenstein series $G_k(z)$. Set $g_2 = 60G_2$ and $g_3 = 140G_3$. Define

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

Δ is a weight-12 modular form. Importantly, the space of weight-12 modular forms is a one dimensional vector space.

A modular form is said to be a *cusp form* if it has a value of 0 at infinity. It is straightforward to verify that Δ is a cusp form. Indeed, our previous work with Eisenstein series shows

$$g_2(\infty) = 120\zeta(4) = \frac{4}{3}\pi^4 \quad \text{and} \quad g_3(\infty) = 280\zeta(6) = \frac{8}{27}\pi^6,$$

from which we deduce $\Delta(\infty) = 0$. It turns out that Δ is not identically 0 (see [1]. This shows that Δ is indeed a nontrivial cusp form. This function arises in the study of elliptic curves as the discriminant. A natural expansion of Δ is given in Serre [2],

$$\Delta = g_2^3 - 27g_3^2 = (2\pi)^{12}2^{-6}3^{-3}(E_2^3 - E_3^2) = (2\pi)^{12}(q - 24q^2 + 252q^3 - 1472q^4 + ...).$$

This expansion serves to motivate the theorem given to us by Jacobi, whose proof can be found in [2, Chapter 7].

Theorem 3.3.3 (Jacobi)**.** *We have*

$$\Delta = (2\pi)^{12}q\prod_{n=1}^{\infty}(1-q^n)^{24}.$$

We can write

$$\sum_{n=1}^{\infty}\tau(n)q^n = q\prod_{n=1}^{\infty}(1-q^n)^{24},$$

in which $\tau(n)$ is the Ramanujan τ-function, for which Ramanujan conjectured the following properties.

(i) $|\tau(p)| \leq 2p^{11/2}$.

(ii) $\tau(mn) = \tau(m)\tau(n)$ if $\gcd(m, n) = 1$.

(iii) $\tau(p^{n+1}) = \tau(p)\tau(p^n) - p^{11}\tau(p^{n-1})$ when p is prime and $n \geq 1$.

We can associate the Dirichlet series

$$L(\Delta, s) = \sum \tau(n) n^{-s}$$

to Δ. Properties (i) and (ii) are equivalent to stating that this Dirichlet series emits a certain Euler product:

$$L(\Delta, s) = \prod_{p \text{ prime}} \frac{1}{1 - \tau(p)p^{-s} + p^{11-2s}}. \tag{3.4}$$

Milne delivers a proof of this equivalence in chapter 5 of [5].

Now, we might ask if we can construct operators that act on the space of modular forms under the $\mathrm{SL}_2(\mathbb{Z})$ action that satisfy the same recurrence relations as the Ramanujan τ-function. We leave the details, as well the definition of the inner product, for the next chapter. For now, we merely look at what some of the consequences might be if we were able to construct operators that look as follows,

$$T(n) \colon M_k \to M_k$$

Endow this operator with the following properties.

(a) $T(m) \circ T(n) = T(mn)$ if $\gcd(m, n) = 1$.

(b) $T(p) \circ T(p^n) = T(p^{n=1}) + p^{2k-1}T(p^{n-1})$ for prime p.

(c) $T(n)$ preserves the space of cusp forms and is Hermitian, that is, $\langle T(n)f, g\rangle = \langle f, T(n)g\rangle$ where f and g are cusp forms.

Since $T(n)$ forms a family of commuting operators, there exist joint eigenforms of $T(n)$ with real eigenvalues that form an orthogonal basis for the space of cusp forms. From here, Milne presents a result that relates eigenforms of $T(n)$ to modular forms.

Theorem 3.3.4 (Milne)**.** *Let $f(z) = \sum c(n)q^n$ be a weight $2k$ modular form that is not identically zero. If f is an eigenforms for all $T(n)$ then $c(1) = 0$, and rescaling the function so that $c(1) = 1$ results in $T(n)f = c(n)f$.*

A function rescaled in this way, to have a constant coefficient of 1 in its Fourier series, is known as a *normalized eigenform*. It follows that if we have a normalized eigenform f for all of the $T(n)$, then the eigenvalues $c(n)$ would enjoy properties that are strikingly similar to Ramanujan's conjectured properties of the τ-function. Indeed, if we take the Dirichlet series and the associated Euler product of f we see that

$$L(f, s) = \sum c(n)n^{-s} = \prod \frac{1}{1 - c(p)p^{-s} + p^{2k-1-2s}},$$

which matches (3.4).

Recall that the space of cusp forms of weight 12 is one dimensional. This implies that Δ is an eigenfunction for all $T(n)$, since each $T(n)$ takes cusp forms of weight 12 to cusp forms of weight 12. That is, properties (ii) and (iii) of the Ramanujan function will follow from the existence of operators with the above properties. We will move to construct these operators shortly, but first we require a slightly more sophisticated definition of modular forms. We introduce a key piece of notation.

Definition 3.3.5. For $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ and $k \in \mathbb{Z}$ define the weight-k operator $[\gamma]_k$ on functions $f\colon H \to \mathbb{C}$ by

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

Now, recalling the congruence subgroups developed in Chapter 2.4, we can define modular forms that have their actions restricted to certain subgroups of $\mathrm{SL}_2(\mathbb{Z})$.

Definition 3.3.6. Let Γ be a congruence subgroup of $SL(2, \mathbb{Z})$ and let $k \in \mathbb{Z}$. A function $f\colon H \to \mathbb{C}$ is said to be a *modular form with respect to* Γ provided

(1) f is holomorphic;

(2) f is weight-k invariant under Γ, that is, $f(z) = f[\Gamma]_k(z)$;

(3) $f[\alpha]_k$ is holomorphic at ∞ for all $\alpha \in \mathrm{SL}_2(\mathbb{Z})$.

With this definition we are able to give the proper definition of a crucial class of modular forms, cusp forms.

Definition 3.3.7. A cusp form of weight-k with respect to Γ to be a function satisfying the above conditions with $a_0 = 0$ in the Fourier expansion of $f[\alpha]_k$ for all $\alpha \in SL(2, \mathbb{Z})$. Note that this is an equivalent definition to the form vanishing at infinity.

We are now ready to develop operators on the space $M_k(\Gamma)$ of modular forms, with the goal of uncovering a basis for $S_k(\Gamma_1(N))$, the space of cusp forms.

Chapter 4

Hecke Operators

In the previous chapter we primarily worked with Eisenstein series. These modular forms were easy to work with, as we saw, but cusp forms are not as easy to write down. We cannot calculate them directly in the same way we calculated Eisenstein series. This is why we will need to develop more sophisticated tools in this chapter. This chapter is a summary of the material in [1, 3].

4.1 The Double Coset Operator

Our first operator of interest will be the *double coset operator*. So far, we have been working with the group $\mathrm{SL}_2(\mathbb{Z})$, but we will want to expand to a larger group, $\mathrm{GL}_2^+(\mathbb{Q})$, the group of 2×2 matrices with rational entries and positive determinants. Since congruence subgroups of $\mathrm{SL}_2(\mathbb{Z})$ are subgroups of $\mathrm{GL}_2^+(\mathbb{Q})$, we can define double cosets as follows.

Definition 4.1.1. Let Γ_1, Γ_2 be two congruence subgroups of $\mathrm{SL}_2(\mathbb{Z})$. A *double coset* in $\mathrm{GL}_2^+(\mathbb{Q})$ is of the form

$$\Gamma_1 \alpha \Gamma_2 = \{\gamma_1 \alpha \gamma_2 : \gamma_1 \in \Gamma_1, \gamma_2 \in \gamma_2\}$$

for each $\alpha \in \mathrm{GL}_2^+(\mathbb{Q})$.

The action of Γ_1 partitions the double coset into orbits through left multiplication. For a fixed coset representative $\beta = \gamma_1 \alpha \gamma_2$, its orbit is $\Gamma_1 \beta$ and the orbit space $\Gamma_1 \Gamma_1 \alpha \Gamma_2$ is a disjoint union $\bigcup \Gamma_1 \beta_j$. This union is finite, as discussed in [1, Section 5.1]. With a finite orbit space the double coset can act on the space of modular forms. First we will need to tweak our earlier definition of the weight-k operator to account for the fact that we are working over $\mathrm{GL}_2^+(\mathbb{Q})$.

Definition 4.1.2. For $\beta \in \mathrm{GL}_2^+(\mathbb{Q})$, the *weight-k operator* $[\beta]_k$ on functions $f \colon H \to \mathbb{C}$ is

$$(f([\beta]_k)(z) = (\det \beta)^{k-1}(cz + d)^{-k} f(\beta(z))$$

for $\beta \in \mathrm{GL}_2^+(\mathbb{Q})$.

With this definition we are able to define our initial coset operator on $M_k(\Gamma)$.

Definition 4.1.3. Let Γ_1, Γ_2 be congruence subgroups of $\mathrm{SL}_2(\mathbb{Z})$ and let $\alpha \in \mathrm{GL}_2^+(\mathbb{Q})$. Set $\Gamma_1\alpha\Gamma_2 = \bigcup_j \Gamma_1\beta_j$ with $\beta_j = \gamma_{1,j}\alpha\gamma_{2,j}$ and $\gamma_{i,j} \in \Gamma_i$. The *weight-k double coset operator* $[\Gamma_1\alpha\Gamma_2]\colon M_k(\Gamma_1) \to M_k(\Gamma_2)$ is

$$f[\Gamma_1\alpha\Gamma_2]_k = \sum_j f[\beta_j]_k$$

The sum is well-defined since the orbit space is finite. The fact that the image actually lands in $M_k(\Gamma_2)$ is handled in [1, Section 5.1]. To show that this operator is well-defined we must show that the action of $\Gamma_1\alpha\Gamma_2$ does not depend on the choice of orbit representative. To this end, write $\Gamma_1\alpha\Gamma_1 = \bigcup_j \Gamma_1\beta_j$ for a differing choice of orbit representative $\{\tau_j\}$. For each j there exists γ_j such that $\gamma_j\beta_j = \tau_j$. Then

$$f[\tau_j]_k = f[\gamma_j\beta_j]_k = (f[\gamma_j]_k)[\beta_j]_k = f[\beta_j]_k,$$

as desired

Example 4.1.4. Three special cases of this operator are given in [1] which will serve to illustrate the definition.

(1) If $\Gamma_2 \subset \Gamma_1$, then taking $\alpha = I$ implies that $\Gamma_1\alpha\Gamma_2 = \Gamma_1$, and so $f[\Gamma_1\alpha\Gamma_2]_k = f[I]_k = f$. This means the natural inclusion from $M_k(\Gamma_1)$ to $M_k(\Gamma_2)$ is an injection.

(2) Suppose that $\Gamma_1 \subset \Gamma_2$. Let $\alpha = I$ and let $\{\gamma_{2,j}\}$ be a set of coset representatives of $\Gamma_1\Gamma_2$. Then, $f[\Gamma_1\alpha\Gamma_2]_k = \sum_i f[\gamma_{2,i}]_k$, the trace map, is a surjection from $M_k(\Gamma_1)$ to $M_k(\Gamma_2)$.

(3) If $\alpha^{-1}\Gamma_1\alpha = \Gamma_2$ then $f[\Gamma_1\alpha\Gamma_2]_k = f[\alpha]_k$ since $\Gamma_1\alpha\Gamma_2 = \Gamma_1\alpha$. This gives an isomorphism from $M_k(\Gamma_1)$ to $M_k(\Gamma_2)$. The inverse map is $([\Gamma_2\alpha^{-1}\Gamma_1])_k = [\alpha^{-1}]_k$.

4.2 The Diamond Operator

We will now use define more specific operators. Specifically, we will use the double operators to define the Hecke operators on $M_k(\Gamma_1(N))$.

The map

$$\Gamma_0 \to (\mathbb{Z}/N\mathbb{Z})^\times, \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto d \bmod N$$

defines a surjective homomorphism. This implies that $\Gamma_1(N)$ is normal in $\Gamma_0(N)$ and induces an isomorphism from $\Gamma_0(N)/\Gamma_1(N) \to (\mathbb{Z}/N\mathbb{Z})^\times$.

Now, let $\Gamma_1 = \Gamma_2 = \Gamma_1(N)$ and take $\alpha \in \Gamma_0(N)$. This is case (3) from Example 4.1.4. In particular, for $f \in M_k(\Gamma_1(N))$, we have $f[\Gamma_1\alpha\Gamma_2]_k = f[\alpha]_k$, which again lands in $M_k(\Gamma_1(N))$. That is, we have an action of $\Gamma_0(N)$ on $M_k(\Gamma_1(N))$ taking f to $f[\alpha]_k$. Since the subgroup $\Gamma_1(N)$ acts trivially this is an action of $\Gamma_0(N)/\Gamma_1(N) \cong (\mathbb{Z}/N\mathbb{Z})^\times$ on $M_k(\Gamma_1(N))$. This action defines our next operator.

Definition 4.2.1. The *diamond operator* $\langle d\rangle\colon M_k(\Gamma_1(N)) \to M_k(\Gamma_1(N))$ is defined by

$$\langle d\rangle f = f[\alpha]_k.$$

Here, $d \in (\mathbb{Z}/N\mathbb{Z})^\times$ and

$$\alpha = \begin{pmatrix} a & b \\ c & \delta \end{pmatrix} \in \Gamma_0(N) \text{ with } \delta = d \bmod N.$$

This is the first *Hecke operator*. We can extend the definition to be for all positive integers n.

Definition 4.2.2. Let $n \in \mathbb{Z}^+$ such that n and N are coprime. Then, $n \bmod N \in (\mathbb{Z}/N\mathbb{Z})^\times$, and so the diamond operator, $\langle n\rangle\colon M_k(\Gamma_1(N)) \to M_k(\Gamma_1(N))$, can be defined according to this reduction. If n and N are not coprime, then we define $\langle n\rangle = 0$.

We can use Dirichlet characters to understand the eigenvalues of this operator. Let χ be a Dirichlet character modulo N. Consider $\gamma \in \Gamma_0(N)$ and let d_γ mark the lower-right entry of γ. This way, $f[\gamma]_k = \langle d_\gamma\rangle f$ for all $f \in M_k(\Gamma_1(N))$.

Definition 4.2.3. With notation as above, the *χ-eigenspace of $M_k(\Gamma_1(N))$* is denoted by

$$M_k(N,\chi) = \{f \in M_k(\Gamma_1(N)) : \langle d_\gamma\rangle f = \chi(d_\gamma) f \text{ for all } \gamma \in \Gamma_0(N)\}.$$

More explicitly, $M_k(N,\chi))$ gathers modular forms which are eigenvectors for the diamond operator with eigenvalues χ. This works because the diamond operator respects the decomposition $M_k(\Gamma_1(N)) = \bigoplus_\chi M_k(N,\chi)$. This decomposition is sensible because $\Gamma_1(N)$ is a normal subgroup of $\Gamma_0(N)$, and by construction $M_k(\Gamma_1(N))$ is a module over the quotient $\Gamma_0(N)/\Gamma_1(N)$. Any representation over $\mathbb{C}$ of a finite abelian group, in this case of $\Gamma_0(N)/\Gamma_1(N)$, can be written as a direct sum of one-dimensional representations of the finite abelian group. The one dimensional representations of $\Gamma_0(N)/\Gamma_1(N)$ are given by the characters χ. We refer the readers to chapters 4 and 5 of [1] for more details.

4.3 The T_p Operator

We can define the second type of Hecke operator, again using the double coset operator. More precisely, in Definition 4.1.1 we take $\Gamma_1 = \Gamma_2 = \Gamma_1(N)$ as before, but choose a special α.

Definition 4.3.1. For p prime, define operators $T_p\colon M_k(\Gamma_1(N)) \to M_k(\Gamma_1(N))$ by

$$T_p f = \begin{cases} \displaystyle\sum_{j=0}^{p-1} f\left[\begin{pmatrix} 1 & j \\ 0 & p \end{pmatrix}\right]_k & \text{if } p \mid N, \\ \displaystyle\sum_{j=0}^{p-1} f\left[\begin{pmatrix} 1 & j \\ 0 & p \end{pmatrix}\right]_k + f\left[\begin{pmatrix} m & n \\ N & p \end{pmatrix}\begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix}\right]_k & \text{if } p \nmid N, \text{ where } mp - nN = 1. \end{cases}$$

Now we will connect this operator to Fourier expansions. We have already seen that $f \in M_k(\Gamma_1(N))$ gives rise to a Fourier expansion

$$f(z) = \sum_{n=0}^{\infty} a_n(f) q^n \qquad (q = e^{2\pi i z}).$$

Theorem 4.3.2. *Let $1_N \colon (\mathbb{Z}/N\mathbb{Z})^\times \to \mathbb{C}^\times$ be a trivial character modulo N. Then, $T_p f$ has the following Fourier expansion,*

$$\begin{aligned}(T_p f)(z) &= \sum_{n=0}^{\infty} a_{np}(f) q^n + 1_N(p) p^{k-1} \sum_{n=0}^{\infty} a_n(\langle p \rangle f) q^{np} \\ &= \sum_{n=0}^{\infty} (a_{np}(f) + \chi(p) p^{k-1} a_{n/p}(f)) q^n\end{aligned}$$

Equivalently, we have

$$a_n(T_p f) = a_{np}(f) + \chi(p) p^{k-1} a_{n/p}(f)$$

for $f \in M_k(N, \chi)$.

We can extend the T_p operator to be defined for all $n \in \mathbb{Z}^+$ using this formula for the Fourier coefficients.

Definition 4.3.3. For p prime and $r \geq 2$, set

$$T_{p^r} = T_p T_{p^{r-1}} - p^{k-1} \langle p \rangle T_{p^{r-2}}.$$

(When $r = 1$, T_p is defined as in Definition 4.3.1.) Then define

$$T_n = \begin{cases} 1 & \text{if } n = 1, \\ \prod T_{p_i^{r_i}} & \text{if } n = \prod p_i^{r_i} \text{ is the prime factorization of } n \geq 2. \end{cases}$$

We can bundle the T_n and the diamond operator, our two Hecke operators, into a single object: the Hecke algebra.

Definition 4.3.4. The *Hecke algebra* $\mathbb{T}_\mathbb{Z}$ over $\mathbb{Z}$ is the algebra of endomorphisms of $M_k(\Gamma_1(N))$ generated by the Hecke operators, that is,

$$\mathbb{T}_\mathbb{Z} = \mathbb{Z}[\{T_n, \langle n \rangle : n \in \mathbb{Z}^+\}].$$

An important result is that the Hecke operators commute.

Theorem 4.3.5. *Let $d, e \in (\mathbb{Z}/N\mathbb{Z})^\times$ and let p, q be prime. Then we have*

(1) $\langle d \rangle T_p = T_p \langle d \rangle$.

(2) $\langle d \rangle \langle e \rangle = \langle e \rangle \langle d \rangle = \langle de \rangle$.

(3) $T_p T_q = T_q T_p$.

Proving this requires a fair amount of computational work, but the details can be found in [1, Chapter 5]. Our extended operators $\langle n \rangle$ and T_n inherit the commutativity, which means we have the following result as well,

Theorem 4.3.6. *The Hecke algebra $\mathbb{T}_\mathbb{Z}$ is commutative.*

4.4 The Peterson Inner Product

Recall from section 3.3 the space of modular forms, $M_k(\Gamma)$, and the space of cusp forms, $S_k(\Gamma)$. In the previous sections we defined double coset operators for any pair of congruence subgroups Γ_1 and Γ_2 of $SL_2(\mathbb{Z})$ which took $M_k(\Gamma_1)$ to $M_k(\Gamma_2)$ and cusp forms to cusp forms. We then specialized to the case where $\Gamma_1 = \Gamma_2 = \Gamma_1(N)$ and found more precise operators, $\langle n \rangle$ and T_n, defined for all $n \in \mathbb{Z}^+$. These are the Hecke operators, commuting endomorphisms on the space $M_k(\Gamma_1(N))$ and $S_k(\Gamma_1(N))$. Now we will define an inner product on the subspace $S_k(\Gamma_1(N))$. This will allow us to study the eigenvalues and eigenvectors of our operators, which turn out to be of key importance. Note that quite a bit of theory goes into the material of the next two sections. We not give many of the details, instead choosing to focus on the overarching picture, which will conclude with the construction of a remarkable set of functions.

Definition 4.4.1. Let Γ and Γ' be congruence subgroups such that $\Gamma' \subseteq \Gamma$. Let $F(\Gamma') \subseteq \mathbb{C}$ be the fundamental domain for the modular curve $X(\Gamma')$ (2.4). The *Peterson inner product* $\langle\ ,\ \rangle \colon S_k(\Gamma) \times S_k(\Gamma) \to \mathbb{C}$ is given by

$$\langle f, g \rangle = \frac{1}{[\overline{\Gamma} : \overline{\Gamma'}]} \int_{F(\Gamma')} f(z)\overline{g(z)}\, \mathrm{Im}(z)^k \, \frac{dxdy}{\mathrm{Im}(z)^2},$$

where $\overline{\Gamma} = \Gamma/\{\pm \mathrm{Id}\}$ and $z \in \mathbb{C}$.

Note the use of the hyperbolic measure $dxdy/\mathrm{Im}(z)^2$ on H, which is $\mathrm{SL}_2(\mathbb{Z})$ invariant. Also, note that the integral formula converges if at least one of f, g is a cusp form. Finally, the inner product is linear in f, conjugate linear in g, antisymmetric, positive definite, and so it turns $S_k(\Gamma)$ into an inner product space. These details, including the fact that this is well-defined, are handled in [1].

We now have a family of commuting operators, namely the Hecke operators $\langle n \rangle$ and T_n, on the inner product space $S_k(\Gamma_1(N))$. We have the following theorem on that relation, again leaving the details to chapter 5 of [1],

Theorem 4.4.2. *In $S_k(\Gamma_1(N))$, the Hecke operators $\langle p \rangle$ and T_p for $p \nmid N$ have adjoints*

$$\langle p \rangle^* = \langle p \rangle^{-1} \quad \textit{and} \quad T_p^* = \langle p \rangle^{-1} T_p.$$

Thus, the Hecke operators $\langle n \rangle$ and T_n are normal in $S_k(\Gamma_1(N))$ whenever $\gcd(n, N) = 1$. It follows from the spectral theorem that The space $S_k(\Gamma_1(N))$ has an orthogonal basis of simultaneous eigenforms of the Hecke operators $\{\langle n \rangle, T_n : \gcd(n, N) = 1\}$.

Our next step will be to eliminate the restriction that n and N are coprime. This will require decomposing $S_k(\Gamma_1(N))$ into an "old" subspace and a "new" subspace. The new subspace will have an orthogonal basis of eigenforms for all Hecke operators.

4.5 The Space of Newforms

So far we have worked with a fixed level N. We will now introduce results that move between levels. Specifically, we will be concerned with moving forms to higher levels. Then, we will

find ourselves interested in forms which do not originate from a lower level, and are in some sense new at level N.

First, we can notice that if $M \mid N$, a form $f \in S_k(\Gamma_1(M))$ can be seen as a form in $S_k(\Gamma_1(N))$. This is because f being weight-k invariant with respect to matrices in $\Gamma_1(M)$ immediately implies that the same holds for matrices in $\Gamma_1(N)$.

A more technical approach involves embedding $S_k(\Gamma_1(M))$ into $S_k(\Gamma_1(N))$. To do so, consider any d which is a factor of N/M. Then set

$$\alpha_d = \begin{pmatrix} d & 0 \\ 0 & 1 \end{pmatrix}$$

and consider the multiply-by-d map, $[\alpha_d]$, which gives us

$$f[\alpha_d](z) = (\det \alpha_d)^{k-1} f(dz) = d^{k-1} f(dz).$$

Since for any $f \in S_k(\Gamma_1(N))$ we have $f(pz) \in S_k(\Gamma(pN))$, this map takes $S_k(\Gamma_1(M))$ to $S_k(\Gamma_1(N))$, listing the level of f from M to N. It is natural to consider splitting $S_k(\Gamma_1(N))$ into two parts, one part made up of forms that can be lifted from lower levels and one part containing forms that are new to the level N.

Definition 4.5.1. With α_d as above, define $i_d \colon S_k(\Gamma_1(Nd^{-1})) \times S_k(\Gamma_1(Nd^{-1})) \to S_k(\Gamma_1(N))$ by

$$(f, g) \mapsto f + g[\alpha_d]_k$$

The subspace of *oldforms at level N* is

$$S_k(\Gamma_1(N))^{\text{old}} = \sum_{\substack{p \text{ prime} \\ p \mid N}} i_p((S_k(\Gamma_1(Np^{-1})))^2).$$

We denote its the orthogonal complement with respect to the Peterson inner product by

$$S_k(\Gamma_1(N))^{\text{new}} = (S_k(\Gamma_1(N))^{\text{old}})^{\perp}.$$

The subspaces $S_k(\Gamma_1(N))^{\text{new}}$ and $S_k(\Gamma_1(N))^{\text{old}}$ are stable with respect to all Hecke operators $\langle n \rangle$ and T_n for all $n \in \mathbb{Z}^+$. In other words, both subspaces have orthogonal bases consisting of eigenforms of the Hecke operators $\{T_n, \langle n \rangle : \gcd(n, N) = 1\}$. However, importantly, the condition that $\gcd(n, N) = 1$ can be dropped in the case of newforms, which we can now define.

Definition 4.5.2. A *newform* is a modular form f such that $f \in S_k(\Gamma_1(N))^{\text{new}}$ and f is a normalized Hecke eigenform, which is to say an eigenform for all Hecke operators $T_n \in \mathbb{T}_{\mathbb{Z}}$.

We will not give the details surrounding the construction of these functions, since it requires a fair bit of theory, but we will state the key result.

Theorem 4.5.3. *The set of newforms gives an orthogonal basis for $S_k(\Gamma_1(N))^{\text{new}}$. Each newform f is such that $f \in S_k(N, \chi)$ and satisfies*

$$T_n f = a_n(f) f$$

for all $n \in \mathbb{Z}^+$.

The above theorem says that the T_n-eigenvalues of f aligns with the Fourier coefficients of f. We can use these Fourier coefficients to generate a number field (see appendix A for the necessary background).

Theorem 4.5.4. *Let f be a newform and consider the following field,*

$$K_f = \mathbb{Q}[a_n(f)]$$

for all $n \in \mathbb{Z}^+$. Then, K_f is a number field.

We call K_f the number field of f. We are also interested in the ring of integers of K_f.

Theorem 4.5.5. *Let f be a newform. Then, $a_n(f)$ are algebraic integers. That is,*

$$\mathbb{Z}[a_n(f)] \subseteq \mathcal{O}_{K_f}$$

These newforms will ultimately serve as our bridge back to elliptic curves. In the next section we will attach Galois representations to elliptic curves and newforms and describe the conditions under which those representations agree. This will allow us to state the Galois theoretic version of the modularity theorem.

Chapter 5

The Modularity Theorem

We will now attach what are known as Galois representations to elliptic curves and newforms, allowing us to state the Galois theoretic version of the modularity theorem. We will begin by introducing some of the algebraic background which will allow us to precisely define Galois representations. This includes describing the absolute Galois group in terms of certain nice subgroups, and then discussing the generators, known as Frobenius elements. We will give overviews of the construction of Galois representations for elliptic curves and newforms and conclude with a statement of the modularity theorem. More details of these constructions can be found in chapter 9 of [1].

5.1 Galois Representations

We will begin by discussing some of the algebraic material necessary to understand the constructions in the following section. This material comes from chapter 9 in [1]. This chapter uses freely results from appendices A and A.2

We first pay attention to the absolute Galois group.

The action of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ on $\overline{\mathbb{Q}}$ restricts to an action on $\overline{\mathbb{Z}}$. This allows us to define a key set of subgroups. Select some prime $p \in \mathbb{Z}$ and consider a prime $\mathfrak{p}$ over p. We can define the following subgroup of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$,

Definition 5.1.1. The decomposition group of $\mathfrak{p}$ is given by,

$$D_{\mathfrak{p}} = \{\alpha \in \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) : \alpha(\mathfrak{p}) = \mathfrak{p}\}$$

Now, since $\sigma \in D_{\mathfrak{p}}$ fixes $\mathfrak{p}$ we can think about the action on the closure of $\mathbb{F}_p$.

$$\sigma(x + \mathfrak{p}) = \sigma(x) + \mathfrak{p}$$

Note that this action on $\overline{\mathbb{F}_p}$ also fixed $\mathbb{F}_p$. Working mod $\mathfrak{p}$ we can consider the map $D_p \to \mathrm{Gal}(\overline{\mathbb{F}_p}/\mathbb{F}_p)$. We will prove this map is a surjection by showing it in the finite case and then lifting the result to the infinite case. To do so we will need to introduce a couple of initial results for the case of finite extensions. Note that the decomposition subgroup can be defined in the same way for finite extensions.

Theorem 5.1.2. *The fixed field K^D of $D_\mathfrak{p}$,*

$$K^D = \{a \in K : \sigma(a) = a\}$$

for all $\sigma \in D_\mathfrak{p}$, is the smallest subfield $F \subset K$ such that $\mathfrak{p} \cap F$ does not split in K.

Proof. Let $F = K^D$. Without proof we have the following isomorphism,

$$\mathrm{Gal}(K/F) \cong D_\mathfrak{p},$$

and the fact that $D_\mathfrak{p}$ acts transitively on the primes of K lying over $\mathfrak{p} \cap F$. Now, $\mathfrak{p}$ is a prime over $\mathfrak{p} \cap F$, and $D_\mathfrak{p}$ fixes $\mathfrak{p}$, so $\mathfrak{p}$ must be the only prime of K over $\mathfrak{p} \cap F$, which implies that $\mathfrak{p} \cap F$ does not split in K.
Next, if we suppose that we have $F \subset K$ such that $\mathfrak{p} \cap F$ does not split in K, then $\mathrm{Gal}(K/F)$ fixes $\mathfrak{p}$, and so $\mathrm{Gal}(K/F) \subset D_\mathfrak{p}$, and thus $K^D \subset F$. This concludes the proof. □

We have shown that p does not split when moving from K^D to K. The next theorem describes the behavior moving from $\mathbb{Q}$ to K^D.

Theorem 5.1.3. *Consider the fixed field K^D and a Galois extension $K/\mathbb{Q}$. We will show that $e = f = 1$ and $g = [K^d : \mathbb{Q}]$ for the extension $K^D/\mathbb{Q}$. That is, p splits completely in K^D.*

Proof. The group $\mathrm{Gal}(K/\mathbb{Q})$ acts on the set of primes $\mathfrak{p}$ over p. The decomposition group is the stabilizer in G of $\mathfrak{p}$. By the orbit-stablizer theorem, $[G : D_\mathfrak{p}]$ is equal to the orbit of $\mathfrak{p}$, which is exactly the number of primes over p. This shows that $g = [K^D : \mathbb{Q}]$. Then we have,

$$\begin{aligned} e(K/K^D) \cdot f(K/K^D) &= [K : K^D] \\ &= [K : \mathbb{Q}]/[K^D : \mathbb{Q}] \\ &= \frac{e(K/\mathbb{Q}) \cdot f(K/\mathbb{Q}) \cdot g(K/\mathbb{Q})}{[K^d : \mathbb{Q}]} \\ &= e(K/\mathbb{Q}) \cdot f(K/\mathbb{Q}) \end{aligned}$$

This implies that $e(K/K^D) = e(K/\mathbb{Q})$ and $f(K/K^D) = f(K/\mathbb{Q})$. Finish the proof by noting that $e(K/\mathbb{Q}) = e(K/K^D) \cdot e(K^D/\mathbb{Q})$ and $f(K/\mathbb{Q}) = f(K/K^D) \cdot f(K^D/\mathbb{Q})$. It follows that $e = f = 1$. □

Now we are ready to proof our main result.

Theorem 5.1.4. *the reduction map from $\phi : D_\mathfrak{p} \to Gal(\overline{\mathbb{F}_\mathfrak{p}}/\mathbb{F}_\mathfrak{p})$ is surjective.*

Proof. We will prove the result for finite extensions before lifting it to the infinite case. Consider the reduction map from $D_\mathfrak{p} \to \mathrm{Gal}(\mathbb{F}_\mathfrak{p}/\mathbb{F}_p)$.

Consider the reduction map from $D_\mathfrak{p} \to \mathrm{Gal}(\mathbb{F}_\mathfrak{p}/\mathbb{F}_p)$.

Choose some $\tilde{a} \in \mathbb{F}_\mathfrak{p}$ such that $\mathbb{F}_\mathfrak{p} = \mathbb{F}_p(\tilde{a})$. Lift $\tilde{a}$ to an algebraic integer $a \in mathcalO_K$ and consider the characteristic polynomial of a over K^D,

$$f = \prod_{\sigma \in D_p} (x - \sigma(a)).$$

We can reduce this polynomial to the minimal polynomial of $\tilde{a}$,

$$\tilde{f} = \prod (x - \tilde{\sigma(a)}).$$

Theorem 28 ensures that this is indeed the minimal polynomial, as we know that the coefficients of $\tilde{f}$ are in F_p, $\tilde{a}$ satisfies $\tilde{f}$m and the degree of $\tilde{f}$ equals the degree of the minimal polynomial of $\tilde{a}$. Now, examining the roots of $\tilde{f}$ we see they are of the form $\tilde{\sigma}(a)$. Now, $\mathrm{Frob}_p(a)$ is also a root of $\tilde{f}$, and so it is of the form $\tilde{\sigma(a)}$. Since $\mathrm{Frob}_p(a)$ generates $\mathrm{Gal}(F_\mathfrak{p}/F_p)$ and is in the image we conclude that the reduction map is surjective.

Now we would like to take this result and use it to prove the same in the infinite case. The absolute Galois group can be realized as the inverse limit of finite Galois groups $\mathrm{Gal}(K/\mathbb{Q})$ (A). In the same way we can see $\mathrm{Gal}(\overline{\mathbb{F}_\mathfrak{p}}/\mathbb{F}_\mathfrak{p})$ as an inverse limit of finite Galois groups $\mathrm{Gal}(\mathbb{F}_q/\mathbb{F}_p)$. This means we can select some $\tilde{\sigma} \in \mathrm{Gal}(\overline{\mathbb{F}_\mathfrak{p}}/\mathbb{F}_\mathfrak{p})$ and represent it with a compatible system $(\tilde{\sigma_n})_{n \in I}$ of elements $\tilde{\sigma_n} \in \mathrm{Gal}(\mathbb{F}_{p^n}/\mathbb{F}_p)$. We just saw that for each $\mathfrak{p}_n$ we can find a $\sigma_n \in D_\mathfrak{p}|_{K_n}$ such that σ_n reduces to $\tilde{\sigma_n}$. The collection of σ_n form a compatible system for the groups $\mathrm{Gal}(K_n/\mathbb{Q})$ since they came from the compatible system $(\tilde{\sigma_n})_{n \in I}$. The inverse limit knits together the $(\sigma_n)_{n \in I}$ into an element of $\sigma \in \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ which reduces to $\tilde{\sigma}$, and so the reduction map $\phi : D_\mathfrak{p} \to \mathrm{Gal}(\overline{\mathbb{F}_\mathfrak{p}}/\mathbb{F}_\mathfrak{p})$ is surjective. □

The kernel of this map gives us the definition of the inertia group of $\mathfrak{p}$.

$$I_p = \{\sigma \in D_\mathfrak{p} : \sigma(x) = x(\mathfrak{p})$$

Note that the above is for all $x \in \overline{\mathbb{Z}}$. This gives the following short exact sequence,

$$1 \to I_p \to D_p \to \mathrm{Gal}(\overline{\mathbb{F}_p}/\mathbb{F}_P) \to 1$$

Recall the Frobenius automorphism of $\overline{\mathbb{F}_p}$: $\phi_p(x) = x^p$.

Definition 5.1.5. A Frobenius element, denoted $\mathrm{Frob}_\mathfrak{p}(x)$, is any $\sigma \in D_\mathfrak{p}$ which acts as the Frobenius automorphism. That is, $\mathrm{Frob}_\mathfrak{p}(x) = \sigma(x) = \phi_p(x) = x^p(\mathfrak{p})$ for $x \in \overline{\mathbb{Z}}$.

By the aforementioned short exact sequence (5.1) we know that these Frobenius elements are only defined up to multiplcation by $I_\mathfrak{p}$. This presents two troubles with our definition. One, the intertial subgroup, the kernel, is large. Two, these elements do not just depend on p, but on the $\mathfrak{p}$ we select over p. While not perfect, we are able to think of them up to conjugation. That is, if $\mathfrak{p}$ and $\mathfrak{p}'$ are both over p, they have conjugate Frobenius elements,

$$\sigma \mathrm{Frob}_\mathfrak{p} \sigma^{-1} = \mathrm{Frob}_{\sigma(\mathfrak{p})} = \mathrm{Frob}_{\mathfrak{p};}$$

Similarly, the subgroups $D_{\mathfrak{p}}$ and all $I_{\mathfrak{p}}$ are also conjugate by elements of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. That is, we have the following identities,

$$D_{\sigma(\mathfrak{p})} = \sigma D_{\mathfrak{p}} \sigma^{-1}$$
$$I_{\sigma(\mathfrak{p})} = \sigma I_{\mathfrak{p}} \sigma^{-1}$$

Frobenius elements are particularly useful because they form a dense subset of much of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. While not as desirable as having a dense subset of all of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, we are able to say something,

Theorem 5.1.6. *Let $\mathbb{Q}^U$ be the maximal extension of $\mathbb{Q}$ in which all but finitely many primes ramify. Then, the Frobenius elements of unramified primes are dense with respect to the Krull topology in $Gal(\mathbb{Q}^U/\mathbb{Q})$.*

We leave this without proof as it requires too much material that we have not developed. This statement allows us to more easily work with continuous maps on $\mathrm{Gal}(\mathbb{Q}^U/\mathbb{Q})$, since such a map's image will be determined by their image on the Frobenius elements. With this background in hand we are able to discuss the representations of interest.

Definition 5.1.7. A $d-$dimensional l-adic Galois representation is a continuous homomorphism

$$\rho : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to GL_d(L)$$

where L is a finite extension of $\mathbb{Q}_l$.

We give one of the critical examples of a Galois representation to illustrate the definition.

Example 5.1.8. We will define the $l-$adic cyclotomic character. Consider the set of the nth roots of unity in $\overline{\mathbb{Q}}^{\times}$,

$$\mu_n = \{\zeta \in \overline{\mathbb{Q}}^{\times} : \zeta^n = 1\}$$

There exists the following isomorphism,

$$\mathrm{Gal}(\mathbb{Q}(\mu_n)/\mathbb{Q}) \to (\mathbb{Z}/n\mathbb{Z})^{\times}$$

given by $\sigma \to a$ such that $\sigma(\zeta_n) = \zeta_n^a$ for any $\zeta_n \in \mu_n$. We can take the inverse limit with respect to the natural restriction maps. Let l be prime. Then,

$$\mathrm{Gal}(\mathbb{Q}(\mu_{l^\infty}/\mathbb{Q}) \cong \varprojlim_{n \geq 1} \mathrm{Gal}(\mathbb{Q}(\mu_{l^n})/\mathbb{Q})$$

We can give another characterization of this inverse limit using the original isomorphism. For $m \leq n$ we have the following commutative diagram,

$$\begin{array}{ccc} \mathrm{Gal}(\mathbb{Q}(\mu_{l^n})/\mathbb{Q}) & \longrightarrow & \mathrm{Gal}(\mathbb{Q}(\mu_{l^m})/\mathbb{Q}) \\ \downarrow & & \downarrow \\ (\mathbb{Z}/l^n\mathbb{Z})^{\times} & \xrightarrow[\text{mod } l^m]{} & (\mathbb{Z}/l^m\mathbb{Z}) \end{array}$$

This gives the following set of relations,

$$\mathrm{Gal}(\mathbb{Q}(\mu_{l^\infty}/\mathbb{Q}) \cong \varprojlim_{n\geq 1} \mathrm{Gal}(\mathbb{Q}(\mu_{l^n})/\mathbb{Q}) \cong \varprojlim_{n\geq 1} (\mathbb{Z}/l^n\mathbb{Z})^\times \cong \mathbb{Z}_l^\times$$

The $l-$adic cyclotomic character is the map,

$$\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{Gal}(\mathbb{Q}(\mu_{l^\infty})/\mathbb{Q}) \to \mathbb{Z}_l^\times$$

However, our definition requires a field, and so we embed $\mathbb{Z}_l^\times$ into $\mathbb{Q}_l^\times$.

$$\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathbb{Z}_l^\times \to GL_1(\mathbb{Q}_l)$$

Now that we have definition 5.1.7 we can say what it means for a Galois representation to be unramified.

Definition 5.1.9. Let ρ be a Galois representation as defined above. To say ρ is unramified at p is to say that $I_\mathfrak{p} \subset \ker\rho$ for any $\mathfrak{p} \subset \overline{\mathbb{Z}}$ over p.

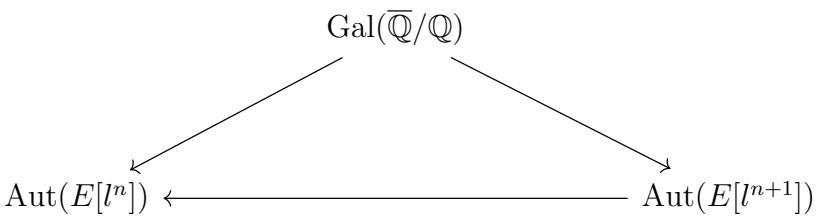
We would like to connect this notion to the idea of a number field being unramified at p. It turns out that ρ being unramified at a prime p implies that p will be unramified in $\overline{\mathbb{Q}}^{\ker\rho}$. We care about this since our earlier density theorem requires that we exclude a finite set of primes. In the next section we will construct Galois representations for elliptic curves and newforms. We will do so using the Galois representations defined in 5.1.7.

5.2 The Galois Representation of an Elliptic Curve

Now we will construct the 2-dimensional Galois representations associated to an elliptic curve.

Consider elliiptic curve E over $\mathbb{Q}$. We are going to work with the natural action of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ on E. That is, if $[x_1, x_2, x_3]$ with $x_i \in \overline{\mathbb{Q}}$ is a point on E then we have $\sigma([x_1, x_2, x_3]) = [\sigma(x_1), \sigma(x_2), \sigma(x_3)]$.

Recall the multiplication-by-m map given in definition 1.4.7 and remember that said map can be described in terms of rational polynomials. The coefficients of these polynomials are fixed by automorphisms in $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. This implies that the action of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ commutes with multiplication by m, which is equivalent to saying that $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ maps torsion subgroups into themselves. The action then restricts to $E[l^n]$ for all positive l and n. We can view the action of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ on l^n-torsion as a map $\sigma_n : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{Aut}(\mathbb{Z}/l^n\mathbb{Z})$, giving rise to the following commutative diagram,



We know that $E[l^n] \cong (\mathbb{Z}/l^n\mathbb{Z})^2$, and so $\mathrm{Aut}(E[l^n]) \cong GL_2(\mathbb{Z}/l^n\mathbb{Z})$. This gives rise to the following maps, after applying the inverse limit,

$$\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{Aut}(\mathrm{Ta}_l(E)) \to GL_2(\mathbb{Z}_l) \to GL_2(\mathbb{Q}_l)$$

Here, the presence of the Tate module gives us our dependence on our specific choice of E. Taking the composition gives us the map we were looking for,

Definition 5.2.1. The l-adic Galois representation associated to an elliptic curve E is,

$$\rho_{E,l} : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to GL_2(\mathbb{Q}_l)$$

However, our definition of a Galois representation requires continuity, which we have not yet shown. Proving continuity will require theorem 36 in A.2. Specifically that the following sets form a basis for the Krull topology,

$$\{U_M(n)\}_{M \in GL_2(\mathbb{Q}_l), n \in \mathbb{Z}^+}$$

where $U_M(n)$ is defined as $M(I + l^n M_2(\mathbb{Z}_l))$ and $M_d(\mathbb{Z}_l)$ is the set of d by d matrices in $\mathbb{Z}_l$. It is sufficient to check the maps behavior on basis sets to determine continuity. The question boils down to showing that the inverse image $\rho_{E,l}^{-1}(U_M(n))$ is open for all M and n. We know that multiplication by $M \in GL_2(\mathbb{Q}_l)$ gives a continuous map on $GL_2(\mathbb{Q}_l)$, and so all we must show is that $\rho_{E,l}^{-1}(U_1(n))$ is open for all n.

For $M \in U_1(n)$ and $e \leq n$ we have $M = I(l^e)$. That is, working mod l^e,

$$M = (I, ..., I, M_{n+1}, M_{n+2}...)$$

Now, pulling M back along $\rho_{E,l}^{-1}$ to $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ we have,

$$(1, ..., 1, \sigma_{i_{n+1}}, \sigma_{i_{n+2}}, ...) \in \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$$

These identity entries correspond to finite Galois extensions $\mathrm{Gal}(K_{i_j}/\mathbb{Q})$ with the l^j-torsion coordinates. Crucially, each element of the second kind maps to an element of the first, and so,

$$\begin{aligned} W &= \rho_{E,l}^{-1}(U_M(n)) \\ &= \{\sigma \in \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) : \sigma = (1, ..., 1, \sigma_{i_{n+1}}, \sigma_{i_n+2}, ...)\} \end{aligned}$$

which is a subgroup of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ whose elements fix the number fields K_{i_j}. But then $\overline{\mathbb{Q}}^W = \cup_{j=1}^n K_{i_j}$ is a finite extension of $\mathbb{Q}$ such that $W = \mathrm{Gal}(K^W/\mathbb{Q})$. In particular, theorem 36 in A.2 says that W is open in the Krull topology. This tells us that our map is continuous, and so we do indeed have a Galois representation. There is lots to be said about this representation, including a large number of properties that are important for the connection to modular forms, but they go beyond the needs of this paper. We will state one key theorem before moving on.

Theorem 5.2.2. *Let E be an elliptic curve over $\mathbb{Q}$ with conductor N and some fixed prime l. Then, $\rho_{E,l}$ is unramified for all primes which do not divide lN and for any $\mathfrak{p}$ over p the image $\rho_{E,l}(Frob_{\mathfrak{p}})$ obeys the following equation,*

$$x^2 - a_p(E)x + p = 0$$

where $a_p(E) = p + 1 - \#E(F_p)$. In this case we have an irreducible representation.

Now we move to construct the Galois representations attached to newforms.

5.3 The Galois Representation of a Newform

This construction is much more involved than the one for elliptic curves. Unfortunately, we are not able to fit all of the requisite background material into this paper, and so we settle for giving a broad overview of the construction given in section 9.5 of [1].

We will be concerned only with weight-2 forms, and so let $f \in M_2(\Gamma_1(N))$ be a newform and consider $X_1(N)$.

The construction of this Galois representation follows three main stages. First, we want to take $X_1(N)$ over $\mathbb{C}$ and realize it as a curve over $\mathbb{Q}$. After doing so we are able to construct a Galois representation,

$$\rho_{X_1(N),l} : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to GL_{2g}(\mathbb{Q}_l)$$

Where g is the genus of $X_1(N)$ over $\mathbb{Q}$.

From here our goal will be to attach a d-dimensional complex torus A_f to f. We use this to construct another Galois representation,

$$\rho_{A_f,l} : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to GL_{2d}(\mathbb{Q}_l)$$

We are able to prove that this gives a representation by making use of 5.3

Finally, we decompose 5.3 in such a way to obtain representations for any prime l in the number field K_f,

$$\rho_{f,l} : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to GL_2(\mathbb{Q}_l)$$

Constructing 5.3 and 5.3 involves a similar strategy to the elliptic curve case. We define the Tate module of $X_1(N)$ and A_f using l^n-torsion and from there the construction of the representation is much the same. However, there is a problem in that the points on $X_1(N)$ and A_f do not define a group, and so we have to appeal to the Picard group of the objects. We will omit the technical details. In the end, the Tate modules of $X_1(N)$ and A_f are the inverse limits of torsion, as expected. This yields the following isomorphisms,

$$Ta(X_1(N)) \cong \mathbb{Z}_l^{2g}$$
$$Ta(A_f) \cong \mathbb{Z}_l^{2d}$$

This gives us representations of dimension $2g$ and $2d$. We can decompose the second of these representations. We can tensor $\mathbb{Z}_l^{2d}$ with $\mathbb{Q}_l$ over $\mathbb{Q}_l$. This results in,

$$V_l(A_f) = Ta(A_f) \otimes_{z_l} \mathbb{Q}_l \cong \mathbb{Q}_l^{2d}$$

Now, this is the vector space underlying the representation, and so to decompose $\rho_{A_f,l}$ is to decompose $V_l(A_f)$. It turns out that we have the following decomposition,

$$V_l(A_f) \cong (K_f \otimes_{\mathbb{Q}} \mathbb{Q}_l)^2$$

Now we must decompose the right hand side, which we can do for an arbitrary number field.

Let K be a number field and let $\mathcal{O}_K$ be its ring of integers. Fix a prime l and for each ideal λ over l define the ring of λ-adic integers $\mathcal{O}_{K,\lambda}$ through an inverse limit,

$$\mathcal{O}_{K,\lambda} = \varprojlim_n \mathcal{O}_K/\lambda^n$$

Further, define the field of λ-adic numbers K_λ as the field of fractions of $\mathcal{O}_{K,l}$. This matches the definition of the $l-$adic numbers in the appendix except we have taken $K = \mathbb{Q}$. Since $l\mathbb{Z} \subset \lambda\mathcal{O}_K$ there is an embedding from Z_l to $\mathcal{O}_{K,\lambda}$. This implies the existence of an embedding from $\mathbb{Q}_l$ to K_λ. That is, K_λ is a finite extension of $\mathbb{Q}_l$. We can state our desired decomposition as follows,

$$K \otimes_{\mathbb{Q}} \mathbb{Q}_l \cong \prod_{\lambda|l} K_\lambda$$

Putting this together gives,

$$V_l(A_f) \cong (\prod_{\lambda|l} K_{f,\lambda})^2$$

This tells us that $\rho_{A_{f,\lambda}}$ gives a homomorphism,

$$\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to GL_2(\prod_{\lambda|l} K_{f,\lambda})$$

If we fix λ then a projection map gives,

$$\rho_{f,\lambda} : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to GL_2(K_{f,\lambda})$$

This homomorphism is the l-adic Galois representation attached to the newform f.

As before, we state only a single key property of this representation.

Theorem 5.3.1. *Let $f \in S_2(N,\chi)$ be a newform with number field K_f and fix some prime l. For each prime in $\mathcal{O}_{K_f}$ over l we have the Galois representation attached to f: $\rho_{f,\lambda} : Gal(\overline{\mathbb{Q}}/\mathbb{Q}) \to GL_2(K_{f,\lambda})$. This representation is unramified at every prime that does not divide lN. For a prime $\mathfrak{p}$ over p, $\rho_{f,\lambda}(Frob_\mathfrak{p})$ satisfies the following equation,*

$$x^2 - a_p(f)x + \chi(p)p = 0$$

We can compare our two Galois representations. For elliptic curves, $\rho_{E,l}$ requires that $L = \mathbb{Q}_l$ in our definition of l-adic Galois representations. However, for newforms, $\rho f, \lambda$ requires a finite extension $L = K_{f,\lambda}$. For our correspondence to work we will need $\mathbb{Q}_l = K_{f,\lambda}$. Our definitions tell us that this occurs when $K_f = \mathbb{Q}$. The above theorem also tells us that if we want our representations to line up then they will need to agree at $\text{Frob}_\mathfrak{p}$. The following theorem describes when this occurs.

Theorem 5.3.2. *Let E be a semi-stable elliptic curve over $\mathbb{Q}$ with conductor N. There exists a newform $f \in S_2(\Gamma_0(N))$ with number field $K_f = \mathbb{Q}$ such that $\rho_{E,l} \sim \rho_{f,l}$ for all l.*

Here we have the Galois theoretic version of the Taniyama-Shimura conjecture, otherwise known as the modularity theorem. This is the version that was proven by Andrew Wiles to complete the proof of Fermat's Last Theorem. We will conclude with a brief summary of this history.

Conclusion

In 1637 Pierre De Fermat famously conjectured, in the margins of his notebook, that the equation

$$x^n + y^n = z^n$$

has no integer solutions for $n > 2$. Fermat claimed to have a proof of his conjecture, but none ever materialized. In the years that followed progress was made on the conjecture by the likes of Kummer and Germaine, but it was not until the mid 20th century that the chain of events which would ultimately lead to a proof began.

In 1955 Taniyama conjectured a version of the modularity theorem at the international symposium on algebraic number theory in Tokyo. His conjecture was flawed, however, and it took the help of Shimura to fix. Their conjecture, which they made in 1957, is the Taniyama-Shimura conjecture.

Nine years later Gerhard Frey pointed us towards the following curve,

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

It happened that if this curve were to have a solution it would be a solution to Fermat's Last Theorem,

$$a^n + b^n = c^n$$

for $n \geq 3$. Indeed, any solution to Fermat's Last Theorem would neccesitate a solution to this curve. Serre and Ribet helped Frey by showing that any solution to this elliptic curve would be non-modular. This means that Fermat's Last Theorem followed directly from the modularity theorem. If every elliptic curve is modular there could be no solution to Frey's proposed function, and therefore no solution to Fermat's last theorem.

Andrew Wiles, with help from Richard Taylor, was able to prove the special case of the modularity theorem we developed in this paper. This special case was just enough to prove Fermat's Last Theorem. The full modularity theorem was not shown until 2001. The result was ultimately published by Breuil, Conrad, Diamond, and Taylor.

Appendix A

Results from Algebraic Number Theory

A.1 Fields and Ideal Factorization

This appendix serves to lay out, without proof, some of the results from algebraic number theory that this paper will require. Note that while many of these definitions extend to arbitrary fields we specialize to $\mathbb{Q}$ when possible as it is our primary interest as number theorists.

Definition A.1.1. If there exists a ring homomorphism $\phi : F \to K$, we say that K is a field extension of F. This is generally denoted $F \subseteq K$ or K/F.

Note that K is automatically a vector space over F. We denote the dimension of such a space by,

$$[K : F] := \dim_F K$$

With this notion of a field extension we are able to define the set of algebraic numbers, specializing to $\mathbb{Q}$

Definition A.1.2. A number $z \in \mathbb{C}$ is algebraic if there exists a nonzero $f \in \mathbb{Q}[x]$ such that $f(z) = 0$. The set of algebraic numbers is as follows:

$$\overline{\mathbb{Q}} = \{\alpha \in \mathbb{C} : \alpha \text{ is algebraic over } \mathbb{Q}\}$$

Note that this is equivalent to the notion of the algebraic closure of $\mathbb{Q}$ and that $\overline{\mathbb{Q}}$ constitutes a field. There is an analogous definition of algebraic integers, denoted by $\overline{\mathbb{Z}}$, made by simply exchanging $\mathbb{Q}$ with $\mathbb{Z}$ in the above. The algebraic integers form a ring.

We can also consider finite subfields of $\overline{\mathbb{Q}}$. This will turn out to be an extremely useful idea.

Definition A.1.3. A number field is an extension $F/\mathbb{Q}$ such that $[F : \mathbb{Q}] < \infty$.

Later, we will find ourselves interested in a certain kind of extensions, known as Galois extensions.

Definition A.1.4. Let K and F be number fields. Let $|\mathrm{Aut}(K)|$ be the group of automorphisms of K that fix F. The extension K/F is Galois if $|\mathrm{Aut}(K)| = [K : \mathbb{Q}]$. Then, we say that $\mathrm{Gal}(K/F) = \mathrm{Aut}(K/F)$.

Example A.1.5. For instance, quadratic extensions are Galois, since for $F(\sqrt{a})$ with $a \in F$ we can define a nontrivial embedding,

$$\sqrt{a} \to -\sqrt{a},$$

which guarantees the existence of a nontrivial automorphism.

Definition A.1.6. When F is a number field, let $\mathcal{O}_F := F \cap \overline{\mathbb{Z}}$ be called the ring of integers of F. This represents the integral closure of $\mathbb{Z}$ in F.

Example A.1.7. The only rational numbers which are roots of monic polynomials with integers coefficients are integers, and so the ring of integers of $\mathbb{Q}$ is $\mathbb{Z}$.

We now state two key results on the factorization of ideals.

Theorem A.1.8. *Let F be a number field and consider $\mathcal{O}_F$. Any ideal $\mathfrak{a} \subset \mathcal{O}_F$ has the following unique factorization,*

$$\mathfrak{a} = \prod_{i=1}^{g} \mathfrak{p}_i^{e_i}$$

Here, $\mathfrak{p}_i$ is a maximal prime ideal and $e_i > 1$. This implies that every ring of integers of a number field is a Dedekind domain.

Theorem A.1.9. *Let $\mathbb{Q}$, F, and L be number fields and consider the tower of extension $\mathbb{Q} \subseteq F \subseteq K$ with rings of integers $\mathcal{O}_F$ and $\mathcal{O}_K$ respectively. Let $\mathfrak{p}$ refer to a maximal prime ideal in $\mathcal{O}_F$ and let $\mathfrak{P}$ denote a maximal prime ideal in $\mathcal{O}_K$. By the previous theorem we have the following unique factorization,*

$$\mathfrak{p}\mathcal{O}_L = \prod_{i=1}^{g} \mathfrak{P}_i^{e_i}$$

We can consider the set of prime ideals $\mathfrak{p}'$ lying over a prime $\mathfrak{p}$ in $\mathcal{O}_F$. If we fix a prime $\mathfrak{p} \subset \mathcal{O}_F$ we can write $\mathfrak{p}\mathcal{O}_K = \mathfrak{P}_1^{e_i}, ..., \mathfrak{P}_g^{e_i}$, letting $S_\mathfrak{p} = \{\mathfrak{P}_1^{e_i}, ..., \mathfrak{P}_g^{e_i}\}$. Importantly, the Galois group $\mathrm{Gal}(K/F)$ acts on $S_\mathfrak{p}$, which we will see shortly. We can consider the fields $\mathcal{O}_F/\mathfrak{p}$ and $\mathcal{O}_K/\mathfrak{P}$ to define the residue class degree of $\mathfrak{P}$.

Definition A.1.10. Let $f_i = [\mathcal{O}_K/\mathfrak{P} : \mathcal{O}_F/\mathfrak{p}]$ denote the residue class degree of $\mathfrak{P}$. Then,

$$\sum^{g} e_i f_i = [K : F].$$

Consider $\alpha \in \mathrm{Gal}(K/F)$ and $\mathfrak{P} \in S_\mathfrak{p}$. Then, α induces the following isomorphism between finite fields,

$$\mathcal{O}_L/\mathfrak{P} \to \mathcal{O}_K/\alpha(\mathfrak{P}),$$

which fixes the subfield $\mathcal{O}_F/\mathfrak{p}$. This says that the residue class degrees of $\mathfrak{P}$ and of $\alpha(\mathfrak{P})$ are equal. In fact, if K/F is a Galois extension of number fields, $\mathrm{Gal}(K/F)$ act transitively on the set $S_{\mathfrak{p}_i}$, and so,

$$e_1 =\ldots = e_g$$
$$f_1 =\ldots = f_g$$

and $efg = [K : F]$.

The first of these results gives us unique factorization of ideals. In the second we are taking a maximal prime ideal in F and constructing an ideal in K. Of course, $\mathfrak{p}\mathcal{O}_K$ is the cheapest way to generate an ideal in K. These ideals factor according to the first theorem. Then, we described the action of the Galois group on the set of prime ideals lying over a given prime, though we did so without proof.

We can now define three distinct kinds of factorization. Consider a prime ideal $\mathfrak{p}$ and its factorization,

$$\mathfrak{p}\mathcal{O}_K = \mathfrak{P}_1^{e_1}\ldots\mathfrak{P}_g^{e_g}$$

If $e_i > 1$ for any i then $\mathfrak{p}$ is said to ramify in $\mathcal{O}_K$. In this case e_i is the ramification index, and for all $\mathfrak{P}$ occurring in the factorization of $\mathfrak{p}$, we say that $\mathfrak{P}$ divides $\mathfrak{p}$.

If $g = [K : F]$ then $e_i = f_i = 1$ for all i and we say $\mathfrak{p}$ splits in K.

If $\mathfrak{p}$ remains prime once it has been lifted to $\mathcal{O}_K$ we say that it is inert. This occurs when $g = e = 1$.

We can apply these results to the case of quadratic extensions.

Example A.1.11. Consider the quadratic extension $K/\mathbb{Q}$. For each prime $p \in \mathbb{Z}$ we have $2 = efg$ and so there are three possibilities,
1. $e = 2$, $f = g = 1$. In this case, p ramifies in $\mathcal{O}_K$. That is, $p\mathcal{O}_K = \mathfrak{p}^2$. Note that number fields only contain finitely many such primes.
2. $e = f = 1$, $g = 2$. Here, p splits in $\mathcal{O}_K$, and so $p\mathcal{O}_K = \mathfrak{p}_1\mathfrak{p}_2$ and $\mathfrak{p}_1 \neq \mathfrak{p}_2$.
3. $e = 1$, $f = 2$, $g = 1$. This final case is when p is inert in K, which means that $p\mathcal{O}_K = \mathfrak{p}$ remains prime.

We can specialize to $K = \mathbb{Q}(\sqrt{5})$ and $\mathcal{O}_K = \mathbb{Z}[(1 + \sqrt{5})/2]$. We see that 5 is a ramified prime since $p\mathcal{O}_K = \sqrt{5}^2$. Now, $\mathbb{Z}[\sqrt{5}]$ constitutes an order with index 2, which means the factorization of all primes $p \neq 2$ will depend on the factorization of $x^2 - 5 \in \mathbb{F}_p[x]$. We see that $x^2 - 5$ splits if and only if $e = f = 1$ and $g = 2$, and for $p \neq 2, 5$ this occurs if and only if 5 is a square in $\mathbb{F}_p$. This can be described in terms of quadratic reciprocity. Recalling that

$\left(\frac{5}{p}\right) = 1$ if 5 is a square mod p and $\left(\frac{5}{p}\right) = -1$ if 5 is not a square mod p we have,

$$\left(\frac{5}{p}\right) = \left(\frac{p}{5}\right) = \begin{cases} 1 & \text{if } p = \pm 1 \mod 5 \\ -1 & \text{if } p = \pm 2 \mod 5 \end{cases}$$

And so the behavior of p is given by its residue class mod 5. In this example we see that an arbitrary prime remains inert around half the time and splits the other half of the time. Ramification is rare, but will be our main case of interest.

A.2 The Absolute Galois Group

The extension $\overline{\mathbb{Q}}/\mathbb{Q}$ is Galois: it is normal and separable as a result of being a characteristic 0 extension. This allows for the following definition,

Definition A.2.1. Let $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ be known as the absolute Galois group of $\mathbb{Q}$.

Our plan for this section is to see $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ as the limit of finite-degree Galois groups. To do so, we will need to construct the inverse limit.

Consider the tower of extensions given by $\overline{\mathbb{Q}}/K$ and $K/\mathbb{Q}$. Here, $K/\mathbb{Q}$ is Galois. We can define the following restriction: for $g \in \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ we can define $g_K \in \mathrm{Gal}(K/\mathbb{Q})$ by restricting g to K, that is $g_K = g|_K$. This gives a natural surjection from $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{Gal}(K/\mathbb{Q})$. We can go the other way as well. If we are given all $g_K \in \mathrm{Gal}(K/\mathbb{Q})$ then for any $q \in \overline{\mathbb{Q}}$ we can pick some K such that $q \in K$ and thus $g(q) = g_K(q)$. This gives a pairing between $g \in \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ and the collections $\{g_{K_i}\}$ where K_i ranges over the Galois number fields $K_i/\mathbb{Q}$. The following two points hold:

1. For all i, j, and considering $g_{K_i} \in \mathrm{Gal}KL_i/\mathbb{Q})$, we have $g_{K_i} = g_{K_j}$ on $K_i \cap K_j$.
2. The automorphism g restricts to g_{K_i} on K_i

This shows that we can relate elements of the absolute Galois group to elements in finite extensions. We will need the following definitions to make full sense of this phenomena.

Definition A.2.2. Consider a directed poset $(I, \leq)$. Let (G_i) be a family of groups alongside a family of homormophisms given by $f_{ij} : A_j \to A_i$ for $i \leq j$. These homomorphisms obey the following properties,
1. f_{ii} gives the identity on G_i.
2. $f_{ik} = f_{ij} \circ f_{jk}$ for $i \leq j \leq k$.
This pair of families, $((G_i), f_{ij}$, over the same index set, is known as an inverse system over I.

With this idea of an inverse system we can define the inverse limit:

Definition A.2.3. For some inverse system $((G_i), f_{ij})$ the inverse limit is a subgroup of the direct products of the G_i's.

$$G = \varprojlim_{i \in I} G_i = \{g \in \prod_{i \in I} G_i | g_i = f_{ij}(g_j) \text{ for all } i \leq j \in I$$

Note that in this definition the condition $g_i = f_{ij}(g_j)$ ensures that the elements of the inverse system are compatible with the reduction maps.

We can see how this definition interacts with our main interest $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. Let $(\{K_i\}, \leq)$ be our partially ordered set. Here, (K_i) is the collection of all Galois number fields and $\leq$ is given by a natural inclusion,

$$K_i \leq K_j \iff K_i \subset K_j$$

Let G_i be the corresponding Galois group $\mathrm{Gal}(K_i/\mathbb{Q})$ and let the homomorphisms f_{ij} be given by restriction. This satisfies the requisite properties and gives rise to an inverse system. Ultimately, this leads us to the following statement,

$$\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \cong \varprojlim_{i \in I} \mathrm{Gal}(K_i/\mathbb{Q})$$

This is to say that the absolute Galois group can be seen as the inverse limit of finite Galois groups. We will now construct the $l-$adic integers.

Example A.2.4. We will construct another inverse system. Consider the positive integers $\mathbb{Z}^+$ and fix some prime $l \in \mathbb{Z}^+$. For $n \in \mathbb{Z}^+$ set $G_n = \mathbb{Z}/l^n\mathbb{Z}$ and define $f_{nm} : \mathbb{Z}/l^m\mathbb{Z} \to \mathbb{Z}/l^n\mathbb{Z}$ by reduction mod l^m for $n \geq m$. This gives rise to an inverse system,

$$\mathbb{Z}/l\mathbb{Z} \leftarrow \mathbb{Z}/l^2\mathbb{Z} \leftarrow \mathbb{Z}/l^3\mathbb{Z} \leftarrow ... \leftarrow \mathbb{Z}/l^n\mathbb{Z} \leftarrow ...$$

This gives rise to the following limit,

$$\mathbb{Z}_l = \varprojlim_{n \in \mathbb{Z}^+} \mathbb{Z}/l^n\mathbb{Z}$$

There is a natural embedding from $\mathbb{Z}$ to this limit, $\mathbb{Z}_l$. For some $x \in \mathbb{Z}$ map x to the sequence $(x_1, x_2, x_3, ...x_i...)$ where each entry x_n is given by the reduction of x by l^n. This sequence is in $\mathbb{Z}_l$ as it satisfies the condition,

$$f_{mn}(x \mod l^m) = x \mod l^n$$

for $m \geq n$. This map is injective since it has a trivial kernel: if $x = 0 \mod l^n$ for all n then $x = 0$.

Definition A.2.5. The above $\mathbb{Z}_l$ is known as the ring of $l - adic$ integers. Its field of fractions, denoted $\mathbb{Q}_l$, is the field of $l - adic$ numbers, and $\mathbb{Z}_l$ is the ring of integers in $\mathbb{Q}_l$.

Now we return to our questions surrounding $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. If we equip each finite Galois group with the discrete topology then the inverse limit hands $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ a topology of its own, known as the Krull topology. We are interested in the following open sets,

Definition A.2.6. Consider the topological space X and some fixed element $x \in X$. A neighborhood base, $\mathcal{N}$, of x is a set of open neighborhoods such that any neighborhoods of x contains some $N \in \mathcal{N}$.

Theorem A.2.7. *For the Krull topology on Gal($\overline{\mathbb{Q}}/\mathbb{Q}$) the following set gives a neighborhood base of the identity,*

$$\{Gal(\overline{\mathbb{Q}}/M) : M/\mathbb{Q} \text{ is a finite Galois extension}\}$$

In this case, the subgroups Gal($\overline{\mathbb{Q}}/M$) are open.

The goal of this paper is to construct continuous maps between elliptic curves and modular forms, and it is the Krull topology on Gal($\overline{\mathbb{Q}}/\mathbb{Q}$) that we will use to determine continuity.

A.3 Projective Space

We will give a brief introduction to projective space. Our first goal is to compactify the standard euclidean space by gluing the ends—positive infinity and negative infinity—together. This creates a "point at infinity" that allows for the intersection of parallel lines. We will start by considering the one-dimensional real line. Let (a, b) and (x, y) be two nonzero vectors over $\mathbb{R}$. Define an equivalence relation by $(a, b) \sim (x, y)$ if $(x, y) = \lambda(a, b)$ for some real scalar $\lambda \in \mathbb{R}$. Let $[x, y]$ denote the equivalence class of (x, y) under this relation. With this we define the projective line,

$$\mathbb{P}^1(\mathbb{R}) = \{[x, y] : x, y \in \mathbb{R},\ (x, y) \neq (0, 0)\}.$$

That is, $\mathbb{P}^1(\mathbb{R})$ is the set of all lines through the origin. Now, if $y \neq 0$, it must be true that $(x, y) \sim (\frac{x}{y}, 1)$. Since $\frac{x}{y} \in \mathbb{R}$ and each class can be written uniquely in this way we have another way of expressing $\mathbb{P}^1(\mathbb{R})$,

$$\mathbb{P}^1(\mathbb{R}) = \{[x, 1] : x \in \mathbb{R}\} \cup \{[1, 0]\}$$

Here, $[x, 1]$ is our real line and $[1, 0]$ is our "point at infinity".

We can extend this construction to a two-dimensional plane. Our goal is to end with a real plane and a projective line of points at infinity. The process is almost identical. This time, take (a, b, c) and (x, y, z) to be non-zero vectors. As before, define an equivalence relation by $(x, y, z) \sim (a, b, c)$ if $(x, y, z) = \lambda(a, b, c)$ for some $\lambda \in \mathbb{R}$. Now, define the projective plane as follows,

$$\mathbb{P}^2(\mathbb{R}) = \{[x, y, z] : x, y, z \in \mathbb{R}, (x, y, z) \neq (0, 0, 0)\}.$$

which is again the collection of lines through the origin. We can equivalently realize this as

$$\mathbb{P}^2(\mathbb{R}) = \{[x, y, 1] : x, y \in \mathbb{R}\} \cup \{[a, b, 0] : a, b \in \mathbb{R}\}.$$

We can see that $[x, y, 1]$ gives us a copy of the real plane, and $[a, b, 0]$, itself a projective line, gives us our "points at infinity". In this setting, any two parallel lines will intersect at a point at infinity.

These ideas can be carried over to any arbitrary field. Over a field K the Euclidean plane is given by

$$\mathbb{A}^2(K) = \{(x, y) : x, y \in K\}.$$

The corresponding projective plane is given by,

$$\mathbb{P}^2(K) = \{[x, y, z] : x, y, z \in K, (x, y, z) \neq (0, 0, 0)\},$$

where the equivalence relation is the same as before.

Denote by $K[x, y]$ the ring of polynomials in two variables x and y with coefficients in the field K. Given a polynomial $f \in K[x, y]$, its zero set defines a curve C in euclidean space, which we denote by

$$C : f(x, y) = 0$$

We can extend $C \subseteq \mathbb{A}^2(K)$ to a curve $C' \subseteq \mathbb{P}^2(K)$. For example, take $C : y^2 - x^3 - 1 = 0$. We can view points (x, y) along the curve to be in the plane $[\frac{x}{z}, \frac{y}{z}, 1]$. After substituting and multiplying through by z we get

$$C' : F(x, y, z) = zy^2 - x^3 - z^3 = 0.$$

Direct verification shows that C' is C extended to the projective plane. The former contains all of the points of the latter, in additional to potentially more at infinity. The points at infinity are found by setting $z = 0$. In this case, the only point at infinity is $[0, 1, 0]$.

This example is a special case of the more general result. Let d be the highest degree of a monomial in f. Then, for a curve $C \subseteq \mathbb{A}^2(K)$ given by $f(x, y) = 0$, its extension to C' is given by

$$C' : F(x, y, z) = 0 \text{ where } F(x, y, z) = z^d \cdot f\left(\frac{x}{z}, \frac{y}{z}\right).$$

Note that this method allows us to go in the other direction as well. If we take $C = y - x^2 = 0$, a parabola, then the extension is given by

$$C' : F(x, y, z) = z^2 f\left(\frac{x}{z}, \frac{y}{z}\right) = zy - x^2 = 0$$

with a unique point at infinity, $[0, 1, 0]$. If we take $C : x^2 - y^2 = 1$ to be a hyperbola we get

$$C' : x^2 - y^2 - z^2 = 0,$$

which has two points at infinity: $[1, 1, 0]$ and $[1, -1, 0]$. Note that in the case of a parabola we are gluing the arms together and forming an ellipse. The same is happening in the case of a hyperbola: the four arms meet at two points and thus result in an ellipse in the projective plane.

We say a curve C is singular at a point $P \in C$ if and only if the following is true,

$$\frac{\partial F}{\partial x}(P) = \frac{\partial F}{\partial y}(P) = \frac{\partial F}{\partial z}(P) = 0$$

Else, we call a curve non-singular at P. If a curve is non-singular at every point we call it smooth.

An elliptic curve, $E : y^2 = x^3 + Ax + B$, which can also be seen over projective space as $E' : zy^2 + x^3 + Axz^2 + Bz^3$, is smooth if and only if $4A^3 + 27B^2 \neq 0$. The discriminant of E is given by $\Delta = -16 \cdot (4A^3 + 27B^2)$.

References

[1] F. I. Diamond and J. Shurman, *A First Course in Modular Forms*, Graduate Texts in Mathematics, 228, Springer, New York, 2005. MR2112196

[2] J.-P. Serre, *A Course in Arithmetic*, translated from the French, Graduate Texts in Mathematics, No. 7, Springer, New York, 1973. MR0344216

[3] Lozano-Robledo, *Elliptic Curves, Modular Forms, and Their L-functions*, Student Mathematical Library, 58, Amer. Math. Soc., Providence, RI, 2011. MR2757255

[4] J. H. Silverman, *The Arithmetic of Elliptic Curves*, second edition, Graduate Texts in Mathematics, 106, Springer, Dordrecht, 2009. MR2514094

[5] J.-S. Milne, *Modular Functions and Modular Curves*, graduate course notes, 2017

[6] J.-S. Milne, *Algebraic Number Theory*, graduate course notes, 2020

[7] J.-S. Milne, *Fields and Galois Theory*, graduate course notes, 2022

[8] P. Deligne, La conjecture de Weil. I, Inst. Hautes Études Sci. Publ. Math. No. 43 (1974), 273–307. MR0340258

[9] A. J. Wiles, Modular elliptic curves and Fermat's last theorem, Ann. of Math. (2) **141** (1995), no. 3, 443–551. MR1333035

[10] C. L. Siegel, Über einige Anwendungen diophantischer Approximationen, On some applications of Diophantine approximations, 81–138

[11] A. Weil, L'arithmétique sur les courbes algébriques, Acta Math. **52** (1929), no. 1, 281–315. MR1555278

[12] B. C. Mazur, Rational isogenies of prime degree (with an appendix by D. Goldfeld), Invent. Math. **44** (1978), no. 2, 129–162. MR0482230

[13] E. Lutz, Sur l'équation $y^2 = x^3 - Ax - B$ dans les corps p-adiques, J. Reine Angew. Math. **177** (1937), 238–247. MR1581558