Modular Forms, Elliptic Curves, and their Connection to Fermat's Last Theorem

Kyrie Johnson

ABSTRACT

Fermat's Last Theorem (FLT) states that if n is an integer greater than three, the equation $x^n + y^n = z^n$ has no integer solutions with $xyz \neq 0$. This incredible statement eluded proof for over three-hundred years: in that time, mathematicians developed numerous tools which finally proved FLT in 1995. In this paper, we introduce some of the essential objects which enter the proof — especially modular forms, elliptic curves, and Galois representations — with an emphasis on precisely stating the Shimura-Taniyama Conjecture and explaining how its proof finally settled FLT. We offer proofs whenever they clarify a definition or elucidate an idea, but generally prefer examples and exposition which make concrete a truly beautiful body of mathematical theory.

Contents

0 Introduction

Proven in 1995, Fermat's Last Theorem (FLT) remains a celebrity of twentieth century mathematics. FLT states that for $n \geq 3$, the equation $x^n + y^n = z^n$ has no integer solutions with $xyz \neq 0$. The saga of FLT began in 1637 when Pierre de Fermat — a French lawyer who enjoyed mathematics in his spare time — conjectured his theorem in the margin of an old math book. Fermat wrote that he had found a "truly marvelous proof" of the theorem, but that the book's margin was simply "to narrow to contain it". Because Fermat never published a formal proof — and it took mathematicians over three hundred years to devise one — it seems almost certain that Fermat never actually proved his own theorem. Nevertheless, he sparked a wildfire: countless mathematicians developed incredible mathematics in an effort to prove FLT.

Early attempts relied on a crucial simplifying observation: a non-trivial solution — i.e. a solution with $xyz \neq 0$ — to the equation $x^{pd} + y^{pd} = z^{pd}$ for the exponent pd yields a non-trivial solution $(x^d)^p + (y^d)^p = (z^d)^p$ for the exponent p. Because Fermat did indeed prove his theorem in the case of $n = 4$, the observation shows that proving FLT for any exponent $n \geq 3$ reduces to proving FLT for equations $x^p + y^p = z^p$ with p an odd prime. So the first progress on FLT involved checking the statement for $p = 3$ (by Euler in 1770), $p = 5$ (by Legendre and Dirichlet, independently, around 1825), and $p = 7$ (by Lamé in 1865).

The first substantial case of FLT came from the work of Sophie Germain in 1823. Germain introduced a much more general strategy for attacking the problem which ultimately split the problem into two cases:

1. Case 1 is the non-existence of $x^p + y^p = z^p$ for which p doesn't divide xyz; and
2. Case 2 is the same but when p does divide xyz.

Together with Legendre, Germain applied her techniques to prove the first case of FLT for all primes less than or equal to 197 (see [Rid09] for additional details). In particular, Germain showed that the first case of FLT holds for odd "Germain primes": an odd prime p such that $2p + 1$ is also prime. It remains unknown whether or not there exist infinitely-many Germain primes, so we still aren't sure if Germain's results gave an infinitude of results on FLT.

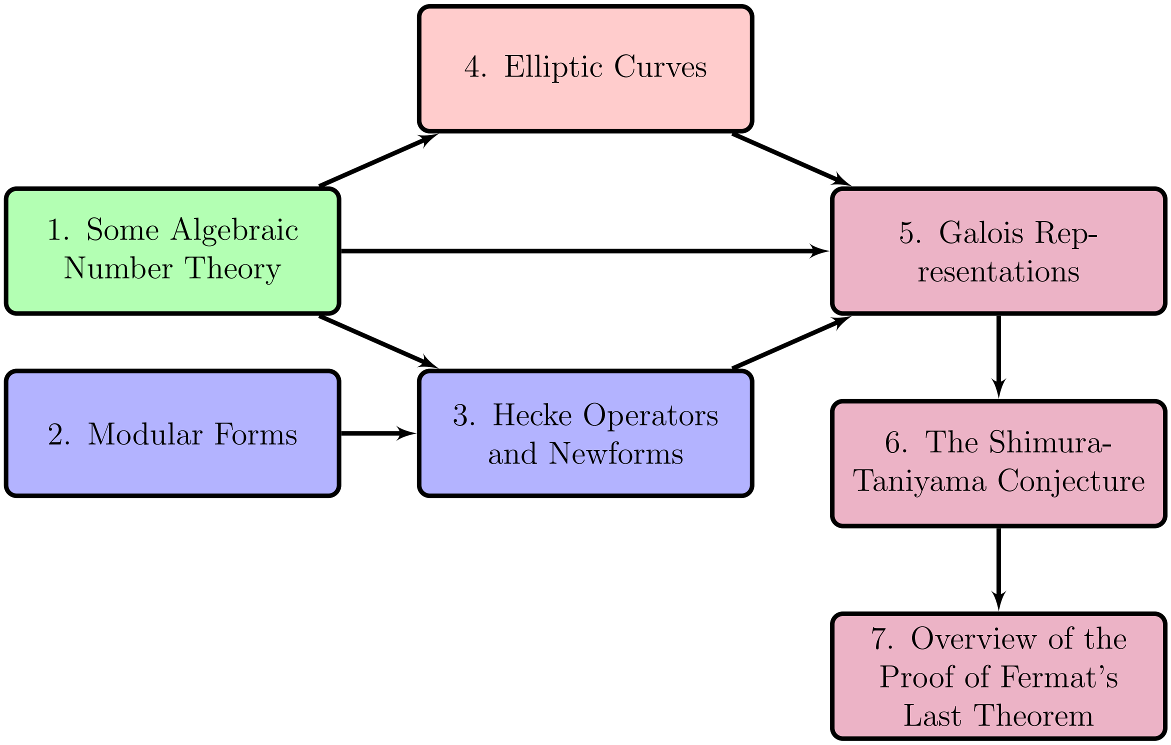
The next substantial case of FLT came from Ernst Kummer who drew upon ideas by Lamé on unique factorisation. Kummer proved (both case 1 and case 2) of FLT for so-called "regular primes" around 1850. The precise definition of a regular prime relies on the notion of factorisation of ring ideals; in section 1, we introduce some of these ideas (and defer to [Mil17] for the rest). Although certain heuristics predict that roughly 61% of primes are regular, whether or not there are infinitely-many remains a mystery. So two hundred years after Fermat conjectured his theorem, mathematicians still weren't sure about infinitely many cases.

The twentieth century witnessed the development of a wonderful body of mathematics which would go on to prove Fermat's Last Theorem. Of central importance are the ideas of Gorō Shimura and Yutaka Taniyama — who conjectured a precise relationship between "modular forms" and "elliptic curves" — and Jean-Pierre Serre — who conjectured a precise relationship between "modular forms" and "Galois representations". Ultimately, the link between modular forms and elliptic curves became an invaluable tool and by 1990 it was known that Fermat's Last Theorem would follow from the Shimura-Taniyama Conjecture. Andrew Wiles thus proved FLT by proving (most of) Shimura-Taniyama.

In this paper, we offer a broad overview of the twentieth century mathematics which proved FLT; we emphasise the role of the Shimura-Taniyama Conjecture (STC) in the proof and indeed develop the necessary language to precisely state STC. Along the way, we introduce some of number theory's most important tools and techniques. While we provide many proofs (especially in sections 2 and 5), we generally prefer illustrative examples to technical arguments.

Section 1 introduces some essential algebraic number theory (material from [Mil17] and [Mil18]), and is intended mostly as a reference for later sections. Section 2 introduces modular forms and their relationship with subgroups of $\mathrm{SL}_2(\mathbb{Z})$ (material from chapters 1 and 2 of [DS05]). Section 3 introduces Hecke operators as well as newforms, the modular forms which play a role in STC (material from chapter 5 of [DS05]). Section 4 introduces elliptic curves, including the Tate module of an elliptic curve and reduction of curves over $\mathbb{Q}$ (material from chapters 3 and 7 of [Sil86]).

Section 5 shifts to the construction of Galois representations for both newforms and elliptic curves, as well as discusses the structure of the absolute Galois group $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ (material from chapter 8 of [Mil17] and chapter 9 of [DS05]). Finally, section 6 states the Shimura-Taniyama Conjecture and goes through an explicit example to illustrate the statement (material from chapter 9 of [DS05] and [Wes99]). Section 7 then brings everything together by outlining the stepping stones which go into proving FLT (material from [DDT07]).


4. Elliptic Curves
1. Some Algebraic Number Theory
5. Galois Representations
2. Modular Forms
3. Hecke Operators and Newforms
6. The Shimura-Taniyama Conjecture
7. Overview of the Proof of Fermat's Last Theorem

1 Some Algebraic Number Theory

Developed throughout the nineteenth and early twentieth century, basic algebraic number theory forms the foundation for not only Lamé and Kummer's early attempts at FLT, but also for the Shimura-Taniyama Conjecture and Wiles's ultimate proof. In this section, we highlight the essential results we will need later, and intend that this section serves only as a reference for subsequent sections.

1.1 Unique Factorisation and the Ring of Integers

One of the greatest structural features of the integers is that they permit unique factorisation. Indeed, a common strategy for attacking problems over $\mathbb{Z}$ is to first consider the problem for irreducible (prime) integers — which are often easier to understand — before then assembling information for a general integer from information on its prime divisors. But unique factorisation fails in rings "not too much larger" than the integers: for example, $6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$ in $\mathbb{Z}[\sqrt{-5}]$. Roughly speaking, the algebraic number theory of the late 1800s worked to recover a notion of unique factorisation in more general rings. Ultimately, the solution is to consider factorisation of ring ideals, rather than of ring elements.

We begin with some essential algebraic notions. Throughout this section, we will work over $\mathbb{Q}$ — as $\mathbb{Q}$ is our principal concern as number theorists — keeping in mind that many of these ideas naturally generalise to arbitrary rings/fields.

Definition 1.1. A number $\alpha \in \mathbb{C}$ is an **algebraic number** if there exists a polynomial

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0, a_i \in \mathbb{Q}$$

such that $f(\alpha) = 0$.

We denote by $\overline{\mathbb{Q}}$ the set of all algebraic numbers, where the notation comes from the algebraic closure of a field; that is, $\overline{\mathbb{Q}}$ may equivalently be thought of as the algebraic closure of $\mathbb{Q}$, just as $\mathbb{C}$ is the algebraic closure of $\mathbb{R}$. Similarly, we define an algebraic integer by replacing $\mathbb{Q}$ with $\mathbb{Z}$:

Definition 1.2. A number $\alpha \in \mathbb{C}$ is an **algebraic integer** if there exists a polynomial

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0, a_i \in \mathbb{Z}$$

such that $f(\alpha) = 0$.

We denote by $\overline{\mathbb{Z}}$ the set of all algebraic integers, where the notation comes from a more general notion of "integral closure". Finally, rather than consider all algebraic numbers $\overline{\mathbb{Q}}$, we will want to consider finite subfields.

Definition 1.3. Consider the tower of field extensions $\mathbb{Q} \subset K \subset \overline{\mathbb{Q}}$. We call K an **(algebraic) number field** if it is a finite algebraic extension of $\mathbb{Q}$. In this case, we call $\mathcal{O}_K := K \cap \overline{\mathbb{Z}}$ the ring of integers of K.

As hinted at in the introductory paragraph, the terminology comes from the prototypical example: $\mathbb{Z}$ is the ring of integers in $\mathbb{Q}$. Indeed, let $\frac{p}{q} \in \mathbb{Q}$ satisfy the polynomial $f(x) = x^n + \sum_{k=0}^{n-1} a_k x^k$. Then

$$f\left(\frac{p}{q}\right) = \frac{p^n}{q^n} + \sum_{k=0}^{n-1} a_k \frac{p^k}{q^k} = 0 \implies p^n + \sum_{k=0}^{n-1} a_k p^k q^{n-k} = 0 \implies q \text{ divides } p$$

so an integral element $\frac{p}{q}$ is indeed an integer. The following two theorems summarise the results we need.

Theorem 1.4. *Let L be an algebraic number field with ring of integers $\mathcal{O}_L$. Then any ideal $\mathfrak{a} \subset \mathcal{O}_L$ uniquely factors into a unique, finite product*

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

with each $\mathfrak{p}_i$ a prime ideal and $e_i \geq 1$. Moreover, every prime ideal $\mathfrak{p}_i$ is maximal.

Theorem 1.5. *Let $\mathbb{Q} \subset K \subset L$ be a tower of algebraic number field with ring of integers $\mathcal{O}_K$ and $\mathcal{O}_L$. Further, let $\mathfrak{p}$ denote a prime (maximal) ideal in $\mathcal{O}_K$ and $\mathbf{k}_\mathfrak{p}$ the field $\mathcal{O}_K/\mathbf{k}_\mathfrak{p}$. By the previous theorem, we have a unique factorisation*

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

into a product of prime ideals $\mathfrak{p}_i \subset \mathcal{O}_L$. Then letting $\mathbf{l}_i$ be the field $\mathcal{O}_L/\mathfrak{p}_i$ and f_i the index $[\mathbf{l}_i : \mathbf{k}_\mathfrak{p}]$, we have

$$\sum_{k=1}^{g} e_i f_i = [L : K].$$

Moreover, if L/K is Galois, then $e_i = e_j =: e$ and $f_i = f_j =: f$ for all i, j; in particular, $efg = [L : K]$.

In theorem 1.4, we have the promised result: unique factorisation of ideals. In theorem 1.5, we control the way an ideal may factor when we promote it to an ideal in a larger ring of integers. Momentarily, we will give names to various types of factorisations as well as specialise to the case when L/K is Galois. But we first return to our motivating example: $\mathbb{Z}[\sqrt{-5}]$.

Recall that $6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$ represents a failure of unique factorisation of elements in $\mathbb{Z}[\sqrt{-5}]$. As in theorem 1.5, with $K = \mathbb{Q}$, $\mathcal{O}_K = \mathbb{Z}$, $L = \mathbb{Q}[\sqrt{-5}]$, and $\mathcal{O}_K = \mathbb{Z}[\sqrt{-5}]$, we may regard (2) and (5) — prime ideals in $\mathcal{O}_K$ — as ideals in $\mathcal{O}_L$. In this case, we obtain factorisations

$$(2) = (2, 1 + \sqrt{-5})^2$$
$$(3) = (3, 1 + \sqrt{-5})(3, 1 - \sqrt{-5}).$$

Moreover, the ideals $(1+\sqrt{-5}), (1-\sqrt{-5}) \subset \mathcal{O}_L$ factor as

$$(1+\sqrt{-5}) = (2, 1+\sqrt{-5})(3, 1+\sqrt{-5})$$
$$(1-\sqrt{-5}) = (2, 1+\sqrt{-5})(3, 1-\sqrt{-5})$$

so we obtain a completely unique factorisation of the ideal $(6) \subset \mathbb{Z}[\sqrt{-5}]$:

$$(6) = (2, 1+\sqrt{-5})^2(3, 1+\sqrt{-5})(3, 1-\sqrt{-5}).$$

In this example, we appealed to an extremely crucial case of the theorem 1.5: when $K = \mathbb{Q}$ and L is some number field. We will almost exclusively consider this case.

Now, we enumerate various types of factorisations:

Definition 1.6. Use the setup as in theorem 1.5. If $g = 1$ and $e_i = 1$, then we have

$$\mathfrak{p}\mathcal{O}_L = \mathfrak{p}_1$$

so $\mathfrak{p}$ remains prime in $\mathcal{O}_L$ and we say that $\mathfrak{p}$ is **inert**. If $g = [L : K]$ — so that $e_i = 1 = f_i$ for all i — then we have

$$\mathfrak{p}\mathcal{O}_L = \prod_{i=1}^{g} \mathfrak{p}_\mathfrak{i}$$

and we say that $\mathfrak{p}$ **splits** (or **splits completely**). Finally, if there is some j such that $e_j > 1$, we say that $\mathfrak{p}$ **ramifies** in $\mathcal{O}_L$. In all cases, we say that the primes $\mathfrak{p}_i$ **divide** $\mathfrak{p}$ — or that the primes $\mathfrak{p}_i$ **lie above** $\mathfrak{p}$ — and naturally denote this by $\mathfrak{p}_i|\mathfrak{p}$.

So for $L = \mathbb{Q}[\sqrt{-5}]$ the number field from before, we see that the ideal (2) ramifies in L while the ideal (3) splits completely. Contrast this with the situation for $\mathbb{Z}[i]$, the ring of integers in $L = \mathbb{Q}[i]$. In this case, $(2) = (2, 1+i)^2$ is the only prime to ramify, while (3) remains inert and $(5) = (5, 2+i)(5, 2-i)$ splits completely. In particular, $(5, 2+i)$ lies above (5) in $\mathbb{Q}[i]$.

For our purposes, ramification will be the case of greatest interest, owing in large part to the rarity of ramification:

Theorem 1.7. *Use the setup as in theorem 1.5. Let $R(\mathcal{O}_L) = \{\mathfrak{p} \subset \mathcal{O}_K : \mathfrak{p}$ ramifies in $\mathcal{O}_L\}$. Then $R(\mathcal{O}_L)$ is finite; succinctly, only finitely-many primes ramify.*

For our favourite example $L = \mathbb{Q}[\sqrt{-5}]$ over $\mathbb{Q}$, we've previously seen that (2) ramifies; the only other prime to ramify is $(5) = (5, \sqrt{-5})^2$. In fact, for any square-free integer D, the quadratic extension $\mathbb{Q}[\sqrt{D}]$ has ramification exclusively at the primes dividing D and sometimes at (2).

1.2 Inverse Limits and the Absolute Galois Group

Recall that a field extension L/K is Galois if the extension is both normal — every irreducible polynomial over K either (i) remains irreducible over L or (ii) splits completely over L — and separable — every minimal polynomial over K of an element of L is separable. In particular, a field of characteristic zero is automatically separable, so a characteristic-zero field extension L/K is Galois if it is normal. It follows that the extension $\overline{\mathbb{Q}}/\mathbb{Q}$ is Galois, so it makes sense to speak of the extension's Galois group.

Definition 1.8. The group $G_{\mathbb{Q}} := \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) := \mathrm{Aut}(\overline{\mathbb{Q}}/\mathbb{Q})$ is called the **absolute Galois group (of $\mathbb{Q}$)**.

We will realise the absolute Galois group as a natural limit of finite-degree Galois groups. Along the way, we'll define a construction called the inverse limit, a fundamental technique in mathematics and especially useful for defining a number of objects we'll need later.

Consider the tower of field extensions $\overline{\mathbb{Q}}/L/\mathbb{Q}$ with $L/\mathbb{Q}$ a Galois number field. Then there is a natural surjection $G_{\mathbb{Q}} \to \mathrm{Gal}(L/\mathbb{Q})$ given by restriction: for any $\sigma \in G_{\mathbb{Q}}$, define $\sigma_L \in \mathrm{Gal}(L/\mathbb{Q})$ by $\sigma_L := \sigma|_L$. Conversely, given all such σ_L, we can reconstruct σ: for any $s \in \overline{\mathbb{Q}}$, pick a Galois number field L such that $s \in L$ so that $\sigma(s)$ must be $\sigma_L(s)$. We thus have a natural pairing between elements of $\sigma \in G_{\mathbb{Q}}$ and collections $\{\sigma_{L_i}\}_{i\in I}$ — where I is an index set and L_i ranges over the Galois number fields $L_i/\mathbb{Q}$ — such that

- for all i and j, the automorphism $\sigma_{L_i} \in \mathrm{Gal}(L_i/\mathbb{Q})$, and $\sigma_{L_j} = \sigma_{L_j}$ on $L_i \cap L_j$; and
- the automorphism σ restricts to σ_{L_i} on L_i.

The first bullet is a sort of "compatibility" requirement which forces automorphisms in distinct Galois groups to knit together in a natural way. The second bullet connects elements of the massive Galois group $G_{\mathbb{Q}}$ to elements of "small", finite extensions. This is a special case of the following construction.

Definition 1.9. Let $(I, \leq)$ be a directed partially-ordered set. Let $\{G_i\}_{i\in I}$ be a collection of groups with maps $r_{j,i} : G_j \to G_i$ such that $r_{j,i} \circ r_{k,j} = r_{k,i}$ for all $k \geq j \geq i$. The pair $(G_i)_{i\in I}$ and $(r_{j,i})_{i,j\in I}$ make up an **inverse system** of groups and bonding morphisms over I.

Definition 1.10. The **inverse limit** of an inverse system $(G_i)_{i\in I}$ and $(r_{j,i})_{i,j\in I}$ is defined by

$$\varprojlim_{i\in I} G_i = \left\{ (a_i) \in \prod_{i\in I} G_i : r_{j,i}(a_j) = a_i \text{ for all } j \geq i \right\},$$

a subgroup of the direct product $\prod_{i\in I} G_i$. The condition that $r_{j,i}(a_j) = a_i$ is called the **compatibility condition** of the system as it ensures that the elements of the inverse system are compatible with the reduction maps. If each G_i has a topology, then we endow $\varprojlim_{i\in I} G_i$ with the subspace topology of the product topology on $\prod_{i\in I} G_i$.

Let's apply this definition to our motivating example $G_{\mathbb{Q}}$. In this case, our partially ordered set is the set $\{L_i\}_{i\in I}$ of all Galois number fields $L_i/\mathbb{Q}$ with an ordering $\leq$ given by

$$L_i \leq L_j \iff L_i \subset L_j.$$

Further, set $G_i := \mathrm{Gal}(L_i/\mathbb{Q})$ and for $L_j \geq L_i$ define the bonding morphism $r_{j,i} : L_j \to L_i$ by restriction. Then any $L_k \geq L_j \geq L_i$ satisfy $r_{j,i} \circ r_{k,j} = r_{k,i}$ and we indeed have an inverse system. The following diagram shows an excerpt of this massive inverse system, where the notation $\mathrm{Gal}(L)$ denotes the Galois group $\mathrm{Gal}(L/\mathbb{Q})$ for various number fields L.

$$\begin{tikzcd}
\ddots\vdots\iddots \arrow[dd] & \iddots \arrow[d] \arrow[ddl, bend right] & \ddots\vdots\iddots & \ddots\vdots\iddots & \\
& \mathrm{Gal}(\mathbb{Q}(\sqrt{2},\sqrt{3})) \arrow[dd] \arrow[ddr] & \mathrm{Gal}(\mathbb{Q}(\sqrt{2},\sqrt{5})) \arrow[ddl] \arrow[ddr] & \mathrm{Gal}(\mathbb{Q}(e^{2\pi i/5})) \arrow[dd] & \\
\mathrm{Gal}(\mathbb{Q}(\cos\left(\frac{2\pi}{7}\right))) \arrow[ddrr, bend right] & & & & \ddots\vdots\iddots \arrow[d] \\
& \mathrm{Gal}(\mathbb{Q}(\sqrt{2})) \arrow[dr] & \mathrm{Gal}(\mathbb{Q}(\sqrt{3})) \arrow[d] & \mathrm{Gal}(\mathbb{Q}(\sqrt{5})) \arrow[dl] & \cdots \arrow[dll, bend left] \\
& & \mathrm{Gal}(\mathbb{Q}) & &
\end{tikzcd}$$

The earlier discussion (together with Zorn's Lemma if we want to be entirely formal) justifies that

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

so, as promised, we have realised $G_{\mathbb{Q}}$ as a limit of finite Galois groups. Momentarily, we will discuss the topology this yields. But we first discuss another essential example.

Example 1.11. Let I denote the set of positive integers with their natural ordering and fix ℓ an integer prime. For $n \in I$, set $G_n := \mathbb{Z}/\ell^n\mathbb{Z}$ and for $n \geq m$ define $r_{n,m} : \mathbb{Z}/\ell^n\mathbb{Z} \to \mathbb{Z}/\ell^m\mathbb{Z}$ by reduction mod ℓ^m. Then we have an inverse system

$$\mathbb{Z}/\ell\mathbb{Z} \leftarrow \mathbb{Z}/\ell^2\mathbb{Z} \leftarrow \mathbb{Z}/\ell^3\mathbb{Z} \leftarrow \cdots \leftarrow \mathbb{Z}/\ell^n\mathbb{Z} \leftarrow \cdots$$

where all maps are given by reduction mod some power of ℓ. But this time we get an object we have not yet encountered:

$$\mathbb{Z}_\ell := \varprojlim_{n\in I} \mathbb{Z}/\ell^n\mathbb{Z}.$$

An important structural feature of $\mathbb{Z}_\ell$ is that we have an embedding $\mathbb{Z} \hookrightarrow \mathbb{Z}_\ell$: send an integer a to the sequence $(a, a, a, a, a, \dots)$ where n^{th} entry "a" denotes the reduction of a mod ℓ^n. The sequence is in $\mathbb{Z}_\ell$ because we certainly have the compatibility condition

$$r_{m,n}(a \mod \ell^m) = a \mod \ell^n$$

for all $m \geq n$. Moreover, the map is injective because an integer a such that $a \equiv 0 \mod \ell^n$ for all n must itself equal 0. So we indeed have a (canonical) embedding of $\mathbb{Z}$ into $\mathbb{Z}_\ell$.

For an example of a non-integer element of $\mathbb{Z}_\ell$, consider the sequence $(a_1, a_2, a_3, a_4, \dots)$ given by

$$a_n = 1 + \ell + \cdots + \ell^{n-1}$$

so that $a_1 = 1$, $a_2 = 1 + \ell$, $a_3 = 1 + \ell + \ell^2$, and so forth. For $n \geq m$, reducing a_n mod ℓ^m yields a_m so once again the sequence satisfies the compatibility condition and is in $\mathbb{Z}_\ell$. For $\ell > 2$, the sequence does not represent an integer because it never stabilises (when $\ell = 2$, the sequence represents -1 under the aforementioned embedding $\mathbb{Z} \hookrightarrow \mathbb{Z}_\ell$).

Definition 1.12. The object $\mathbb{Z}_\ell$ constructed in the preceding example is the **ring of ℓ-adic integers**. Its field of fractions is $\mathbb{Q}_\ell$, the **field of ℓ-adic numbers**. As the names suggest, $\mathbb{Z}_\ell$ is in fact the ring of integers inside $\mathbb{Q}_\ell$.

If we endow each finite Galois group $\mathrm{Gal}(L_i/\mathbb{Q})$ with the discrete topology, for L_i a Galois number field, then the inverse limit $\varprojlim_i \mathrm{Gal}(L_i/\mathbb{Q})$ equips $G_\mathbb{Q}$ with its own topology. We call this topology the Krull topology on $G_\mathbb{Q}$ and for our purposes we only need a particularly important collection of open subgroups in $G_\mathbb{Q}$.

Definition 1.13. Let X be a topological space and fix $x \in X$. A **neighbourhood base** $\mathcal{N}$ for x is a set of (open) neighbourhoods of x such that any neighborhood U of x contains some $N \in \mathcal{N}$.

Theorem 1.14. *In the Krull topology on $G_\mathbb{Q}$, the collection*

$$\{\mathrm{Gal}(\overline{\mathbb{Q}}/M) : M/\mathbb{Q} \textit{ finite and Galois}\}$$

constitutes a neighbourhood base of the identity. In particular, the subgroups $\mathrm{Gal}(\overline{\mathbb{Q}}/M)$ *are open for $M/\mathbb{Q}$ finite and Galois.*

Our ultimate goal is to associate to "elliptic curves" and "modular forms" certain continuous maps $G_\mathbb{Q} \to \mathrm{GL}_2(\mathbb{Q}_\ell)$. Continuity will be determined with respect to the Krull topology on $G_\mathbb{Q}$ and the subgroups in theorem 1.14 will figure prominently. But before we do any of that we must first define both modular forms (section 2) and elliptic curves (section 4).

2 Modular Forms

Broadly speaking, modular forms are functions on the upper-half of the complex plane which are holomorphic and exhibit significant symmetry. To make precise this notion of symmetry, we'll study an important group action before then linking the group action to modular forms.

2.1 The Group Action

We begin by defining an essential group action, which appears not only in the theory of modular forms, but also in complex analysis, hyperbolic geometry, geometric group theory, the theory of continued fractions, and elsewhere. Our acting group will be $\mathrm{SL}_2(\mathbb{Z})$, the group of 2 by 2 integer matrices with determinant one:

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

In the context of modular forms, we will often refer to $\mathrm{SL}_2(\mathbb{Z})$ as the "modular group". Our set on which the modular group will act is $\mathbb{H}^* := \mathbb{H} \cup \mathbb{Q} \cup \{\infty\}$: the upper-half of the complex plane,

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

together with its "rational limits", the points $\mathbb{Q}$ and infinity.

In words, we will refer to $\mathbb{H}^*$ as the "extended upper-half-plane". To distinguish elements of $\mathbb{H}$, we will always use τ to denote a complex number with positive imaginary part (unless otherwise specified); similarly, we will always use s to denote a point of $\mathbb{Q} \cup \{\infty\}$ and use γ to denote a matrix in $\mathrm{SL}_2(\mathbb{Z})$. The action takes place as follows: a matrix $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$ acts on $\mathbb{H}$ by taking τ to

$$\gamma(\tau) := \frac{a\tau + b}{c\tau + d}. \tag{1}$$

Because

$$\mathrm{Im}(\gamma(\tau)) = \frac{\mathrm{Im}(\tau)}{|c\tau + d|^2},$$

that the $\mathrm{Im}(\tau)$ is greater than 0 implies that $\mathrm{Im}(\gamma(\tau)) > 0$. So the action indeed sends elements of $\mathbb{H}$ to elements of $\mathbb{H}$. Analogously, a matrix $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$ acts on $\mathbb{Q} \cup \{\infty\}$ by taking s to

$$\gamma(s) := \frac{a\tau + s}{cs + d}.$$

In particular, ∞ maps to the rational number $\frac{a}{c}$ unless $c = 0$, in which case ∞ maps to itself. Similarly, a number $s \in \mathbb{Q}$ maps to another rational number unless $cs + d = 0$, in which case s maps to ∞. This shows that the action of $\mathrm{SL}_2(\mathbb{Z})$ not only shuffles around the points of $\mathbb{H}$, but also shuffles around the rational limits $\mathbb{Q} \cup \{\infty\}$.

Notice that for all $z \in \mathbb{H}^*$, the identity $I := \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ satisfies $I(z) = z$. Thus, for $z \mapsto \gamma(z)$ to define a group action, one need only check that $\gamma_1(\gamma_2(z)) = (\gamma_1\gamma_2)(z)$. Our main objective is to define and study "modular forms", holomorphic functions which play nicely with the action of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathbb{H}^*$. But first we study the action of the modular group in its own right for which we introduce an incredible lemma regarding the structure of $\mathrm{SL}_2(\mathbb{Z})$.

Lemma 2.1. *The modular group* $\mathrm{SL}_2(\mathbb{Z})$ *is generated by the matrices* $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ *and* $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$.

Proof. Take $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$ and let S denote the group of matrices generated by $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$; we will exhibit a sequence of matrices in S which, by right multiplication, take $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ to a matrix in S. This will prove that $\mathrm{SL}_2(\mathbb{Z}) \subset S$; because $S \subset \mathrm{SL}_2(\mathbb{Z})$ automatically, the claim will follow.

Now, note that $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)^n = \left(\begin{smallmatrix} 1 & n \\ 0 & 1 \end{smallmatrix}\right) \in S$. Computing

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a & an + b \\ c & cn + d \end{pmatrix} \tag{2}$$

and

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} b & -a \\ d & -c \end{pmatrix} \tag{3}$$

illustrates how the matrices in S act by right multiplication. In particular, we may choose n in equation 2 such that $\left(\begin{smallmatrix} a_1 & b_1 \\ c_1 & d_1 \end{smallmatrix}\right) := \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)\left(\begin{smallmatrix} 1 & n \\ 0 & 1 \end{smallmatrix}\right)$ satisfies $0 \leq |d_1| < |c|$; in fact, we may further apply equation 3 to flip the values of c_1 and d_1 so that $0 \leq |c_1| < |c|$. Repeating this process on $\left(\begin{smallmatrix} a_1 & b_1 \\ c_1 & d_1 \end{smallmatrix}\right)$, we obtain a matrix $\left(\begin{smallmatrix} a_2 & b_2 \\ c_2 & d_2 \end{smallmatrix}\right)$ such that $0 \leq |c_2| < |c_1| < |c|$. In this way, we obtain a sequence of matrices $\left(\begin{smallmatrix} a_i & b_i \\ c_i & d_i \end{smallmatrix}\right)$, $1 \leq i \leq k$ such that $|c_k| < |c_{k-1}| < \cdots < |c_1| < |c|$. Because all of the inequalities are strict, the process eventually terminates with a matrix where $c_k = 0$.

So by repeated right multiplication of matrices from S we have obtained the matrix $\left(\begin{smallmatrix} a_k & b_k \\ 0 & d_k \end{smallmatrix}\right)$, which necessarily has determinant one. This forces $a_k = d_k = \pm 1$. Because $\left(\begin{smallmatrix} \pm 1 & b_k \\ 0 & \pm 1 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 & b_k \\ 0 & 1 \end{smallmatrix}\right)\left(\begin{smallmatrix} \pm 1 & 0 \\ 0 & \pm 1 \end{smallmatrix}\right)$ is a product of matrices in S, the matrix $\left(\begin{smallmatrix} \pm 1 & b_k \\ 0 & \pm 1 \end{smallmatrix}\right)$ is itself in S and we're done. □

We will use knowledge of these generators to make sense of the modular group's action on $\mathbb{H}$. In particular, let $T_n := \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)^n = \left(\begin{smallmatrix} 1 & n \\ 0 & 1 \end{smallmatrix}\right)$, any $n \in \mathbb{Z}$ — noting that this generator has infinite order in $\mathrm{SL}_2(\mathbb{Z})$ — and let $F := \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ — noting that this generator has order four in $\mathrm{SL}_2(\mathbb{Z})$ but has order two as an action on $\mathbb{H}^*$. Then for $\tau \in \mathbb{H}$, equation (1) yields $T_n(\tau) = \tau + n$ so T_n acts on $\mathbb{H}$ as a translation by n. Similarly, $F(\tau) = -\frac{1}{\tau}$, so $|F(\tau)| = \frac{1}{|\tau|}$ and F "flips" elements of $\mathbb{H}$ over the upper-half of the circle $\{z \in \mathbb{C} : |z| = 1\}$: more precisely, for $z = re^{i\theta}$ with $\theta \in (0, \pi)$ and $r > 0$, we have $F(z) = \frac{1}{r}e^{i(\pi - \theta)} \in \mathbb{H}$. These observations break $\mathbb{H}$ into some natural regions. First, we have the strip $\{\tau \in \mathbb{H} : -\frac{1}{2} < \tau \leq \frac{1}{2}\}$ which covers all of $\mathbb{H}$ by the translations T_n. Second, we have the region $\{\tau \in \mathbb{H} : |\tau| \geq 1\}$ which covers $\mathbb{H}$ after flipping once by F. Hopefully this serves to motivate the following proposition.

Proposition 2.2. *Let $D := \{\tau \in \mathbb{H} : -\frac{1}{2} \leq \mathrm{Re}(\tau) \leq \frac{1}{2}, |\tau| \geq 1\}$, as pictured in figure 1. Let L denote the part of the boundary of D with real part less than 0 and R the same but larger than 0. Then*

(a) for all $\tau \in \mathbb{H}$, there exists $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ and $r \in D$ such that $\gamma(\tau) = r$; and

(b) modding D by the action of $\mathrm{SL}_2(\mathbb{Z})$ identifies L and R, and identifies nothing else.

Proof. See the proofs of lemmas 2.3.1 and 2.3.2 in [DS05]. □

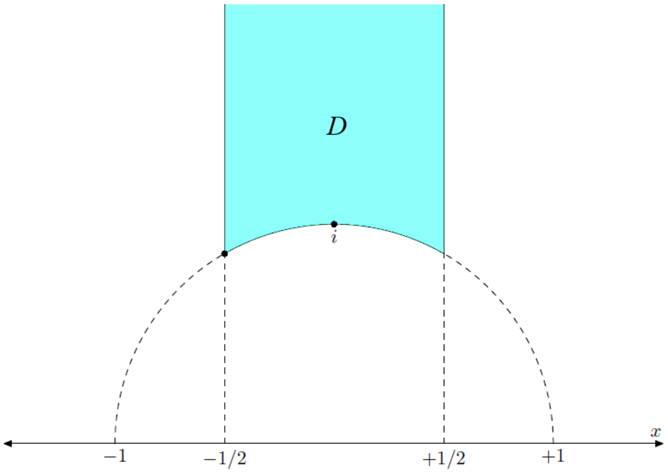


Figure 1: A fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$. The lower-left boundary point of D is a primitive sixth root of unity. Image adapted from [Sch18].

We call the region D a fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$, so-called because it bears a unique (up to boundary identifications) representative from the orbit of any $\tau \in \mathbb{H}$.

2.2 An Early Definition and Examples

For now, we'll set aside the action on $\mathbb{Q} \cup \{\infty\}$ and return to it in section 2.3. With the action of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathbb{H}$, however, we proceed to the notion of a modular form, a function which plays nicely with this action. For this we need some handy notation.

Definition 2.3. Let f be a function from $\mathbb{H}$ to $\mathbb{C}$ and $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$. For $k \in \mathbb{Z}$, define the **weight-k operator** $[\gamma]_k$ by defining its action $f[\gamma]_k$ on f:

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

And now we can state the definition of a modular form.

Definition 2.4. A function $f : \mathbb{H} \to \mathbb{C}$ is a **modular form of weight-k (with respect to $\mathrm{SL}_2(\mathbb{Z})$)** if

(i) f is holomorphic on $\mathbb{H}$;

(ii) the limit $\lim_{\mathrm{Im}(\tau)\to\infty} f(\tau)$ exists and is finite;

(iii) and $f(\tau) = f[\gamma]_k(\tau)$ — so $f(\gamma(\tau)) = (c\tau + d)^k f(\tau)$ — for all $\gamma \in \mathrm{SL}_2(\mathbb{Z})$.

In a certain precise sense, condition (ii) means that f is "holomorphic at infinity", in line with our increasing regard for infinity as a point in its own right. In particular, the fundamental domain D is not compact under the subspace topology inherited from $\mathbb{C}$; adding the point at infinity (as well as adding some natural open sets to the topology) will make $D \cup \{\infty\}$ compact. We thus require condition (ii) so that modular forms make sense on the (nicer) compact sets. To motivate condition (iii), consider the cases of $k = 0$ and $k = 2$. A weight-0 modular form f satisfies $f(\gamma(\tau)) = f(\tau)$ for all $\gamma \in \mathrm{SL}_2(\mathbb{Z})$; as such, weight-0 modular forms give $\mathrm{SL}_2(\mathbb{Z})$-invariant functions on $\mathbb{H}$. A weight-2 modular form f satisfies $f(\gamma(\tau)) = (c\tau + d)^2 f(\tau)$; computing

$$\int f(\gamma(\tau))\, d(\gamma(\tau)) = \int (c\tau + d)^2 f(\tau) \frac{1}{(c\tau + d)^2}\, d\tau = \int f(\tau)\, d\tau$$

shows that weight-2 modular forms give $\mathrm{SL}_2(\mathbb{Z})$-invariant integration on $\mathbb{H}$. Indeed, the proof of the Modularity Theorem — a key ingredient in the proof of Fermat's Last Theorem — necessitates the theory of weight-2 modular forms.

Naturally, $f = 0$ is a weight-k modular form for all k. A weight-0 modular form f satisfies $f(\tau) = f(\gamma(\tau))$ for all $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ and all $\tau \in \mathbb{H}$. In particular, we can regard f as a function on the space $\mathbb{H} \cup \{\infty\}$ mod the action of $\mathrm{SL}_2(\mathbb{Z})$. We will think more about this space later; for now, take for granted that it is compact (as discussed) "complex one-manifold". Just as any holomorphic function on $\mathbb{C} \cup \{\infty\}$ is constant (by Liouville's Theorem), a holomorphic function on a compact one-manifold is constant. So $f = c$ for some $c \in \mathbb{C}$, and all weight-0 modular forms are constant.

For more interesting examples of modular forms, we consider so-called Eisenstein series.

Definition 2.5. For k an integer larger than 2, the **weight-k Eisenstein series**, denoted $G_k(\tau)$, is

$$G_k(\tau) = \sum_{(c,d)\in\mathbb{Z}^2}{}' \frac{1}{(c\tau+d)^k}$$

where the prime on the summand denotes that the sum excludes $(c,d) = (0,0)$.

Because we require $k > 2$, the sum defining an Eisenstein series converges absolutely on $\mathbb{H}$, converges uniformly on compact subsets of $\mathbb{H}$, and is bounded as the $\mathrm{Im}(\tau)$ approaches infinity (refer to exercise 1.1.4 in [DS05] for a proof). By absolute convergence, we may freely rearrange the terms in the sum: using this fact and applying the definition of the weight-k operator, it follows that $G_k[\gamma]_k(\tau) = G_k(\tau)$ for any $\gamma \in \mathrm{SL}_2(\mathbb{Z})$. As such, the Eisenstein series give examples of modular forms of weight-k, as promised. Note, however, that when k is odd, $G_k(\tau) = 0$, so we obtain interesting examples for $k \geq 4$ and even. Indeed, this reflects a more general fact: there are no modular forms of odd weight with respect to all of $\mathrm{SL}_2(\mathbb{Z})$. To see this, note that $-I = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$ and the condition $f(\tau) = f[-I]_k(\tau)$ forces $f(\tau) = (-1)^k f(\tau)$ for all τ.

The Eisenstein series are of particular use as a way of generating other functions of interest.

Definition 2.6. Set $g_2(\tau) := 60G_4(\tau)$ and $g_3(\tau) := 140G_6(\tau)$. Then we define the **discriminant function** $\Delta : \mathbb{H} \to \mathbb{C}$, also called the Ramanujan-Delta function, by

$$\Delta(\tau) := g_2(\tau)^3 - 27g_3(\tau)^2$$

a modular form of weight-12. Similarly, we construct the **j-function** $j : \mathbb{H} \to \mathbb{C}$ by defining

$$j(\tau) := \frac{g_2(\tau)^3}{\Delta(\tau)},$$

a "meromorphic modular form" of weight-0; although j satisfies conditions (i) and (iii) to be a modular form, Δ has exactly one zero at infinity (and g_2 does not vanish at infinity) so j has a pole at infinity. These functions arise in the study of "elliptic curves", where they correspond to the "discriminant" and "j-invariant", respectively, of elliptic curves over $\mathbb{C}$.

2.3 Congruence Subgroups

So far, we have only considered modular forms of weight-k with respect to the entire modular group. We will ultimately want a slightly more refined notion of modular form, in which we restrict to the action of certain subgroups of $\mathrm{SL}_2(\mathbb{Z})$. To motivate this transition we introduce a delightful problem in classical number theory: the four squares problem. Let $r(n)$ denote the cardinality of $\{(x_1, x_2, x_3, x_4) \in \mathbb{Z}^4 : n = x_1^2 + x_2^2 + x_3^2 + x_4^2\}$. Then the function

$$f(\tau) := \sum_{n=0}^{\infty} r(n)e^{2\pi i n\tau}$$

satisfies conditions (i) and (ii) to be a modular form (in fact satisfies a stronger condition than condition (ii), as we will discuss), but doesn't quite satisfy condition (iii) – see [DS05] section 1.2 for details. Nevertheless, $f(\tau)$ does satisfy the transformation laws

$$f(\tau+1) = f(\tau)$$

and

$$f\left(\frac{\tau}{4\tau+1}\right) = (4\tau+1)^2 f(\tau)$$

so $f(\gamma\tau) = f[\gamma]_2$ for $\gamma = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and $\gamma = \left(\begin{smallmatrix} 1 & 0 \\ 4 & 1 \end{smallmatrix}\right)$; let Γ_f denote the subgroup of $\mathrm{SL}_2(\mathbb{Z})$ generated by these two matrices. Then f satisfies condition (iii) when we replace $\mathrm{SL}_2(\mathbb{Z})$ with Γ_f, which motivates the following definition.

Definition 2.7. Let Γ be a subgroup of $\mathrm{SL}_2(\mathbb{Z})$ and for all $s \in \mathbb{Q} \cup \{\infty\}$, let $\alpha_s \in \mathrm{SL}_2(\mathbb{Z})$ be such that $\alpha_s(s) = \infty$. A function $f : \mathbb{H} \to \mathbb{C}$ is a **modular form of weight-k with respect to Γ** if

(i) f is holomorphic on $\mathbb{H}$;

(ii) the limit $\lim_{\mathrm{Im}(\tau)\to\infty} f[\alpha_s]_k(\tau)$ exists and is finite for all s;

(iii) and $f(\tau) = f[\gamma]_k(\tau)$ — so $f(\gamma(\tau)) = (c\tau+d)^k f(\tau)$ — for all $\gamma \in \Gamma$.

We tweak condition (ii) so that f isn't just holomorphic at infinity — as we required for a modular form with respect to $\mathrm{SL}_2(\mathbb{Z})$ — but is also holomorphic at all other rational limits $\mathbb{Q} \cup \{\infty\}$. When dealing with $\mathrm{SL}_2(\mathbb{Z})$, it makes no difference because the action of $\mathrm{SL}_2(\mathbb{Z})$ identifies all points of $\mathbb{Q} \cup \{\infty\}$; for a subgroup, however, we can wind up with rational numbers r such that $\gamma(r) \neq \infty$ for all $\gamma \in \Gamma$. So condition (ii) transfers such a rational r to infinity — using the operator $[\alpha_r]$ — so that the holomorphicity of $f[\alpha_r]$ at infinity yields the holomorphicity of f at r. Later examples should clarify these ideas, especially geometrically.

Recall our function $f(\tau) := \sum r(n)e^{2\pi i n\tau}$ and its associated subgroup Γ_f. With notation as before, f is a modular form of weight 2 with respect Γ_f. In fact, the subgroup Γ_f will have a special name once we introduce some more notation.

Definition 2.8. The **principal congruence subgroup of level** N is

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

where the equivalence mod N is taken entry-wise. A subgroup $\Gamma < \mathrm{SL}_2(\mathbb{Z})$ is a **congruence subgroup of level** N if $\Gamma(N) \subset \Gamma$. The most important congruence subgroups of level N are

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

and

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

where $*$ represents an unrestricted congruence mod N.

Note that for a fixed level N, we have the containments $\Gamma(N) < \Gamma_1(N) < \Gamma_0(N)$, so $\Gamma_1(N)$ and $\Gamma_0(N)$ are indeed congruence subgroups of level N. Note also that we have already seen an example of a congruence subgroup, as $\Gamma_f = \Gamma_0(4)$ – refer to exercise 1.2.4 in [DS05]. An important property of all congruence subgroups is that they have finite index.

Lemma 2.9. *A congruence subgroup Γ of level N has finite index in* $\mathrm{SL}_2(\mathbb{Z})$.

Proof. Consider the natural map $\phi : \mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$ given by reduction mod N. Then $\ker\phi = \Gamma(N)$ so $\mathrm{SL}_2(\mathbb{Z})/\Gamma(N) \cong \mathrm{Im}\,\phi < \mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$, a finite group. So $\Gamma(N)$ has finite index in $\mathrm{SL}_2(\mathbb{Z})$; because $\Gamma(N) < \Gamma$, the lemma follows. □

Now, we introduce some final details about the action of the modular group, but specialised to a congruence subgroup Γ. First note that we can define a fundamental domain for Γ as we did for the entire modular group: a fundamental domain for Γ is a set $D \subset \mathbb{H}$ such that $\Gamma(D) = \mathbb{H}$ and such that no two elements of D (up to boundary identifications) are equivalent under the action of Γ. Because $\Gamma \subset \mathrm{SL}_2(\mathbb{Z})$, a fundamental domain for Γ always contains a fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$. For example, figure 2 depicts the earlier fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$ together with eleven other (hyperbolic) triangles; the union of all twelve triangles constitutes a fundamental domain for $\Gamma(3)$ — where twelves comes from the fact that $[\mathrm{SL}_2(\mathbb{Z})/\{\pm I\} : \Gamma(3)] = 12$.

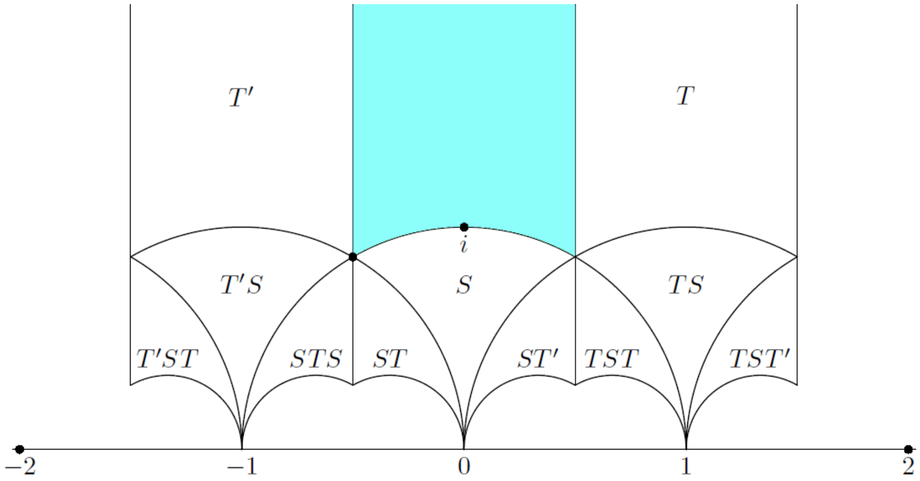


Figure 2: A fundamental domain for $\Gamma(3)$, with a fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$ in blue. The (hyperbolic) triangles are labelled with the transformation which takes the blue region to that triangle, where T denotes translation to the right, T' denotes translation to the left, and S denotes flips over the unit circle. Image adapted from [Sch18].

Definition 2.10. Let Γ be a congruence subgroup and write $\mathbb{Q}^* = \bigsqcup_i \Gamma(s_i)$ for distinct $s_i \in \mathbb{Q}^*$. We call the s_i the **cusps** of Γ, where it is understood that any other representative of the same coset works equally well.

The geometry of a fundamental domain motivates the term cusp, because they appear at the "limits", or "cusps", of a fundamental domain. For example, $\mathrm{SL}_2(\mathbb{Z})$ has just the cusp at ∞, while the cusps of $\Gamma(3)$ are -1, 0, 1, and ∞, as seen in figure 2. Both of these examples have finitely-many cusps; this is a general phenomenon.

Theorem 2.11. *A congruence subgroup Γ has finitely-many cusps.*

Proof. By proposition 2.9, there exist n and $\beta_i \in \mathrm{SL}_2(\mathbb{Z})$ such that $\mathrm{SL}_2(\mathbb{Z}) = \bigsqcup_{i=1}^{n} \Gamma\beta_i$. Thus,

$$\mathbb{Q}^* = \mathrm{SL}_2(\mathbb{Z})(\infty) = \bigcup_{i=1}^{n} \Gamma\beta_i(\infty)$$

so the cusps of Γ are some subset of $\beta_1(\infty), \ldots, \beta_n(\infty)$. □

As remarked earlier, the cusps are of interest because they appear at the limits of a fundamental domain; by including the cusps, therefore, we make the fundamental domain compact. Because of their incredible importance, these compactified domains get a special name.

Definition 2.12. Set $\mathbb{H}^* := \mathbb{H} \cup \mathbb{Q} \cup \{\infty\}$, the extended hyperbolic plane. Denote by $X(N)$ the quotient space $\Gamma(N)\backslash\mathbb{H}^*$ i.e. a fundamental domain for $\Gamma(N)$ with cusps added and boundaries identified (note that we place $\Gamma(N)$ to the left of the backslash because $\Gamma(N)$ acts on the left). Similary, define $X_1(N) := \Gamma_1(N)\backslash\mathbb{H}^*$ and $X_0(N) := \Gamma_0(N)\backslash\mathbb{H}^*$.

For example, $X(1)$ is the blue region in figure 1 with the left and right halves of the boundary identified and a point added at infinity; this makes $X(1)$ into a sphere, the simplest example of a compact complex manifold of dimension one.

Indeed, any of $X(N)$, $X_1(N)$, and $X_0(N)$ permits the structure of a compact complex manifold of dimension one, which in turn makes modular forms — and their cousins automorphic forms — into holomorphic — respectively, meromorphic — functions on the manifold. But we won't need these details for our discussion. An understanding of the example of $X(1)$, the action of congruence subgroups, and the geometry the action yields suffice.

Before we proceed, we need to address a small omission in the discussion thus far. Although we defined only the action of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathbb{H}$, we will want the action of a much larger group. Denote by $\mathrm{GL}_2^+(\mathbb{R})$ the set of two-by-two real matrices with positive detreminant. Then $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{GL}_2^+(\mathbb{R})$ and $\tau \in \mathbb{H}$, the action on $\mathbb{H}$ extends with precisely the same definition as before: $\gamma(\tau) = \frac{a\tau+b}{c\tau+d}$. However, the weight-$k$ operator requires an additional term.

Definition 2.13. Let f be a weight-k modular form and $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{GL}_2^+(\mathbb{R})$. Define the **weight-$k$ operator** $[\gamma]_k$ by $(f[\gamma]_k)(\tau) = (\det\gamma)^{k-1}(c\tau + d)^{-k} f(\gamma(\tau))$.

Notice that this definition agrees with the one given for $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ where $(\det\gamma) = 1$. We require the more general definition because it will appear in the definitions of the operators we next define.

3 Hecke Operators and Newforms

As is standard throughout mathematics, we now want to leverage the power of linear algebra. Specifically, we will define operators on the space $M_k(\Gamma)$ of weight-k modular forms with respect to Γ, before then investigating the eigenvectors of these operators. Ultimately, we will go on to associate certain representations to a special subset of these eigenvectors.

3.1 The Double Coset Operator

We first define a more general operator, before then isolating special cases to define our operators of interest.

Definition 3.1. Let Γ_1, Γ_2 be congruence subgroups and let $\alpha \in \mathrm{GL}_2(\mathbb{Z})$. Write $\Gamma_1\alpha\Gamma_2 = \cup_j \Gamma_1\beta_j$ with $\beta_j = \gamma_{1,j}\alpha\gamma_{2,j}$ with $\gamma_{i,j} \in \Gamma_i$.Then we define the **weight-k double coset operator** $[\Gamma_1\alpha\Gamma_2]_k : M_k(\Gamma_1) \to M_k(\Gamma_2)$ by

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

Although it is not at all clear from the definition, the sum defining the double coset operator is finite and the image indeed lands in $M_k(\Gamma_2)$ (see section 5.1 of [DS05] for details).

Because we made a choice of coset representatives β_j, we must also check if the operator is well-defined. To this end, take $f \in M_k(\Gamma_1)$ and let β and β' represent the same coset in $\Gamma_1\backslash\Gamma_1\alpha\Gamma_2$ — recalling once again that we write Γ_1 to the left of the backslash because the action occurs on the left. Then we need to check that $f[\beta]_k = f[\beta']_k$. Write $\beta = \gamma_1\alpha\gamma_2$ and $\beta' = \gamma_1'\alpha\gamma_2'$, with $\gamma_i \in \Gamma_i$, and write use that β and β' represent the same coset to obtain $\Gamma_1\beta = \Gamma_1\beta'$ because β and β'. Then

$$\Gamma_1\beta = \Gamma_1\beta' \implies \Gamma_1\alpha\gamma_2 = \Gamma_1\alpha\gamma_2' \implies \alpha\gamma_2 \in \Gamma_1\alpha\gamma_2'$$

and the Γ_1 invariance of f implies that $f[\alpha\gamma_2]_k = f[\alpha\gamma_2']_k$. Another application of Γ_1 invariance yields $f[\beta]_k = f[\beta']_k$ so the operator is indeed well-defined.

We illustrate this definition with three examples:

- Say $\Gamma_2 < \Gamma_1$ and $\alpha = I$. Then $\Gamma_1\alpha\Gamma_2 = \Gamma_1$ and $f[\Gamma_1\alpha\Gamma_2]_k = f[I]_k = f$, so we obtain the natural inclusion $M_k(\Gamma_1) \hookrightarrow M_k(\Gamma_2)$.
- Say $\Gamma_1 < \Gamma_2$ and $\alpha = I$. Then $\Gamma_1\alpha\Gamma_2 = \bigsqcup_{i=1}^n \Gamma_1\gamma_{2,i}$ for some coset representatives $\gamma_{2,i} \in \Gamma_2$. In this case, $f[\Gamma_1\alpha\Gamma_2]_k = \sum_i f[\gamma_{2,i}]_k$ gives a surjection $M_k(\Gamma_1) \twoheadrightarrow M_k(\Gamma_2)$: for any $f \in M_k(\Gamma_2) \subset M_k(\Gamma_1)$, we have $\frac{f}{n}[\Gamma_1\alpha\Gamma_2]_k = f$.
- Say $\alpha^{-1}\Gamma_1\alpha = \Gamma_2$. Then $\Gamma_1\alpha\Gamma_2 = \Gamma_1\alpha$ and $f[\Gamma_1\alpha\Gamma_2]_k = f[\alpha]_k$. So in this case we have an isomorphism $M_k(\Gamma_1) \xrightarrow{\sim} M_k(\Gamma_2)$ with inverse map $[\Gamma_2\alpha^{-1}\Gamma_1]_k = [\alpha^{-1}]_k$.

3.2 The diamond Operator and Dirichlet Characters

We now use the double coset operator to define the Hecke operators on $M_k(\Gamma_0(N))$. Consider the situation when $\Gamma_1 = \Gamma_1(N) = \Gamma_2$ and $\alpha \in \Gamma_0(N)$.Then $\Gamma_1\alpha\Gamma_2 = \Gamma_1\alpha$ and for $f \in M_k(\Gamma_1(N))$ we have $f[\Gamma_1\alpha\Gamma_2]_k = f[\alpha]_k$. In this way we obtain an action of $\Gamma_0(N)$ on $M_k(\Gamma_1(N))$, which takes f to $f[\alpha]_k$; because the normal subgroup $\Gamma_1(N)$ of $\Gamma_0(N)$ acts trivially in the preceding construction, we have actually constructed an action of $\Gamma_0(N)/\Gamma_1(N) \cong (\mathbb{Z}/N\mathbb{Z})^\times$ on $M_k(\Gamma_1(N))$. This actions yields the first of two Hecke operators.

Definition 3.2. For $d \in (\mathbb{Z}/N\mathbb{Z})^\times$, the **diamond operator** $\langle d\rangle : M_k(\Gamma_1(N)) \to M_k(\Gamma_1(N))$ is given by

$$\langle d\rangle f = f[\alpha]_k, \text{ any } \alpha = \begin{pmatrix} a & b \\ c & \delta \end{pmatrix} \in \Gamma_0(N) \text{ with } \delta \equiv d \mod N.$$

Note that such an α exists because c, which is $\equiv 0 \mod N$, and δ, which reduces to $d \in (\mathbb{Z}/n\mathbb{Z})^\times$ mod N, are relatively prime. Also note that this the well-defined-ness of this action — i.e lack of dependence on the choice of δ — follows from the triviality of the action of $\Gamma_1(N)$. Recalling that the weight-k operator is multiplicative — so $[\gamma_1\gamma_2]_k = [\gamma_1]_k[\gamma_2]_k$ — the next proposition follows from multiplying the matrices corresponding to two diamond operators.

Proposition 3.3. *Let $d, e \in (\mathbb{Z}/N\mathbb{Z})^\times$ have corresponding diamond operators $\langle d\rangle$ and $\langle e\rangle$ respectively. Then $\langle d\rangle\langle e\rangle = \langle de\rangle$. In particular, diamond operators commute.*

So far, we've only defined the diamond operator for d less than and coprime to the level N. The definition naturally generalises to any positive integer n.

Definition 3.4. Fix a level N and let n be a positive integer. If n and N are coprime, then the reduction $\bar{n}$ of n mod N is in $(\mathbb{Z}/N\mathbb{Z})^\times$ so we may define the **diamond operator** $\langle n\rangle$ of n by $\langle n\rangle := \langle \bar{n}\rangle$. If n and N are not coprime, define $\langle n\rangle := 0$.

Because our original diamond operators commute, our more general diamond opeartors also commute. Next, we introduce a convenient object for discussing eigenvalues of the diamond operator.

Definition 3.5. Let $G_N := (\mathbb{Z}/N\mathbb{Z})^\times$. A **Dirichlet character** $\chi \mod N$ is a homomorphism of multiplicative groups

$$\chi : G_N \to \mathbb{C}^\times.$$

Equivalently, we will think of a Dirichlet character mod N as a multiplicative map

$$\chi : \mathbb{Z} \to \mathbb{C}^\times$$

such that $\chi(k + N) = \chi(k)$ for all k and such that $\chi(n) = 0$ whenever $(n \mod N) \notin G_N$. Either way, we denote by $\widehat{G_N}$ the group (under multiplication) of all Dirichlet characters mod N.

Dirichlet characters satisfy a remarkable number of important properties, some of which we highlight here. Because G_N has finite order, its image in $\mathbb{C}^\times$ has finite order so $\chi(G_N) \subset S^1$, where S^1 denotes the circle group in $\mathbb{C}^\times$. In particular, the image of a Dirichlet character mod N consists entirely of N^{th} roots of unity (but will usually only require d^{th} roots of unity for divisors d of N). As with any object involving levels, Dirichlet characters of a smaller level naturally promote to Dirichlet characters of a higher level, but not all characters go the other way; we call a Dirichlet character primitive if it does not arise from a smaller level in the following precise sense.

Definition 3.6. For any d dividing N, let $\pi_{N,d} : G_N \to G_d$ denote the natural surjection, and let χ denote a Dirichlet character mod N. Let d denote the smallest divisor of N such that some character χ_d mod d satisfies $\chi = \chi_d \circ \pi_{N,d}$ (note that this is equivalent to the condition that χ is trivial on $\ker \pi_{N,d}$). We call d the **conductor** of χ and say that χ is **primitive** if the conductor of χ is N.

In other words, a Dirichlet character is primitive if it cannot be expressed as a character on a smaller level. For example, every nontrivial character mod a prime p is automatically primitive. The notion of the “level” of a Dirichlet character will naturally correspond to the levels of modular forms, as we will see momentarily. But first we use a standard trick involving sums over a group to obtain the so-called orthogonality relations.

Proposition 3.7. *Let $\chi : \mathbb{Z} \to \mathbb{C}^\times$ be a Dirichlet character mod N. Then*

$$\sum_{n \in G_N} \chi(n) = \begin{cases} \phi(N) & \text{if } \chi = 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\sum_{\chi \in \widehat{G_N}} \chi(n) = \begin{cases} \phi(N) & \text{if } n = 1 \\ 0 & \text{otherwise} \end{cases}$$

where ϕ is the Euler totient function.

Proof. Let $S := \sum_{n \in G_N} \chi(n)$. If $\chi = 1$, then $S = \#G_N = \phi(n)$ as desired. If $\chi \neq 1$, then there exists $m \in G_N$ such that $\chi(m) \neq 1$. Then

$$\sum_{n \in G_N} \chi(n) = \sum_{n \in G_N} \chi(mn)$$

because multiplication by m defines an automorphism of G_N, and so

$$S = \sum_{n \in G_N} \chi(mn) = \sum_{n \in G_N} \chi(m)\chi(n) = \chi(m)S.$$

But $\chi(m) \neq 1$, so $S = 0$ as desired. A similar argument with the same trick proves the second relation. □

Let's look at some examples. Let $N = 6$ and note that $G_6 = \{1, 5\}$. Then there is a single nontrivial Dirichlet character χ defined by $\chi(5) = -1$; the character χ has conductor three (so is not primitive) and necessarily satisfies $\chi(1) + \chi(-1) = 0$ as in the proposition. As another example, consider $N = 8$ for which $G_8 = \{1, 3, 5, 7\}$. Define a character χ by $\chi(3) = -1 = \chi(5)$ and $\chi(7) = 1$; then χ is primitive and once again the images sum to 0.

With Dirichlet characters in hand, we connect them with the diamond operator.

Definition 3.8. Let χ be a Dirichlet character mod N. For $\gamma \in \Gamma_0(N)$, let d_γ denote the lower-right entry of γ, so that $f[\gamma]_k = \langle d_\gamma \rangle f$ for all $f \in M_k(\Gamma_1(N))$. We define the χ-eigenspace of $M_k(\Gamma_1(N))$ to be

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

In particular, $M_k(N, \chi)$ collects modular forms which are eigenvectors for the diamond operators with eigenvalues given by χ.

What makes this definition work is that both Dirichlet characters and the diamond operator are multiplicative. More specifically, let $d_1, d_2 \in \mathbb{Z}/N\mathbb{Z}$ and say that a modular form f satisfies $\langle d_i \rangle f = \chi(d_i) f$ for some character $\chi : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}$. Then

$$\langle d_1 d_2 \rangle f = \langle d_1 \rangle \langle d_2 \rangle f = \langle d_1 \rangle \chi(d_2) f = \chi(d_1)\chi(d_2) f = \chi(d_1 d_2) f$$

so $\langle d_1 d_2 \rangle f = \chi(d_1 d_2) f$ as it must. It follows that if f is an eigenvector for each diamond operator d with eigenvalue λ_d, then the map $d \mapsto \lambda_d$ defines a character χ; in this case, we would have $f \in M_k(N, \chi)$.

Example 3.9. Let 1_N denote the trivial character mod N. Then $M_k(N, 1_N)$ consists of all modular forms such that $\langle d_\gamma \rangle f = f$ for all $\gamma \in \Gamma_0(N)$. Because $\langle d_\gamma \rangle f = f[\gamma]_k$, it follows that $M_k(N, 1_N) = M_k(\Gamma_0(N))$.

Example 3.10. Fix a weight k and level N, and take $f \in M_k(\Gamma_1(N))$. Then $-I \in \Gamma_0(N)$ and $\langle -1 \rangle f = f[-I]_k = (-1)^k f$. Thus, if χ is a character such that $\chi(-1) \neq (-1)^k$, then $M_k(N, \chi) = \{0\}$.

We consider χ-eigenspaces $M_k(N, \chi)$ because the full space $M_k(\Gamma_1(N))$ of modular forms decomposes into a direct sum of eigenspaces.

Theorem 3.11. *We have an equality $M_k(\Gamma_1(N)) = \bigoplus_\chi M_k(N, \chi)$ of $\mathbb{C}$ vector spaces, where the direct sum ranges over Dirichlet characters χ of level N.*

Proof. Recall the notation $G_N = (\mathbb{Z}/N\mathbb{Z})^\times$. As in previous arguments, we will "symmetrise" by summing over the whole group G_N: for each Dirichlet character χ, define a linear operator T_χ on $M_k(\Gamma_1(N))$ by

$$T_\chi := \frac{1}{|G_N|} \sum_{d \in G_N} \frac{\langle d \rangle}{\chi(d)}.$$

By construction, T_χ fixes $M_k(N,\chi)$. Moreover, for any $f \in M_k(\Gamma_1(N))$ and any $e \in G_N$,

$$\begin{aligned}\langle e\rangle(T_\chi f) &= \frac{1}{|G_N|}\sum_{d\in G_N}\frac{\langle de\rangle f}{\chi(d)} \\ &= \frac{1}{|G_N|}\sum_{d'\in G_N}\frac{\langle d'\rangle f}{\chi(d'e^{-1})} \\ &= \frac{1}{\chi(e^{-1})}\cdot\frac{1}{|G_N|}\sum_{d'\in G_N}\frac{\langle d'\rangle f}{\chi(d')} \\ &= \chi(e)f,\end{aligned}$$

so $T_\chi(f) \in M_k(N,\chi)$ and T_χ projects to $M_k(N,\chi)$. So far, we have shown that the operators T_χ are projections. To get a direct sum decomposition, it remains to show that

(i) for distinct Dirichlet characters χ and χ', the operator T_χ kills $M_k(N,\chi')$; and

(ii) the operator $\sum_{\chi\in\widehat{G_N}} T_\chi$ is the identity.

Assuming (i) and (ii), let's finish the proof. For any $f \in M_k(\Gamma_1(N))$, statement (ii) implies that

$$f = \sum_{\chi\in\widehat{G_N}} f_\chi,$$

where $f_\chi = T_\chi f \in M_k(N,\chi)$. For a direct sum, however, we need that the representation of f in (3.2) is unique. Represent f with a potentially different sum $f = \sum_{\chi\in\widehat{G_N}} f_\chi$, each $f_\chi \in M_k(N,\chi)$. Then statement (i) shows that for any $\chi' \in \widehat{G_N}$

$$T_{\chi'}f = \sum_{\chi\in\widehat{G_N}} T_{\chi'}(f_\chi) = T'_\chi(f_{\chi'}) = f_{\chi'}$$

so f_χ must equal $T_\chi f$ for all χ, so the sum in (3.2) is unique and we indeed have a direct sum. We now finish up the proofs of (i) and (ii).

To show (i), take $f \in M_k(N,\chi')$ and compute

$$|G_N|\cdot T_\chi f = \sum_{d\in G_N}\frac{\langle d\rangle f}{\chi(d)} = \sum_{d\in G_N}\frac{\chi'(d)f}{\chi(d)} = \sum_{d\in G_N}\psi(d)f$$

where $\psi := \chi'/\chi$ is a nontrivial Dirichlet character mod N (because $\chi \neq \chi'$). By the first orthogonality relation in theorem 3.7, the sum $\sum_{d\in G_N}\psi(d) = 0$ so $T_\chi f = 0$, as desired.

To show (ii), compute

$$\begin{aligned}|G_N|\cdot\sum_{\chi\in\widehat{G_N}} T_\chi &= \sum_{\chi\in\widehat{G_N}}\sum_{d\in G_N}\frac{\langle d\rangle}{\chi(d)} \\ &= \sum_{d\in G_N}\langle d\rangle\sum_{\chi\in\widehat{G_N}}\frac{1}{\chi(d)} \\ &= \sum_{d'\in G_N}\langle d'\rangle\sum_{\chi\in\widehat{G_N}}\chi(d').\end{aligned}$$

By the second orthogonality relation in theorem 3.7, the sum $\langle d' \rangle \sum_{\chi \in \widehat{G_N}} \chi(d')$ is nonzero only when $d = 1$, in which case it is $|G_N|$. Thus, we wind up with

$$|G_N| \cdot \sum_{\chi \in \widehat{G_N}} T_\chi = \langle 1 \rangle \cdot |G_N|$$

from which it follows that $\sum_{\chi \in \widehat{G_N}} T_\chi$ is the identity. This completes the proof. □

3.3 The T_p Operator

To define our second Hecke operator, we appeal once again to the to the case of the double coset operator $[\Gamma_1 \alpha \Gamma_2]$ when $\Gamma_1 = \Gamma_1(N) = \Gamma_2$. But now take $\alpha = \left(\begin{smallmatrix} 1 & 0 \\ 0 & p \end{smallmatrix}\right)$. Then the coset decomposition of $\Gamma_1 \alpha \Gamma_2$ depends on whether or not p divides N and after some effort (see [DS05] section 5.2) one obtains the following.

Definition 3.12. Set $\beta_j := \left(\begin{smallmatrix} 1 & j \\ 0 & p \end{smallmatrix}\right)$ and $\beta_\infty := \left(\begin{smallmatrix} p & 0 \\ 0 & 1 \end{smallmatrix}\right)$. The T_p **operator** on $M_k(\Gamma_1(N))$ is given by, for any $\left(\begin{smallmatrix} m & n \\ N & p \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$ and any $f \in M_k(\Gamma_1(N))$,

$$T_p f = \begin{cases} \sum_{j=0}^{p-1} f[\beta_j]_k, & \text{if } p \mid N \\ \sum_{j=0}^{p-1} f[\beta_j]_k + f[\left(\begin{smallmatrix} m & n \\ N & p \end{smallmatrix}\right)\beta_\infty]_k, & \text{if } p \nmid N. \end{cases}$$

We will make sense of the Hecke actions through their effect on Fourier expansions. In particular, because $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in \Gamma_1(N)$, a modular form $f \in M_k(\Gamma_1(N))$ satisfies $f(\tau) = f(\tau + 1)$. So f admits a Fourier expansion

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

where $q := e^{2\pi i \tau}$.

Theorem 3.13. *Take $f \in M_k(\Gamma_1(N))$ and let $a_n(f)$ denote the n^{th} Fourier coefficient of f. Let 1_N denote the function which is 1 when $p \nmid N$ and 0 when $p \mid N$ (the trivial character mod N). Then we have a formula for the n^{th} Fourier coefficient of $T_p f \in M_k(\Gamma_1(N))$:*

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

where, by convention, $a_{n/p}(f)$ is zero if n/p is not an integer.

If further $f \in M_k(N, \chi)$ for some Dirichlet character χ, then $T_p f \in M_k(N, \chi)$ and

$$a_n(T_p f) = a_{np}(f) + p^{k-1} \chi(p) a_{n/p}(f). \tag{4}$$

Proof. See the proof of proposition 5.2.2 in [DS05]. For μ_p a primitive p^{th} root of unity, the main ingredient in the proof is the following:

$$\sum_{j=0}^{p-1} \mu_p^{Nj} = \begin{cases} p, & \text{if } p \mid N, \\ 0, & \text{otherwise.} \end{cases}$$

To see this, recall that $G_p := (\mathbb{Z}/\mathbb{Z})^\times$ is cyclic for p prime and let a generate G_p. Then the map $G_p \to \mathbb{C}^\times$ defined by $a \mapsto \mu_p$ defines a Dirichlet character mod p. □

As with the diamond operator, we want to make sense of a version of the T_p operator for all positive n. In this case, we appeal to formula 4 for inspiration.

Definition 3.14. Fix a level N and consider a prime factorisation $n = \prod p_i^{r_i}$, with p_i distinct. For p prime and $r \geq 2$, we define

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

Then define the operator T_n by

$$T_n := \prod T_{p_i^{r_i}}.$$

For completeness, we also define T_1 to be the identity.

And with that we have defined all of our Hecke operators. For convenience of study, we collect them into a single object.

Definition 3.15. The **Hecke algebra** is the algebra of operators

$$\mathbb{T}_{\mathbb{Z}} := \mathbb{Z}[T_n, \langle n\rangle : n \in \mathbb{Z}_{>0}] \subset \operatorname{End}(M_k(\Gamma_1(N))),$$

where each integer operates on a modular form by multiplication.

With a bit of computational effort (see section 5.2 in [DS05]), one can show the following important fact.

Theorem 3.16. *The Hecke algebra $\mathbb{T}_{\mathbb{Z}}$ commutes; in other words, all Hecke operators commute with one another (and they naturally commute with multiplication by an integer).*

Although we won't go into the details here, that the Hecke algebra commutes will be an essential ingredient for theorem 3.22 in the next section. Indeed, we proceed to define newforms — the extremely special modular forms to which we will attach "Galois representations", and theorem 3.22 gives some important properties of newforms. A fair bit of theory goes into developing the definitions and propositions which follow, including the commutativity of the Hecke algebra, but we simply isolate the bare essentials, choosing instead to focus on the overarching story.

3.4 Newforms

Throughout this section, fix $\Gamma := \Gamma_1(N)$. Broadly speaking, a newform f of level N is a modular form in $M_k(\Gamma)$ with three essential properties:

(a) f vanishes at the cusps of Γ,

(b) f does not arise from a form at a lower level D dividing N, and

(c) f is an eigenvector (or eigenform) for every operator in the Hecke algebra.

Although only property (c) requires the theory of Hecke operators, we've postponed the discussion of (a) and (b) until now so that all three occur at the same time.

To make sense of condition (a), let n denote the number of cusps of $\Gamma_1(N)$ and choose $\beta_j \in \mathrm{SL}_2(\mathbb{Z})$, $1 \le j \le n$, such that $\beta_j(\infty)$ represent the cusps of Γ; that is, so that $\mathrm{SL}_2(\mathbb{Z}) = \bigsqcup_j \Gamma\beta_j$. Without loss of generality, assume that $\beta_1 = I$ so that β_1 corresponds to the cusp at ∞. Also recall that $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right) \in \Gamma = \Gamma_1(N)$ forces $f(\tau+1) = f(\tau)$ for all $\tau \in \mathbb{H}$, so f has a Fourier expansion

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

where, as always, $q := e^{2\pi i \tau}$. The value of f at infinity is $\lim_{\mathrm{Im}(\tau)\to\infty} = a_0(f)$ so we say that f vanishes at the cusp ∞ if $a_0(f) = 0$. Then, we say that f vanishes at a cusp $\beta_j(\infty)$ if $f[\beta_j{}^{-1}]_k$ vanishes at infinity; in other words, to check that a modular form vanishes at a cusp, we first transfer the cusp back to ∞ and then use the Fourier expansion there to determine vanishing (note the similarity of this definition to that of definition 2.7). Note that this entire process depends on the existence of a Fourier expansion; to see that $f[\beta_j^{-1}]$ has a Fourier expansion, refer to the discussion preceding definition 1.2.3 in [DS05].

We restrict attention to $\Gamma = \Gamma_1(N)$ as this will be our congruence subgroup of interest, even though a similar definition makes sense for much more general congruence subgroups.

Definition 3.17. A modular form $f \in M_k(\Gamma)$ is a **cusp form** if f vanishes at the cusps of Γ. We denote by $S_k(\Gamma)$ the space of cusp forms inside $M_k(\Gamma)$.

By linearity of the weight-k operator, $S_k(\Gamma)$ is itself a vector space. In our new language, we want our newforms to be cusp forms.

To make sense of condition (b), we first remark that there are two natural ways of moving between levels of modular forms.

Lemma 3.18. *Let M divide N. Then every $f \in M_k(\Gamma_1(M))$ may be regarded as a modular form in $M_k(\Gamma_1(N))$. Similarly, every $f \in S_k(\Gamma_1(M))$ may be regarded as a modular form in $S_k(\Gamma_1(N))$.*

Proof. Because $\Gamma_1(N) < \Gamma_1(M)$, that f is weight-k invariant with respect to matrices in $\Gamma_1(M)$ automatically implies the same for matrices in $\Gamma_1(N)$. The result for cusp forms is analogous. □

Lemma 3.19. *Let $f \in M_k(\Gamma_1(N))$ and define $g(\tau) := f(p\tau)$. Then $g \in M_k(\Gamma_1(pN))$. Similarly, if $f \in S_k(\Gamma_1(N))$, then $g \in S_k(\Gamma_1(pN))$.*

Proof. Take $\gamma \in \Gamma_1(pN)$ and write $\gamma = \left(\begin{smallmatrix} a & b \\ pNc & d \end{smallmatrix}\right)$ with $a, d \equiv 1 \mod pN$ and $ad - bpNC = 1$. Compute

$$\begin{aligned} g[\gamma]_k(\tau) &= (pNc\tau + d)^{-k} g\left(\frac{a\tau+b}{pNc\tau+d}\right) \\ &= (Nc(p\tau)+d)^{-k} f\left(\frac{a(p\tau)+bp}{Nc(p\tau)+d}\right) \\ &= f\left[\left(\begin{smallmatrix} a & bp \\ Nc & d \end{smallmatrix}\right)\right]_k (p\tau) \\ &= f(p\tau) \end{aligned}$$

where the final equality follows from the fact that $\left(\begin{smallmatrix} a & bp \\ Nc & d \end{smallmatrix}\right) \in \Gamma_1(N)$. So we have $g[\gamma]_k(\tau) = g(\tau)$ for any $\gamma \in \Gamma_1(pN)$ and $g \in M_k(\Gamma_1(pN))$ as desired. The result for cusp forms is analogous. $\square$

We are going to want our modular form in $M_k(\Gamma_1(N))$ to not arise from a lower level in either of the above senses. Roughly speaking, we want to study modular forms at the very first level for which they appear — the first level at which they are "new" modular forms, rather than "old" modular forms which come from previous levels. In spirit, this relates to the notion of a primitive Dirichlet character, where a primitive Dirichlet character is in some sense new to its level (although we will not put restrictions on whether or not the characters associated to our modular forms are primitive).

To do all of this formally, for each d dividing the level N we set $\alpha_d := \left(\begin{smallmatrix} d & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and consider the multiplication-by-d map $[\alpha_d]_k$. In particular, for $f \in S_k(\Gamma_1(\frac{N}{d}))$, lemma 3.19 shows that $f[\alpha_d]_k = (\det \alpha_d)^{k-1} f(d\tau) = d^{k-1} f(d\tau)$ is in $S_k(\Gamma_1(N))$.

Definition 3.20. For each divisor d of N, let $\alpha_d := \left(\begin{smallmatrix} d & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and define a map

$$i_d : S_k(\Gamma_1(N/d)) \times S_k(\Gamma_1(N/d)) \to S_k(\Gamma_1(N))$$

by

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

Define the subspace of **old modular forms at level N** to be

$$S_k(\Gamma_1(N))^{\text{old}} := \sum_{\substack{d|N \\ d\neq 1}} \operatorname{Im} i_d.$$

And with this define the subspace of **new modular forms at level N** by

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

where the orthogonal complement is taken with respect to the so-called Petersson inner product on $S_k(\Gamma_1(N))$.

We will not define the inner product here, but know that it exists only for the space of cusp forms $S_k(\Gamma_1(N))$ and not for the full space of modular forms. This, in part, motivates the need for condition (a). In our new language, condition (b) becomes that we want our newforms to be new at their level N.

Finally, condition (c) requires that our newform is an eigenvector (from here on we will use "eigenform") for every operator in the Hecke algebra. It turns out that (see theorem 5.8.2 in [DS05]) that $f \in S_k(\Gamma_1(N))^{new}$ is an eigenform for every $T \in \mathbb{T}_{\mathbb{Z}}$ if and only if f is an eigenform for each $T \in \{\langle n \rangle, T_n : \gcd(n, N) = 1\}$. Either way, we can now make a precise definition of newform.

Definition 3.21. A modular form $f \in M_k(\Gamma_1(N))$ is a **newform** if

(a) f is a cusp form i.e. $f \in S_k(\Gamma_1(N))$,

(b) f is new at level N i.e. $f \in S_k(\Gamma_1(N))^{\text{new}}$, and

(c) f is a (normalised) Hecke eigenform i.e. f is an eigenform for all operators $T \in \mathbb{T}_{\mathbb{Z}}$.

By normalised we mean that $a_1(f) = 1$.

We require the condition on normalisation so that any one-dimensional subspace of $S_k(\Gamma_1(N))^{\text{new}}$ has at most one newform. Indeed, newforms enjoy a number of important properties, including that they will constitute a basis for $S_k(\Gamma_1(N))^{\text{new}}$.

Theorem 3.22. *The set of newforms constitutes an orthogonal basis of $S_k(\Gamma_1(N))^{\text{new}}$ (with respect to the Petersson inner product). Each newform f lies in an eigenspace $S_k(N, \chi)$ and satisfies $T_n f = a_n(f)f$ for all positive n. That is, the T_n-eigenvalues of f align with its Fourier coefficients.*

Recall that by convention we take T_1 to be the identity map; this makes sense because we require a newform to satisfy $a_1(f) = 1 =$ eigenvalue of T_1. Moreover, that the the definition of T_n in 3.14 reflects the Fourier coefficient formula in 3.13 in part permits (and indeed contributes to the proof of) theorem 3.22. Hopefully, this provides at least some sense of how the theory allows such remarkable functions to exist.

We require one more property of newforms which relates the Fourier coefficients to the number fields introduced in section 1.

Theorem 3.23. *For f a newform, set $\mathbb{K}_f := \mathbb{Q}[a_n(f) : n \in \mathbb{Z}_{>0}]$ so $\mathbb{K}_f$ is the smallest field extension over $\mathbb{Q}$ generated by the Fourier coefficients of f. Then the extension $\mathbb{K}_f/\mathbb{Q}$ has finite degree; in particular, $\mathbb{K}_f$ is a number field.*

Definition 3.24. For f a newform, the number field $\mathbb{K}_f$ introduced in the previous theorem is the **number field (or coefficient field) of** f.

Whenever we have a number field, we're naturally interested in its ring of integers. In this case, we have that the Fourier coefficients not only generate a number field, but also lie within that number field's ring of integers.

Theorem 3.25. *For f a newform, the Fourier coefficients $a_n(f)$ are algebraic integers; in other words, $\mathbb{Z}[a_n(f) : n \in \mathbb{Z}_{>0}]$ is a subset of the ring of integers of $\mathbb{K}_f$.*

Note that we only have $\mathbb{Z}[a_n(f)] \subseteq \mathcal{O}_{\mathbb{K}_f}$, not strict equality. Some examples will illustrate these theorems as well as demonstrate instances when our containment is (and is not) an equality.

Example 3.26. In the London Modular Forms Database (LMFDB), newform 37.2.b.a has Fourier expansion

$$f(q) = q + 2iq^2 - q^3 - 2q^4 - 2iq^5 - 2iq^6 + 3q^7 - 2q^9 + \cdots,$$

where as always $q = e^{2\pi i \tau}$. The newform f has level 37 and weight 2, so $f \in S_2(37, \chi)$ for some Dirichlet character χ; let's calculate χ.

Recall from definiton 3.14 that $T_{p^2} = T_p T_p - p^{k-1}\langle p \rangle$. By applying the eigenvalue map — and remembering that f is a newform implies that its Fourier coefficients are precisely its T_n eigenvalues — we have

$$a_{p^2}(f) = a_p(f)^2 - p\langle p \rangle$$

(one can also deduce this equation by taking $n = 2 = p$ in theorem 3.13). Taking $p = 2$, we have

$$-2 = (2i)^2 - 2 \cdot \chi(2)$$

from which it follows that $\chi(2) = -1$. Because 2 generates $(\mathbb{Z}/37\mathbb{Z})^\times$, the equality $\chi(2) = -1$ entirely defines χ. Indeed, this agrees with the Dirichlet character reported on LMFDB.

Note also that the Fourier coefficients of f generate the number field $\mathbb{K}_f = \mathbb{Q}[i]$ which has ring of integers $\mathbb{Z}[i]$. But because $\mathbb{Z}[a_n(f)] = \mathbb{Z}[2i]$ — this appears likely from the coefficients we have written above and indeed LMFDB confirms it — we have a proper containment in $\mathbb{Z}[a_n(f)] \subseteq \mathcal{O}_{\mathbb{K}_f}$. The index of $\mathbb{Z}[2i]$ in $\mathcal{O}_{\mathbb{K}_f}$ is 2, sometimes known as the coefficient ring index or Hecke index of f. Lastly, the q^8 coefficient in the Fourier expansion is zero, so $T_8 \in I_f$ (where $I_f \subset \mathbb{T}_\mathbb{Z}$ denotes the kernel of the eigenvalue map); in fact, looking at additional terms on LMFDB reveals that $T_{24}, T_{31}, \ldots \in I_f$ as well.

Example 3.27. In the London Modular Forms Database (LMFDB), newform 16.2.e.a has Fourier expansion

$$f(q) = q + (-1-i)q^2 + (-1+i)q^3 + 2iq^4 + (-1-i)q^5 + 2q^6 - 2iq^7 + (2-2i)q^8 + iq^9 + \cdots .$$

The newform f also has associated Dirichlet character $\chi : (\mathbb{Z}/16\mathbb{Z})^\times \to \mathbb{C}^\times$ defined by $\chi(5) = i$ and weight 2, so $f \in S_2(16, \chi)$.

As before, let's substitute $p = 3$ into

$$a_{p^2}(f) = a_p(f)^2 - p\langle p \rangle$$

to obtain

$$i = (-1+i)^2 - 3\chi(3).$$

It follows that $\chi(3) = -i$; because $3 \equiv (-1) \cdot 5^3 \mod 16$, our calculation of $\chi(3)$ agrees with the definition of χ as it must.

Once again the Fourier coefficients generate the number field $\mathbb{K}_f = \mathbb{Q}[i]$ which has ring of integers $\mathbb{Z}[i] = \mathbb{Z}[a_n(f)]$. But in this case we have an equality in $\mathbb{Z}[a_n(f)] \subseteq \mathcal{O}_{\mathbb{K}_f}$. Finally, we see that $T_n \notin I_f$ for any $n \in \{1, \ldots, 9\}$, but some arithmetic yields $2 \cdot T_3 + T_8$ and $2 \cdot T_2 + T_4 + T_6$ are in I_f.

In section 5, we will attach a "Galois representation" to each newform f as well as attach a Galois representation to each "elliptic curve" over $\mathbb{Q}$. Then the Shimura-Taniyama Conjecture will provide a precise correspondence between newforms and elliptic curves through their Galois representations. To accomplish all of this, we first require the notion of an elliptic curve.

4 Elliptic Curves

Mathematicians studied elliptic curves well before the twentieth century. But the Shimura-Taniyama Conjecture — and its contribution to proving Fermat's Last Theorem — vaulted elliptic curves into the forefront of mathematical interest. Although their precise definition appears most naturally in the abstraction of algebraic geometry, we simplify our study as much as possible to avoid this technicality. Throughout this section, we let $\mathbf{k}$ denote a field of any characteristic with algebraic closure $\overline{\mathbf{k}}$. Eventually, we will of course specialise to the case of $\mathbf{k} = \mathbb{Q}$.

4.1 Projective Space

To give a precise definition of an elliptic curve, we require the notion of projective space.

Definition 4.1. Define n-dimensional **projective space** by

$$\mathbb{P}^n(\mathbf{k}) := \{(x_1 : \cdots : x_{n+1}) : x_i \in \mathbf{k},\ \text{some } x_i \neq 0\}/\sim$$

where $(x_1 : \cdots : x_{n+1}) \sim (y_1 : \cdots : y_{n+1})$ if there exists $\lambda \in \mathbf{k}$ such that $x_i = \lambda y_i$ for all i.

In words, projective space is the collection of nonzero $(n+1)$-tuples modulo the action of multiplication by field elements.

As an example, we show that $\mathbb{P}^1(\mathbb{R})$ is isomorphic to one copy of the real line together with an extra point out at infinity. Indeed, any point $(a : b) \in \mathbb{P}^1(\mathbb{R})$ with $a \neq 0$ is equivalent to $(1 : c)$, where $c := a^{-1}b$. Similarly, any point $(0 : a)$, $a \neq 0$, is equivalent to $(0 : 1)$. Altogether, we have

$$\mathbb{P}^1(\mathbb{R}) := \{(1 : a) : a \in \mathbb{R}\} \cup (0 : 1) \cong \mathbb{R} \cup \{\infty\}$$

where the point $(0, 1)$ plays the role of an extra point out an infinity. This is in fact one of the great advantages of projective space: it allows us to account for points out at infinity. As another example, consider $\mathbb{P}^2(\mathbb{R})$. In the same way as before, we have

$$\mathbb{P}^2(\mathbb{R}) \cong \mathbb{R}^2 \cup \mathbb{P}^1(\mathbb{R}) \cong \mathbb{R}^2 \cup \mathbb{R} \cup \{\infty\}$$

where the extra $\mathbb{R} \cup \{\infty\}$ accounts for the extra points at infinity: each point in $m \in \mathbb{R}$ corresponds to the infinity at the end of a line of slope m, and the point $\{\infty\}$ corresponds to the infinity at the end of a vertical line. Therefore, even parallel lines intersect in $\mathbb{P}^2(\mathbb{R})$: they intersect at their slope's point at infinity! When we turn toward elliptic curves, we will play with this idea of points at infinity. But we first generalise the notions we've already introduced.

Definition 4.2. Let $\mathbf{k}$ be a field. Then n-dimensional **affine space** over $\mathbf{k}$ is

$$\mathbb{A}^n(\mathbf{k}) = \{(x_1, \ldots, x_n) : x_i \in \mathbf{k}\}$$

Theorem 4.3. *Let $\mathbf{k}$ be a field. Then*

$$\mathbb{P}(\mathbf{k})^n = \mathbb{A}^n(\mathbf{k}) \cup \mathbb{P}^{n-1}(\mathbf{k}).$$

So n-dimensional projective space consists of two components: an n-dimensional affine piece together with the points $\mathbb{P}^{n-1}(\mathbf{k})$ out at infinity.

Proof. The proof strategy proceeds identically as in the example of $\mathbb{P}^1(\mathbb{R})$ given above; for the sake of brevity, we omit the details. □

4.2 Elliptic Curves

With the notion of projective space in hand, we turn to elliptic curves, for which we'll require the notion of a "curve" in $\mathbb{P}^2(\mathbf{k})$.

Definition 4.4. A **projective curve in** $\mathbb{P}^2(\mathbf{k})$ is the set of zeroes of a homogeneous — every monomial has the same degree — polynomial in three variables of degree d.

Remark 4.5. We only define projective curves in $\mathbb{P}^2(\mathbf{k})$ because that's all we need for elliptic curves; fortunately, these agree with the natural notion of curve that we're accustomed to seeing in, say, calculus. For future reference, note that curves in $\mathbb{P}^n(\mathbf{k})$, $n \geq 3$, require intersections of zero sets of polynomials; this is analogous to the fact that one can realize a line in $\mathbb{R}^3$ as an intersection of two planes.

We require the condition on homogeneity because it guarantees that equivalent points map to the same thing. More precisely, a homogeneous polynomial P in three variables of degree d gives a well-defined map $P : \mathbb{P}^2(\mathbf{k}) \to \mathbb{P}$ because $(x_1, x_2, x_3) \sim (y_1, y_2, y_3)$ if and only if $y_i = \lambda x_i$ if and only if

$$P(y_1, y_2, y_3) = P(\lambda x_1, \lambda x_2, \lambda x_3) = \lambda^d P(x_1, x_2, x_3) \sim P(x_1, x_2, x_3).$$

Formally, an elliptic curve E (over a field $\mathbf{k}$) is a "smooth projective algebraic curve (over $\mathbf{k}$) of genus one". Rather than work with this abstract definition, which draws upon significant algebraic geometry, we appeal to an incredible and beautiful theorem regarding the structure of elliptic curves.

Theorem 4.6. *Let E be an elliptic curve over a field $\mathbf{k}$. Then E is isomorphic to the projective curve in $\mathbb{P}^2(\mathbf{k})$ given by some equation*

$$Y^2Z + a_1XYZ + a_3YZ^2 = X^3 + a_2X^2Z + a_4XZ^2 + a_6Z^3. \tag{5}$$

Definition 4.7. The **discriminant** of the Weierstrass equation (5) is

$$\Delta := -b_2^2b_8 - 8b_4^3 - 27b_6^2 + 9b_2b_4b_6$$

where

$$\begin{aligned} b_2 &= a_1^2 + 4a_2, \\ b_4 &= 2a_4 + a_1a_3, \\ b_6 &= a_3^2 + 4a_6, \text{ and} \\ b_8 &= a_1^2a_6 + 4a_2a_6 - a_1a_3a_4 + a_2a_3^2 - a_4^2. \end{aligned}$$

The j**-invariant** of an elliptic curve is

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

where

$$c_4 = b_2^2 - 24b_4.$$

Theorem 4.8. *The Weierstrass equation (5) defines an elliptic curve if and only if its discriminant does not equal zero. Two Weierstrass equations define the same (isomorphic) elliptic curves over* $\overline{\mathbf{k}}$ *if and only if they have the same* j*-invariant.*

As such, from here on out we will always regard elliptic curves as the points satisfying some equation (5) with nonzero discriminant, and we will refer to this equation as the (homogeneous) Weierstrass form of the elliptic curve E. Note that E lives inside $\mathbb{P}^2 = \mathbb{A}^2 \cup \mathbb{P}^1$. As such, we should be able to split E into two pieces: an affine piece and a piece out at infinity. After making sense of these pieces, we will look at some examples.

We can obtain infinitely-many different copies of $\mathbb{A}^2$ inside $\mathbb{P}^2$ by choosing one of X, Y, or Z and setting that variable to something nonzero (Note that this isn't the standard way to obtain the de-homogenised Weierstrass equation, as it suggests that we have to make a choice. There is a natural canonical alternative that ends in the same place, but we present the material in this way because we find it more intuitive.) To obtain the simplest possible affine embedding of equation (5), we choose to set Z equal to 1, which leaves the curve

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6 \tag{6}$$

in $\mathbb{A}^2$ (where, by convention, we use lower-case letters to denote variables in affine space). For the component at infinity, theorem 4.3 again shows that we need only set $Z = 0$; this forces $X = 0$, so there is only the point $(0 : 1 : 0)$ out at infinity. Note that from equation (6) we can completely recover (5) by simply multiplying each term by the appropriate power of Z: this process, called "re-homogenisation", shows that we may easily transfer between the projective and affine interpretations. We will most often think of an elliptic curve E as defined by equation (6) in $\mathbf{k}^2$ and having a designated point at infinity $(0 : 1 : 0)$. Figure 3 gives some examples of elliptic curves.

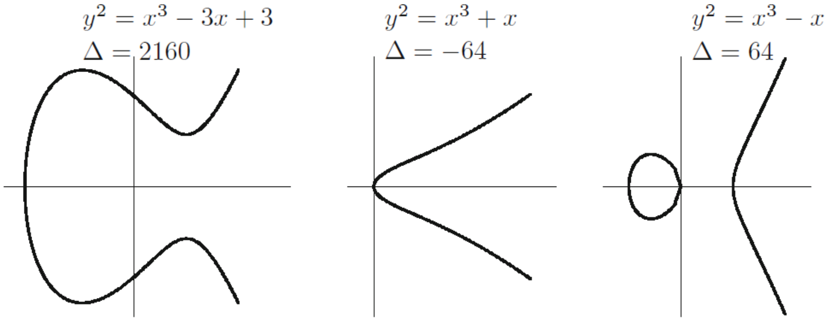


Figure 3: The affine pieces of three elliptic curves over $\mathbb{R}$. Image taken from [Sil86].

We will often use 0 to denote the point $(0 : 1 : 0)$ at infinity because it will act as the identity element for an abelian group structure on the elliptic curve!

Definition 4.9. Let an elliptic curve E over $\mathbf{k}$ have Weierstrass equation (6) and let P, Q be two points on E with coordinates in $\mathbf{k}$. Define a binary operation on E as follows: if the line through P and Q intersects E at a third point (x, y), set $P + Q := (x, -y)$;

otherwise, $P+Q := 0$. Equivalently, we could define the group law by requiring that three collinear points P, Q, R on E satisfy $P+Q+R=0$. Either way, we count intersections with multiplicity so that $P+P$ equals the intersection between the tangent line at P and E. We denote by $E(\mathbf{k})$ the group which results. Figure 4 illustrates the group law.

The two definitions given are certainly equivalent when $P+Q=0$. Otherwise, let $R=(x,y) \in \mathbf{k}^2$ denote $P+Q$. We may assume that our Weierstrass equation is such that (x,y) is on E if and only if $(x,-y)$ is (see section 3.1 for details justifying this claim in [Sil86]). Then consider the vertical line through (x,y) and $(x,-y)$: because it passes through the point at infinity, $(x,y)+(x,-y)=0$ from which it follows that $-R=(x,-y)$. But then $P+Q+R=0$ implies $P+Q=-R=(x,-y)$ as in the second definition. Checking that either definition indeed defines a group law — in particular, checking associativity — is a long, careful exercise but is nevertheless doable.

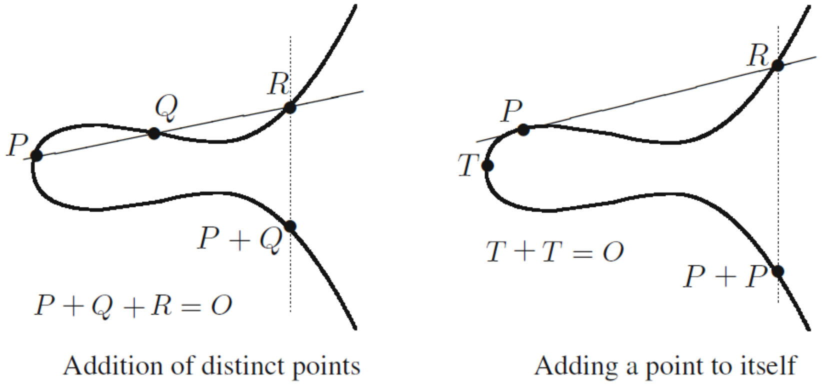


Figure 4: The group law of an elliptic curve. Image taken from [Sil86].

Our principal interest will be in understanding elliptic curves over finite algebraic extensions K of $\mathbb{Q}$. Note that, by the group law definition, if P and Q are are points of E whose coordinates lie in K, then $P+Q$ also has coordinates in K. This holds true for any field, so we can indeed make sense of the group of an elliptic curve over any field (as long as the discriminant is nonzero). Let's compute an example to illustrate these ideas.

Example 4.10. Let E be the elliptic curve given by the equation $y^2 = x^3 + x$ (this is the second elliptic curve in figure 3). Rather than regard E as a curve over $\mathbb{R}$, however, let's regard E as a curve over $\mathbb{F}_3$; this is valid because the Weierstrass equation has discriminant -64 which is nonzero mod 3 (so, in particular, theorem 4.8 says that $y^2 = x^3 + x$ indeed defines an elliptic curve mod 3). Over $\mathbb{F}_3$, there are only four points on the elliptic curve: 0, (0,0), (2,1), and (2,2). So the elliptic curve group $E(\mathbb{F}_3)$ is isomorphic to either $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ or $\mathbb{Z}/4\mathbb{Z}$. To determine which one it is, we'll compute $P+P$ for $P=(2,1)$.

The line which passes through P twice is the tangent line at P; computing that $\frac{dy}{dx} = \frac{1}{2y}$ over $\mathbb{F}_3$ shows that the tangent line at $(2,1)$ has equation $y = 2x$. The tangent line intersects

E at a point (x_0, y_0) satisfying

$$\begin{aligned} y_0 &= 2x_0 \\ (2x_0)^2 &= x_0^3 + x_0 \end{aligned}$$

from which it follows that $x_0 = 0$ or $x_0 = 2$. The point where $x_0 = 2$ is the point P, so we're interested in the point with $x_0 = 0$. This is the point $(0, 0)$. So $(2, 1) + (2, 1)$ is not the identity, from which it follows that $E(\mathbb{F}_3) = \mathbb{Z}/4\mathbb{Z}$.

4.3 Torsion and the Tate Module

As is often the case in group theory, to understand the elliptic curve group structure $E(\mathbf{k})$, we begin with nice subgroups. For us, in particular, we want to study the torsion subgroups.

Definition 4.11. Let E be an elliptic curve. For each positive integer m, define the **multiplication by** m map $[m] : E \to E$ by $[m]P = \underbrace{P + \cdots + P}_{m \text{ times}}$.

With this notation, the m torsion of the elliptic curve group is precisely the kernel of $[m]$. The multiplication by m map, though a bit complicated, may be understood entirely in terms of rational polynomials.

Theorem 4.12. *Let E be an elliptic curve with coefficients in some field $\mathbf{k}$ and let $P = (x, y)$ be a point of E (with x and y potentially in $\overline{\mathbf{k}}$). Then either $[m]P = 0$ or there exist polynomials ψ_i in two variables with coefficients in $\mathbf{k}$ such that*

$$[m]P = \left(\frac{\psi_1(x, y)}{\psi_2(x, y)}, \frac{\psi_3(x, y)}{\psi_4(x, y)} \right).$$

Proof. The proof proceeds by induction. The base case for $m = 1$ follows immediately. For the inductive step, write $[m]P = P + [m-1]P$ and notice that if $[m]P = 0$ or $[m-1]P = 0$, we're done. Otherwise, use the definition of the group law to see that the addition of two points is equivalent to evaluating two rational functions with coefficients in $\mathbf{k}$. Because rational functions (over $\mathbf{k}$) evaluated at rational functions (over $\mathbf{k}$) yield rational functions (over $\mathbf{k}$), the statement follows. □

In the statement and proof of the theorem, we've emphasised the importance of when coefficients/coordinates are in $\mathbf{k}$ or in $\overline{\mathbf{k}}$; in general, we'll think of coordinates are in $\overline{\mathbf{k}}$ unless otherwise specified, as in the notation $E(\mathbf{k})$ introduced earlier or in the following definition.

Definition 4.13. Let E be an elliptic curve over $\mathbf{k}$. We define the m**-torsion** of E to be

$$E[m] := \ker[m] = \{P = (x, y) \in \overline{\mathbf{k}}^2 : P \in E, [m]P = 0\} \cup \{0\}.$$

We refer to points $(x, y) \in E[m]$ with $x, y \in \mathbf{k}$ as *rational* m-torsion points.

The word choice "rational" comes from our motivating example as number theorists: elliptic curves over $\mathbb{Q}$. When we restrict to this case, rational torsion points will quite literally be torsion points with $\mathbb{Q}$-rational coordinates. Either way, we consider algebraic closures $\overline{\mathbf{k}}$ because the m-torsion structure is much simpler over an algebraically closed field.

Theorem 4.14. *Let E be an elliptic curve over a field $\mathbf{k}$ and m a positive integer. If* $\operatorname{char}(\mathbf{k})$ *does not divide m, then*

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

This includes when $\operatorname{char}(\mathbf{k}) = 0$*, as zero divides no positive integers.*

Proof. While we haven't introduced the tools to prove this theorem in full generality, we can prove it for $m = 2$.

Let $P = (x, y) \in \mathbf{k}^2$ be a 2-torsion point on E, and recall that $-P = (x, -y)$. Then P is a 2-torsion point if and only if $P + P = 0$ if and only if $P = -P$ if and only if $y = -y$ if and only if $y = 0$. The points $(x, 0)$ on E correspond to the roots of a third-degree Weierstrass polynomial; because E has coefficients in $\mathbf{k}$, there are three such distinct roots x_1, x_2, x_3 in $\overline{\mathbf{k}}^2$ (they are distinct for otherwise the discriminant would be zero and E would not define an elliptic curve). As such, we have four two-torsion points 0, $(x_1, 0)$, $(x_2, 0)$, and $(x_3, 0)$. A group of order four all of whose elements have order two must be $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, as desired. □

To see this theorem in action, we'll look at some examples.

Example 4.15. Let E be the elliptic curve $y^2 = x^3 - x$ over $\mathbb{R}$, which is graphed in figure 3. Because the graph of $y^2 = x^3 - x$ shows three roots, all of the 2-torsion points are real; in fact, they are the points with x-coordinates -1, 0, and 1. We obtain an isomorphism $E[2] \xrightarrow{\sim} \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ by mapping

$$\begin{aligned} (-1, 0) &\mapsto (1, 0) \\ (0, 0) &\mapsto (0, 1) \\ (0, -1) &\mapsto (1, 1) \end{aligned}$$

which makes sense because $(-1, 0) + (0, 0) = (0, -1)$ in E and $(1, 0) + (0, 1) = (1, 1)$ in $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$.

The other two examples in figure 3 each have a conjugate pair of complex 2-torsion points, so the geometry of the torsion does not show up in the graphs over $\mathbb{R}$. Nevertheless, one could create an analogous isomorphism of groups purely algebraically. Let's return to the elliptic curve in example 4.10 to see what this looks like.

Example 4.16. Let E be the elliptic curve over $\mathbb{F}_3$ given by the equation $y^2 = x^3 + x$. In example 4.10, we computed that $E(\mathbb{F}_3) = \mathbb{Z}/4\mathbb{Z}$, so there is only one rational, two-torsion point; namely, the point $(0, 0)$. To get the other two-torsion points, we have to move up to $\overline{\mathbb{F}_3}$. Indeed, the roots of $x^3 + x = x(x^2 + 1)$ are 0 and $\sqrt{-1}$. But -1 is not a square mod 3, so we need to take an element $i \in \overline{\mathbb{F}_3}$ such that $i^2 = -1$. Then our torsion points become 0, $(0, 0)$, $(i, 0)$, and $(-i, 0)$. Once again, we have a group of order four all of whose elements have order two, so the group must be $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$.

To study the m-torsion groups $E[m]$, we pack all of them into a single algebraic object associated to the elliptic curve: the Tate module. To define the Tate module, first notice that for any positive integer ℓ and each positive n we have a natural map

$$E[\ell^n] \leftarrow E[\ell^{n+1}]$$

given by multiplication by ℓ. This looks reminiscent of example 1.11 in which we constructed the ℓ-adic integers; indeed, the ℓ^n-torsion groups form an inverse system

$$E[\ell] \leftarrow E[\ell^2] \leftarrow E[\ell^3] \leftarrow \cdots \leftarrow E[\ell^n] \leftarrow E[\ell^{n+1}] \leftarrow \cdots$$

with maps given by multiplication by some power of ℓ. The inverse limit of this system is the object we want.

Definition 4.17. We define the **Tate module** of an elliptic curve E by $\mathrm{Ta}_\ell(E) := \varprojlim_n E[\ell^n]$.

Momentarily, we will use the Tate module to define representations in section 5. But while we're on the topic of elliptic curves, we define the essential notion of reduction.

4.4 Reduction of Elliptic Curves over $\mathbb{Q}$

Figure 3 depicts the affine piece of the elliptic curve $E : y^2 = x^3 - x$ over $\mathbb{R}$; by considering only the rational points, we may equally well regard E as an elliptic curve over $\mathbb{Q}$. In both cases, the discriminant is $-64 \neq 0$ so the Weierstrass equation indeed defines an elliptic curve over either field. In examples 4.10 and 4.16, we went even farther and regarded E over $\mathbb{F}_3$, where this again yields an elliptic curve because $-64 \not\equiv 0 \mod 3$. This process, of taking an elliptic curve E over $\mathbb{Q}$ and instead regarding it as a curve over a finite field $\mathbb{F}_p$, is called reduction. A question of well-defined-ness arises, however, because of theorem 4.8: distinct Weierstrass equations might define the same elliptic curve. And those distinct Weierstrass equations could yield different behaviours mod the same prime p. So we must standardise our choice.

Definition 4.18. Let E be an elliptic curve over $\mathbb{Q}$. A **minimal Weierstrass equation of** E is a Weierstrass equation with rational coefficients whose discriminant has the fewest number of prime divisors (counted with multiplicity).

Remark 4.19. This definition of minimality relies on the fact that $\mathcal{O}_{\mathbb{Q}} = \mathbb{Z}$ is a principal ideal domain (and in fact permits unique factorisation) so we can simultaneously minimize the number of all prime divisors. For a general number field, $\mathcal{O}_K$ may not be a P.I.D. in which case minimality is better regarded as a local condition for various primes. For $K = \mathbb{Q}$, this amounts to considering elliptic curves over $\mathbb{Q}_p$ for primes p, but is not necessary for our discussion (which is why we ignore the local perspective; see [Sil86] for elliptic curves over a general "local field").

By the Well-Ordering Principle, minimal Weierstrass equations necessarily exist. But they are not unique: the equations $y^2 = x^3 - x$ and $y^2 = (x+1)^3 - (x+1)$ define the same (up to isomorphism) elliptic curve E, both have discriminant $-64 = 2^6$, and no other Weierstrass equation representing E has five or fewer total prime divisors. For details on the existence and construction of minimal Weierstrass equations, see sections 3.1 and 7.1 in [Sil86]. From here on out, we represent all elliptic curves over $\mathbb{Q}$ with their minimal Weierstrass equations, even when not specifically stated.

Definition 4.20. Let E be an elliptic curve over $\mathbb{Q}$ and let $\tilde{E}$ denote the curve obtained by reducing its (minimal) Weierstrass equation mod p. Then two possibilities arise:

(i) if p does not divide the discriminant of E, then $\tilde{E}$ is an elliptic curve over $\mathbb{F}_p$; or

(ii) if p divides the discriminant of E, then $\tilde{E}$ has a singularity as a curve over $\mathbb{F}_p$.

In the first case, we say that E has **good reduction at** p. In the second case, we say that E has **bad reduction at** p.

Case (ii) — the case of bad reduction — in fact has two sub-cases, as a Weierstrass equation can have one of two distinct types of singularities: a cusp or a node. The terminology comes from the geometry of such singularities over $\mathbb{R}$, as depicted in figure 5. Just as the discriminant Δ detected when an elliptic curve has a singularity, another invariant will detect the type of singularity.

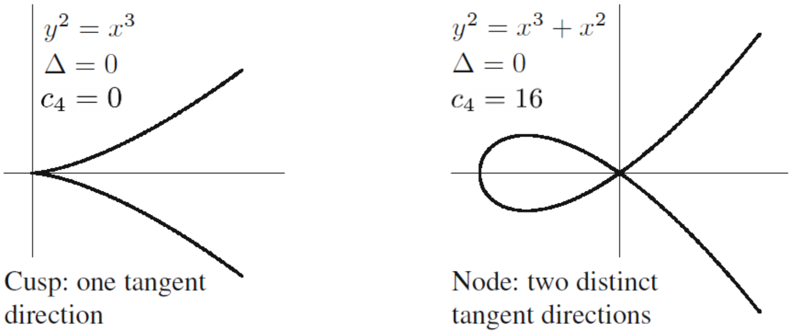


Figure 5: The two types of of non-smooth cubic curves. Image adapted from [Sil86].

Theorem 4.21. *Let C be a projective curve in $\mathbb{P}^2(\mathbf{k})$ given by a homogeneous Weierstrass equation. Then*

- *the curve C is nonsingular if and only if $\Delta \neq 0$, in which case C is an elliptic curve;*
- *the curve C has a node if and only if $\Delta = 0$ and $c_4 \neq 0$; and*
- *the curve C has a cusp if and only if $\Delta = 0 = c_4$.*

In addition, C has at most one singularity.

So in reducing a curve over $\mathbb{Q}$ mod some prime p, we could end up with three very different types of behaviour. The curve $y^2 = x^3$ in figure 5, for example, has a cusp over any finite field (as both its discriminant and c_4-value are always 0). In contrast, the curve $y^2 = x^3 + x^2$ has a cusp only over $\mathbb{F}_2$ and has a node over any other finite field. The following definition enumerates the types of reduction.

Definition 4.22. Let E be an elliptic curve over $\mathbb{Q}$ and let $\tilde{E}$ denote the reduction of E mod some prime p. We say that

- E has **good reduction at** p if $\tilde{E}$ defines an elliptic curve over $\mathbb{F}_p$;

- E has **semi-stable (multiplicative) bad reduction** at p if $\tilde{E}$ has a node over $\mathbb{F}_p$; and
- E has **unstable (additive) bad reduction at** p if $\tilde{E}$ has a cusp over $\mathbb{F}_p$.

The use of the term "stable" refers to the potentially-different behaviour of a curve $E/\mathbb{Q}$ when regarded as a curve some over finite extensions of $\mathbb{Q}$. And the use of the terms "additive" and "multiplicative" refer to the structure of the group of points $\tilde{E}(\mathbb{F}_p)$ in those cases. For details, see [Sil86] section 7.4 (noting that there they consider elliptic curves over $\mathbb{Q}_p \supset \mathbb{Q}$) and [DS05] section 8.3. Finally, we collect the local behaviour of reductions at various primes into a global invariant associated with an elliptic curve over $\mathbb{Q}$.

Definition 4.23. Let E be an elliptic curve over $\mathbb{Q}$ and set

$$e_p := \begin{cases} 0, & \text{if } E \text{ has good reduction at p} \\ 1, & \text{if } E \text{ has semi-stable reduction at p} \\ 2 + \delta_p, & \text{if } E \text{ has unstable reduction at p} \end{cases}$$

where δ_p is a "small" (≤ 6) correction term which is zero for all $p \notin \{2, 3\}$ (for a more detailed discussion of δ_p, see [Sil94]). Define the **conductor** N of E by

$$N := \prod_{p \text{ prime}} p^{e_p}.$$

Ultimately, the conductor will appear as a level of a newform in the Shimura-Taniyama Conjecture, hence the use of the symbol N. Note that for a semi-stable elliptic curve, the conductor of N is square-free; this case is of interest as Wiles first proved the Shimura-Taniyama Conjecture for semi-stable curves.

5 Galois Representations

In this section, we construct the "Galois representations" associated to an elliptic curve, before then doing the same for a newform.

5.1 Frobenius Elements and Galois Representations

We begin by highlighting some essential structural properties of $G_{\mathbb{Q}}$ to facilitate our study of Galois representations. First, we remark that the action of $G_{\mathbb{Q}}$ on $\overline{\mathbb{Q}}$ restricts to an action on $\overline{\mathbb{Z}} \subset \overline{\mathbb{Q}}$. Indeed, for $\sigma \in G_{\mathbb{Q}}$ and $x \in \overline{\mathbb{Z}}$, there exist $a_i \in \mathbb{Z}$ such that

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0$$

so

$$\sigma(x)^n + a_{n-1}\sigma(x)^{n-1} + \cdots + a_1\sigma(x) + a_0 = 0$$

from which it follows that $\sigma(x) \in \overline{\mathbb{Z}}$. We'll use the action of $G_{\mathbb{Q}}$ on $\overline{\mathbb{Z}}$ to define some important subgroups. Fix $p \in \mathbb{Z}$ prime and let $\mathfrak{p} \subset \overline{\mathbb{Z}}$ be a prime ideal lying over p. The decomposition group of $\mathfrak{p}$ is

$$D_{\mathfrak{p}} = \{\sigma \in G_{\mathbb{Q}} : \sigma(\mathfrak{p}) = \mathfrak{p}\}$$

a subgroup of the absolute Galois group. Because elements of the decomposition group fix $\mathfrak{p}$, they induce a natural action on $\overline{\mathbb{F}_p} = \overline{\mathbb{Z}}/\mathfrak{p}$: for $\sigma \in D_\mathfrak{p}$ and $x \in \overline{\mathbb{Z}}$,

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

At the level of $\mathbb{Z}$, we have $(\overline{\mathbb{Z}} \cap \mathbb{Z})/(\mathfrak{p} \cap \mathbb{Z}) = \mathbb{Z}/(p) = \mathbb{F}_p$; because σ fixes $\mathfrak{p} \cap \mathbb{Z}$, it follows that the induced action of σ on $\overline{\mathbb{F}_p}$ fixes $\mathbb{F}_p$. Altogether, in modding everything by $\mathfrak{p}$ we obtain a natural map $D_\mathfrak{p} \to G_{\mathbb{F}_p}$, where $G_{\mathbb{F}_p} = \mathrm{Gal}(\overline{\mathbb{F}_p}/\mathbb{F}_p)$.

Theorem 5.1. *The reduction map $D_\mathfrak{p} \to G_{\mathbb{F}_p}$ is a surjection.*

Proof. We have seen that the absolute Galois group $G_\mathbb{Q}$ may be realised as an inverse limit of finite Galois groups $\mathrm{Gal}(K/\mathbb{Q})$. In the same way, we can realise $G_{\mathbb{F}_p}$ as an inverse limit of finite Galois groups $\mathrm{Gal}(\mathbb{F}_q/\mathbb{F}_p)$. We will prove the theorem first in the case of finite extensions and then pass to the inverse limit to obtain the desired result. To this end, fix a number field $K/\mathbb{Q}$, regard $\mathfrak{p}$ as an ideal in $\mathcal{O}_K$ by considering $\mathcal{P} := \mathfrak{p} \cap \mathcal{O}_K$, and regard $D_\mathfrak{p}$ as the decomposition group in $\mathrm{Gal}(K/\mathbb{Q})$ by considering $D := D_\mathfrak{p}|_K < \mathrm{Gal}(K/\mathbb{Q})$. For our final piece of setup, denote by $\mathbb{F}_\mathfrak{p}$ the finite field $\mathcal{O}_K/\mathfrak{p}$ so that we have a reduction map $D \to \mathrm{Gal}(\mathbb{F}_\mathfrak{p}/\mathbb{F}_p)$.

Let $\overline{a} \in \mathbb{F}_\mathfrak{p}$ be such that $\mathbb{F}_\mathfrak{p} = \mathbb{F}_p(\overline{a})$. Consider the polynomial

$$t(x) = \prod_{\sigma \in D} (x - \sigma(a)),$$

which has a as a root and has coefficients which are fixed by elements of D i.e. $t(x) \in K^D$. Then the reduction $\overline{t}(x) \mod P$ — obtained by reducing the coefficients of $t(x)$ mod P — has $\overline{a}$ as a root and we assert without proof that it is the minimal polynomial of $\overline{a}$ over $\mathbb{F}_\mathfrak{p}$. By Galois theory, every Galois conjugate of $\overline{a}$ is a root of the minimal polynomial of $\overline{a}$; that $\overline{t}(x)$ is the minimal polynomial thus implies that every Galois conjugate of $\overline{a}$ takes the form $\overline{\sigma}(\overline{a})$ for some $\sigma \in D$. Because $\overline{a}$ generates the extension $\mathbb{F}_\mathfrak{p}/\mathbb{F}_p$, it follows that every element of $\mathrm{Gal}(\mathbb{F}_\mathfrak{p}/\mathbb{F}_p)$ appears as some $\overline{\sigma}$ so the reduction map $D \to \mathrm{Gal}(\mathbb{F}_\mathfrak{p}/\mathbb{F}_p)$ surjects.

Now, return to the infinite extensions and take an automorphism $\overline{\sigma} \in G_{\mathbb{F}_p}$. Regard $G_{\mathbb{F}_p}$ as an inverse limit of finite Galois groups so that we may represent $\overline{\sigma}$ with a compatible system $(\overline{\sigma_n})_{n \in I}$ of elements $\overline{\sigma_n} \in \mathrm{Gal}(\mathbb{F}_{p^n}/\mathbb{F}_p)$, where $\mathfrak{p}_n := \mathfrak{p} \cap K_n$ for some number field K_n and $\mathbb{F}_{p^n} = \mathbb{F}_{\mathfrak{p}_n} = \mathcal{O}_{K_n}/\mathfrak{p}_n$. For each $\mathfrak{p}_n$ our work in the finite case shows we can find a $\sigma_n \in D_\mathfrak{p}|_{K_n}$ such that the reduction of σ_n is precisely $\overline{\sigma_n}$. These σ_n necessarily form a compatible system for the groups $\mathrm{Gal}(K_n/\mathbb{Q})$, because they arose from a compatible system $(\overline{\sigma_n})_{n \in I}$. Then the inverse limit knits together the $(\sigma_n)_{n \in I}$ into an element $\sigma \in G_\mathbb{Q}$ which reduces to $\overline{\sigma}$. And so the reduction map $D_\mathfrak{p} \to G_{\mathbb{F}_p}$ surjects. □

We now define the inertia group of $\mathfrak{p}$ to be the kernel of the reduction map $D_\mathfrak{p} \to G_{\mathbb{F}_p}$,

$$I_\mathfrak{p} = \{\sigma \in D_\mathfrak{p} : \sigma(x) \equiv x \mod \mathfrak{p} \text{ for all } x \in \overline{\mathbb{Z}}\},$$

so that we have a short exact sequence

$$0 \to I_\mathfrak{p} \to D_\mathfrak{p} \to G_{\mathbb{F}_p} \to 0.$$

As a field of characteristic p, there is a natural automorphism of $\overline{\mathbb{F}_p}$: the Frobenius automorphism ϕ_p defined by $\phi_p(x) := x^p$. We define a Frobenius element $\mathrm{Frob}_{\mathfrak{p}}$ to be any element of the decomposition group which maps to $\phi_p \in G_{\mathbb{F}_p}$. Equivalently, a Frobenius element $\mathrm{Frob}_{\mathfrak{p}}$ is any element of $G_{\mathbb{Q}}$ such that for all $x \in \overline{\mathbb{Z}}$,

$$\mathrm{Frob}_{\mathfrak{p}}(x) \equiv \phi_p(x) = x^p \mod \mathfrak{p}.$$

Note that, by the short exact sequence above, the Frobenius element $\mathrm{Frob}_{\mathfrak{p}}$ always exists but is only defined up to multiplication by the kernel $I_{\mathfrak{p}}$.

Ideally, we would like to have a notion of Frob_p in which our Frobenius element depends only the prime p and not on our choice of $\mathfrak{p}$ lying over p. While we cannot quite obtain such a unique Frobenius element in general, we can speak of a Frobenius element up to conjugation.

Theorem 5.2. *Let $\mathfrak{p}$ and $\mathfrak{p}'$ be primes lying over p. Then there exists $\sigma \in G_{\mathbb{Q}}$ such that $\sigma(\mathfrak{p}) = \mathfrak{p}'$ and*

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

In particular, any two primes $\mathfrak{p}$ and $\mathfrak{p}'$ lying over p have conjugate Frobenius elements.

Proof. We won't prove the existence of such a $\sigma \in G_{\mathbb{Q}}$ as it isn't our main focus. So assume $\sigma(\mathfrak{p}) = \mathfrak{p}'$. Then the definition of $\mathrm{Frob}_{\mathfrak{p}}$ yields

$$\begin{aligned}
\mathrm{Frob}_{\mathfrak{p}} \text{ is a Frobenius over } \mathfrak{p} &\iff \mathrm{Frob}_{\mathfrak{p}}(\sigma^{-1}x) \equiv (\sigma^{-1}x)^p \mod \mathfrak{p} \\
&\iff \mathrm{Frob}_{\mathfrak{p}}(\sigma^{-1}x) - (\sigma^{-1}x)^p \in \mathfrak{p} \\
&\iff \sigma \mathrm{Frob}_{\mathfrak{p}} \sigma^{-1} x - x^p \in \sigma(\mathfrak{p}) \\
&\iff \sigma \mathrm{Frob}_{\mathfrak{p}} \sigma^{-1} x \equiv x^p \mod \sigma(\mathfrak{p}) \\
&\iff \sigma \mathrm{Frob}_{\mathfrak{p}} \sigma^{-1} \text{ is a Frobenius over } \sigma(\mathfrak{p}) \\
&\iff \sigma \mathrm{Frob}_{\mathfrak{p}} \sigma^{-1} = \mathrm{Frob}_{\sigma(\mathfrak{p})}
\end{aligned}$$

as desired. □

Of course, a natural concern arises: when we write $\sigma \mathrm{Frob}_{\mathfrak{p}} \sigma^{-1} = \mathrm{Frob}_{\sigma(\mathfrak{p})}$, the left-hand-side is only defined up to $I_{\mathfrak{p}}$ while the right-hand-side is only defined up to $I_{\sigma(\mathfrak{p})}$. We allow this ambiguity because once again we have equality up to conjugation:

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

Now, we care so much about Frobenius elements because they form a dense subset of "much" of the absolute Galois group. Indeed, although we would like a dense subset of all of $G_{\mathbb{Q}}$, we have to settle for something a touch weaker.

Theorem 5.3 (A Simple Version of the Chebotarev Density Theorem)**.** *Let S be a finite set of primes in $\mathbb{Z}$ and let $\mathbb{Q}^{unr,S}$ denote the maximal extension of $\mathbb{Q}$ unramified outside of S i.e. the union of all finite extensions $L/\mathbb{Q}$ for which any prime $p \notin S$ does not ramify in $\mathcal{O}_L$. For each prime $p \notin S$, let*

$$F_p = \bigcup_{\mathfrak{p}|p} \{\textit{conjugacy class of Frobenius elements } \mathrm{Frob}_{\mathfrak{p}}|_{\mathbb{Q}^{unr,S}}\}.$$

Then the union of all such F_p is dense in $\mathrm{Gal}(\mathbb{Q}^{unr,S}/\mathbb{Q})$. Succinctly, the Frobenius elements of unramified primes are dense in $\mathrm{Gal}(\mathbb{Q}^{unr,S}/\mathbb{Q})$.

In this statement, density is determined with respect to the Krull topology.

That we have a dense set is of great interest for continuity reasons: the image of a continuous map on $\mathrm{Gal}(\mathbb{Q}^{unr,S}/\mathbb{Q})$ is completely determined by the map's image on Frobenius elements $\mathrm{Frob}_{\mathfrak{p}}$. As we have seen, however, the symbol $\mathrm{Frob}_{\mathfrak{p}}$ is only defined up to conjugation, so it makes sense to study continuous maps which are themselves defined up to conjugation; this motivates our study of representations in the next section which are, by definition, determined only up to conjugation.

Definition 5.4. A **d-dimensional ℓ-adic Galois representation** is a continuous homomorphism

$$\rho : G_{\mathbb{Q}} \to \mathrm{GL}_d(\mathbb{L})$$

where $\mathbb{L}$ is a finite extension of $\mathbb{Q}_\ell$. Continuity is determined with respect to the Krull topology on the domain and the subspace toplogy (of $\mathbb{L}^{d^2}$) on the codomain. Two Galois representations ρ_1, ρ_2 are isomorphic if there exists a matrix $M \in \mathrm{GL}_d(\mathbb{L})$ such that $\rho_1(g) = M\rho_2(g)M^{-1}$ for all $g \in G_{\mathbb{Q}}$.

The Galois representations associated to elliptic curves will have $\mathbb{L} = \mathbb{Q}_\ell$ while the Galois representations associated to a newform f will have $\mathbb{L} = \mathbb{K}_{f,\lambda}$, a finite extension of $\mathbb{Q}_\ell$ obtained from the number field $\mathbb{K}_f$.

Just as we had a notion of ramification for primes in a number field, so too does there exist a notion of ramification for Galois representations. In both cases, ramification depends on the behaviour for a prime $\mathfrak{p}$ lying above a rational prime p.

Definition 5.5. Let $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_d(\mathbb{L})$ be a Galois representation. We say that ρ is **unramified** at a prime $p \in \mathbb{Z}$ if for any prime $\mathfrak{p} \subset \overline{\mathbb{Z}}$ lying over p, the inertia group is killed by ρ i.e. $I_{\mathfrak{p}} \subset \ker \rho$.

So we now have two uses of the word "unramified": a number field in which p is unramified and a Galois representation which is unramified at p. Let's connect these two ideas. Note that ρ necessarily factors through a unique quotient as in the commutative diagram

$$\begin{tikzcd}
G_{\mathbb{Q}} \arrow[rr, "\rho"] \arrow[dr, "r"'] & & \mathrm{GL}_2(\mathbb{Q}_\ell) \\
& \mathrm{Gal}(\overline{\mathbb{Q}}^{\ker \rho}/\mathbb{Q}) \arrow[ur, "i"'] &
\end{tikzcd}$$

such that i is injective, where $\overline{\mathbb{Q}}^{\ker \rho}$ denotes the fixed field of $\ker \rho$ and r denotes the natural restriction map. If ρ is unramified at p, then for all $\mathfrak{p}$ lying above p, we have $\ker \rho \supset I_{\mathfrak{p}}$. It turns out that this implies that p will be unramified in $\overline{\mathbb{Q}}^{\ker \rho}$, so ρ in some sense kills the part of $G_{\mathbb{Q}}$ where p ramifies. We care about ramification because of theorem 5.3, for which we need to exclude finitely-many ramified primes.

5.2 Galois Representation Associated to an Elliptic Curve

Let E be an elliptic curve over $\mathbb{Q}$. To define a Galois representation associated to E, we consider the natural action of $G_{\mathbb{Q}}$ on E. In particular, an automorphism $\sigma \in G_{\mathbb{Q}}$ acts on a point $[x_1 : x_2 : x_3]$ of E, $x_i \in \overline{\mathbb{Q}}$, by acting componentwise: $\sigma([x_1 : x_2 : x_3]) := [\sigma(x_1) : \sigma(x_2) : \sigma(x_3)]$. Note that the action produces a well-defined map on $\mathbb{P}^3(\overline{\mathbb{Q}})$ because homomorphisms are multiplicative.

Now, recall that multiplication by m on E may be written as rational polynomials with rational coefficients; because the coefficients are fixed by $\sigma \in G_{\mathbb{Q}}$, the action of the absolute Galois group commutes with multiplication by m. That is, $G_{\mathbb{Q}}$ sends torsion subgroups into themselves, so the action restricts to $E[\ell^n]$ for all positive ℓ and n. Regard the action of $G_{\mathbb{Q}}$ on ℓ^n-torsion as a map $\sigma_n : G_{\mathbb{Q}} \to \mathrm{Aut}(\mathbb{Z}/\ell^n\mathbb{Z})$ to obtain the following commutative diagram

$$\begin{tikzcd}
 & G_{\mathbb{Q}} \arrow[dl, "\sigma_n"'] \arrow[dr, "\sigma_{n+1}"] & \\
\mathrm{Aut}(E[\ell^n]) & & \mathrm{Aut}(E[\ell^{n+1}]), \arrow[ll]
\end{tikzcd}$$

where the bottom map $\mathrm{Aut}(E[\ell^{n+1}]) \to \mathrm{Aut}(E[\ell^n])$ arises from the natural multiplication-by-ℓ map $E[\ell^{n+1}] \to E[\ell^n]$. But by theorem 4.14, we have an isomorphism $E[\ell^n] \cong (\mathbb{Z}/\ell^n\mathbb{Z})^2$. Thus, $\mathrm{Aut}(E[\ell^n]) \cong \mathrm{GL}_2(\mathbb{Z}/\ell^n\mathbb{Z})$ and the diagram becomes

$$\begin{tikzcd}
 & G_{\mathbb{Q}} \arrow[dl, "\sigma_n"'] \arrow[dr, "\sigma_{n+1}"] & \\
\mathrm{Aut}(E[\ell^n]) \arrow[d, "\wr"'] & & \mathrm{Aut}(E[\ell^{n+1}]) \arrow[ll] \arrow[d, "\wr"'] \\
\mathrm{GL}_2(\mathbb{Z}/\ell^n\mathbb{Z}) & & \mathrm{GL}_2(\mathbb{Z}/\ell^{n+1}\mathbb{Z})) \arrow[ll]
\end{tikzcd}$$

Applying the inverse limit — remembering that $\varprojlim_n E[\ell^n] = \mathrm{Ta}_\ell(E)$ and $\varprojlim_n \mathbb{Z}/\ell^n\mathbb{Z} = \mathbb{Z}_\ell$ — we obtain a series of maps

$$G_{\mathbb{Q}} \to \mathrm{Aut}(\mathrm{Ta}_\ell(E)) \xrightarrow{\sim} \mathrm{GL}_2(\mathbb{Z}_\ell) \hookrightarrow \mathrm{GL}_2(\mathbb{Q}_\ell).$$

Their composition yields a map $\rho_{E,\ell} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Q}_\ell)$.

Definition 5.6. The map $\rho_{E,\ell} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Q}_\ell)$ obtained in the preceding construction is the ℓ-adic Galois representation associated to the elliptic curve E. Note that dependence on E enters in the action of $G_{\mathbb{Q}}$ on the Tate module $\mathrm{Ta}_\ell(E)$.

Recall that for $\rho_{E,\ell} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Q}_\ell)$ to define a Galois representation, it must be continuous. We'll argue continuity assuming theorem 1.14 and the following fact: a basis for the topology on $\mathrm{GL}_2(\mathbb{Q}_\ell)$ is the collection

$$\{U_M(n)\}_{M\in \mathrm{GL}_2(\mathbb{Q}_\ell), n\in\mathbb{Z}_{\geq 0}}, \text{ where } U_M(n) := M(I + \ell^n M_2(\mathbb{Z}_\ell))$$

and $M_d(\mathbb{Z}_\ell)$ denotes the set of d by d matrices with entries in $\mathbb{Z}_\ell$. Because a map is continuous if and only if the inverse image of any basis set is open, we need only show that $\rho_{E,\ell}^{-1}(U_m(n))$ is open for all M and n. Moreover, multiplication by $M \in \mathrm{GL}_2(\mathbb{Q}_\ell)$ defines a continuous map on $\mathrm{GL}_2(\mathbb{Q}_\ell)$, so it actually suffices to show that $\rho_{E,\ell}^{-1}(U_1(n))$ is open for all m.

Take $M \in U_1(n)$ and note that for any $e \leq n$ we have $M \equiv I \mod \ell^e$. In particular,

$$M = (\underbrace{I, \ldots, I}_{n \text{ times}}, M_{n+1}, M_{n+2}, \ldots) \tag{7}$$

where M_e denotes the reduction of M mod ℓ^e. As such, when we pull M back along $\rho_{E,\ell}^{-1}$ to $G_{\mathbb{Q}}$, we wind up with elements of the form

$$(\underbrace{1, \ldots, 1}_{n \text{ times}}, \sigma_{i_{n+1}}, \sigma_{i_{n+2}}, \ldots) \in G_{\mathbb{Q}} \tag{8}$$

where the first n identity entries correspond to finite Galois extensions $\mathrm{Gal}(K_{i_j}/\mathbb{Q})$ containing the ℓ^j-torsion coordinates. Moreover, every element in $G_{\mathbb{Q}}$ of the form in (8) maps to something of the form in (7). So we have

$$H := \rho_{E,\ell}^{-1}(U_m(n)) = \{\sigma \in G_{\mathbb{Q}} : \sigma = (\underbrace{1, \ldots, 1}_{n \text{ times}}, \sigma_{i_{n+1}}, \sigma_{i_{n+2}}, \ldots)\},$$

a subgroup of $G_{\mathbb{Q}}$ whose elements fix the number fields K_{i_j}. But then $\overline{\mathbb{Q}}^H = \bigcup_{j=1}^n K_{i_j}$ is a finite extension of $\mathbb{Q}$ such that $H = \mathrm{Gal}(K^H/\mathbb{Q})$. In particular, theorem 1.14 says that H is open in the Krull topology, so $\rho_{E,\ell}$ is continuous.

Now that we have the Galois representation $\rho_{E,\ell}$, we enumerate some of its properties.

Theorem 5.7. *Let E be an elliptic curve over $\mathbb{Q}$ with conductor N and fix a prime ℓ. Then $\rho_{E,\ell}$ is unramified for all primes $p \nmid \ell N$ and for any $\mathfrak{p}$ a prime lying over p, the image $\rho_{E,\ell}(\mathrm{Frob}_{\mathfrak{p}})$ satisfies the equation*

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

where $a_p(E) := p + 1 - \#E(\mathbb{F}_p)$. Moreover, $\rho_{E,\ell}$ is irreducible.

For clarification, we use # to denote cardinality. There are four important things to remark.

First, note that the characterstic polynomial of $\rho_{E,\ell}(\mathrm{Frob}_{\mathfrak{p}})$ is entirely independent of ℓ, and instead only depends on E and p.

Second, that $\rho_{E,\ell}$ is unramified at all but the finitely many primes dividing ℓN shows that $\rho_{E,\ell}$ factors through a Galois group simpler than all of $G_{\mathbb{Q}}$. To see this, let S denote the set of primes dividing ℓN and denote by H the smallest closed normal subgroup of $G_{\mathbb{Q}}$ containing I_p for all $p \notin S$ — we require that H be normal to avoid the ambiguity of conjugate inertial groups. By virtue of being unramified outside of S, the representation $\rho_{E,\ell}$ is trivial on H. Apply the fundamental theorem of infinite Galois theory — which gives a correspondence between intermediate fields and *closed* subgroups of the Galois group (see [Mil18]) — to see that $H = \mathrm{Gal}(\overline{\mathbb{Q}}/\overline{\mathbb{Q}}^H)$ and $\overline{\mathbb{Q}}^H$ is Galois over $\mathbb{Q}$. The following lemma identifies $\overline{\mathbb{Q}}^H$ more explicity.

Lemma 5.8. *We have $\overline{\mathbb{Q}}^H = \mathbb{Q}^{unr,S}$, the maximal extension of $\mathbb{Q}$ unramified outside of S.*

By the lemma, $H = \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}^{unr,S})$. Because $\rho_{E,\ell}$ kills the generators of H, the continuity of $\rho_{E,\ell}$ and that H is closed imply that $\rho_{E,\ell}$ kills all of $H = \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}^{unr,S})$. So $\rho_{E,\ell}$ is entirely determined by its action at the level of $\mathbb{Q}^{unr,S}/\mathbb{Q}$, rather than $\overline{\mathbb{Q}}/\mathbb{Q}$. And so we have proven our desired result.

Theorem 5.9. *The representation $\rho_{E,\ell} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Q}_\ell)$ factors through $\mathrm{Gal}(\mathbb{Q}^{unr,S}/\mathbb{Q})$. That is, we have a commutative diagram*

$$\begin{tikzcd}
G_{\mathbb{Q}} \arrow[rr, "\rho_{E,\ell}"] \arrow[dr, "r"'] & & \mathrm{GL}_2(\mathbb{Q}_\ell) \\
& \mathrm{Gal}(\mathbb{Q}^{unr,S}/\mathbb{Q}) \arrow[ur, "\rho"'] &
\end{tikzcd}$$

where r denotes the natural restriction map and ρ is the unique continuous map which fills in the diagram.

Third, because the Frobenius element Frob_p is only determined up to conjugation, a characteristic equation $x^2 - a_p(E)x + p = 0$ is the best possible determination of $\rho_{E,\ell}(\mathrm{Frob}_p)$ we could hope for. Two conjugate Frobenius elements might have distinct images in $\mathrm{GL}_2(\mathbb{Q}_\ell)$, but as conjugates they will necessarily have the same characteristic equation.

Fourth, the two previous remarks allow us to appeal to the Chebotarev density theorem (theorem 5.3). Indeed, the density of the Frobenius elements (of unramified primes) in $\mathrm{Gal}(\mathbb{Q}^{unr,S}/\mathbb{Q})$ tells us that ρ is entirely determined by where it sends Frobenii. So the commutative diagram shows that $\rho_{E,\ell}$ is itself determined by its image on (unramified) Frobenii. But theorem 5.7 provides the characteristic equations of these Frobenius elements, so checking the equality of $\rho_{E,\ell}$ and some other representations reduces to checking equality only at Frobenius elements. We will return to these idea in section 6, after constructing similar representations for newforms.

5.3 Galois Representation Associated to a Newform

Let $f \in M_2(\Gamma_1(N))$ be a newform and let $X_1(N)$ denote $\mathbb{H} \cup \mathbb{Q} \cup \{\infty\}$ mod the action of $\Gamma_1(N)$ as in definition 2.12. Note that f has weight $k = 2$; as weight two newforms are the only ones to appear in the STC correspondence, we need only consider Galois representations for weight-2 forms.

The construction of the Galois representation associated to f is much more complicated than that for an elliptic curve, and we simply have not developed the necessary background here. As such, we offer only an outline and refer the reader to section 9.5 of [DS05] for additional details. The construction proceeds in three stages:

1. Regard $X_1(N)$ as a curve over $\mathbb{C}$ and then transfer this to $X_1(N)$ as a curve over $\mathbb{Q}$ (by analyzing the fields of meromorphic functions on each). Then the first step is to construct a Galois representation

$$\rho_{X_1(N),\ell} : G_{\mathbb{Q}} \to \mathrm{GL}_{2g}(\mathbb{Q}_\ell)$$

where g is the "genus" of $X_1(N)$ as a curve over $\mathbb{Q}$.

2. Associate to f a "complex torus" A_f with dimension d. Then construct a Galois representation

$$\rho_{A_f,\ell} : G_{\mathbb{Q}} \to \mathrm{GL}_{2d}(\mathbb{Q}_\ell)$$

where we use the representation $\rho_{X_1(N),\ell}$ in part 1 to prove that $\rho_{A_f,\ell}$ is indeed a representation with some nice properties.

3. Decompose the representation $\rho_{A_f,\ell}$ into some natural pieces to obtain representations
$$\rho_{f,\lambda} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Q}_\ell)$$
where λ is any prime lying over ℓ in the number field $\mathbb{K}_f$.

In steps 1 and 2, the construction of $\rho_{X_1(N),\ell}$ and $\rho_{A_f,\ell}$ proceeds quite similarly to that of $\rho_{E,\ell}$ for E an elliptic curve. Indeed, one can define the Tate module of both (the "Jacobian" of) $X_1(N)$ and A_f using ℓ^n-torsion and then construct the representation as before. In this case, however, the points on either $X_1(N)$ or A_f don't necessarily form a group (in this way, elliptic curves are quite special), so one needs to use the "Picard group" of the curve/surface; for an elliptic curve E, the Picard group of E is isomorphic to the group of points on E so this more general construction agrees with what we did for elliptic curves. Either way, the Tate modules $\mathrm{Ta}(X_1(N))$ and $\mathrm{Ta}(A_f)$ arise as inverse limits of torsion as before. Similar to the elliptic curve case, this process yields isomorphisms $\mathrm{Ta}(X_1(N)) \cong \mathbb{Z}_\ell^{2g}$ and $\mathrm{Ta}(A_f) \cong \mathbb{Z}_\ell^{2d}$ and so obtain representations with dimensions $2g$ and $2d$.

The isomorphism $\mathrm{Ta}(A_f) \cong \mathbb{Z}_\ell^{2d}$ becomes quite important for step 3 as well. Tensoring $\mathbb{Z}_\ell^{2d}$ with $\mathbb{Q}_\ell$ over $\mathbb{Z}_\ell$ in effect inverts the elements of $\mathbb{Z}_\ell$ so that
$$V_\ell(A_f) := \mathrm{Ta}(A_f) \otimes_{\mathbb{Z}_\ell} \mathbb{Q}_\ell \cong \mathbb{Q}_\ell^{2d}$$
is the vector space underying the representation $\rho_{A_f,\ell}$. Thus, to decompose $\rho_{A_f,\ell}$ we need only decompose $V_\ell(A_f)$.

Lemma 5.10. *We have a decomposition $V_\ell(A_f) \cong (\mathbb{K}_f \otimes_{\mathbb{Q}} \mathbb{Q}_\ell)^2$.*

So it remains to decompose $\mathbb{K}_f \otimes_{\mathbb{Q}} \mathbb{Q}_\ell$. We will do this for an arbitrary number field before then specializing to the case of $\mathbb{K}_f$.

Let $\mathbb{K}$ be a number field with ring of integers $\mathcal{O}_{\mathbb{K}}$ and fix a prime ℓ. For each ideal λ lying over ℓ, define the ring of λ-adic integers $\mathcal{O}_{\mathbb{K},\lambda}$ by the inverse limit
$$\mathcal{O}_{\mathbb{K},\lambda} = \varprojlim_n \mathcal{O}_{\mathbb{K}}/\lambda^n$$
and define the field of λ-adic numbers $\mathbb{K}_\lambda$ as the field of fractions of $\mathcal{O}_{\mathbb{K},\lambda}$. Notice that this agrees with the definition of the ℓ-adic integers/numbers in example 1.11 by taking $\mathbb{K} = \mathbb{Q}$. Indeed, because $\ell\mathbb{Z} \subset \lambda\mathcal{O}_{\mathbb{K}}$, there is a natural embedding of $\mathbb{Z}_\ell \hookrightarrow \mathcal{O}_{\mathbb{K},\lambda}$ and thus an embedding $\mathbb{Q}_\ell \hookrightarrow \mathbb{K}_\lambda$. The latter makes $\mathbb{K}_\lambda$ into a field extension of $\mathbb{Q}_\ell$, and the extension is finite because $\mathbb{K}/\mathbb{Q}$ is. Now, we can state our desired result and apply it to our particular case.

Lemma 5.11. *We have a decomposition $\mathbb{K} \otimes_{\mathbb{Q}} \mathbb{Q}_\ell \cong \prod_{\lambda|\ell} \mathbb{K}_\lambda$ where the notation $\lambda|\ell$ denotes that the product ranges over all primes $\lambda \subset \mathcal{O}_{\mathbb{K}}$ lying over ℓ.*

Proof. Refer to section 9.2 of [DS05] the proof. □

For λ a prime in $\mathbb{K}_f$ lying over ℓ, denote by $K_{f,\lambda}$ the field of λ-adic integers. Together the lemmas say that
$$V_\ell(A_f) \cong \left(\prod_{\lambda|\ell} \mathbb{K}_{f,\lambda} \right)^2$$

so that the representation $\rho_{A_f,\lambda}$ yields a homomorphism

$$G_{\mathbb{Q}} \to \mathrm{GL}_2\left(\prod_{\lambda|\ell} \mathbb{K}_{f,\lambda}\right).$$

For a fixed λ, we may apply a projection from the product to obtain a homomorphism

$$\rho_{f,\lambda} : G_{\mathbb{Q}} \to \mathrm{GL}_2\left(\mathbb{K}_{f,\lambda}\right).$$

Definition 5.12. The homomorphism $\rho_{f,\lambda}$ is the ℓ-adic Galois representation associated to the newform f.

Contrast the two representations we have constructed: the Galois representation $\rho_{E,\ell}$ appeals to the case of definition 5.4 when $\mathbb{L} = \mathbb{Q}_\ell$, whereas $\rho_{f,\lambda}$ requires a finite extension $\mathbb{L} = \mathbb{K}_{f,\lambda}$. For these definitions to align — as they must for our desired correspondence — we will need $\mathbb{K}_{f,\lambda} = \mathbb{Q}_\ell$. Based on the definition of $\mathbb{K}_{f,\lambda}$, this occurs when $\mathbb{K}_f = \mathbb{Q}$. As we will see in the next section, the Shimura-Taniyama Conjecture provides the existence of such newforms. But first we again summarise the properties of the representation just constructed; note their incredible similarity to those of $\rho_{E,\ell}$.

Theorem 5.13. *Let $f \in S_2(N, \chi)$ be a newform with number field $\mathbb{K}_f$ and fix a prime ℓ. For each prime λ of $\mathcal{O}_{\mathbb{K}_f}$ lying over ℓ, let $\rho_{f,\lambda} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{K}_{f,\lambda})$ be the Galois representation associated to f. Then $\rho_{f,\lambda}$ is unramified at primes away from ℓ and N (i.e. is unramified for every prime $p \nmid \ell N$). Moreover, for $\mathfrak{p}$ a prime over p, $\rho_{f,\lambda}(\mathrm{Frob}_{\mathfrak{p}})$ satisfies the equation*

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

In particular, if $f \in S_2(\Gamma_0(N))$ then $\chi(p) = 1$ so that the equation becomes $x^2 - a_p(f)x + p = 0$.

Note once again that the characteristic equation of $\rho_{f,\lambda}(\mathrm{Frob}_{\mathfrak{p}})$ does not depend on ℓ! Indeed, for the representations $\rho_{f,\lambda}$ and $\rho_{E,\ell}$ to align, we'll want their characteristic equations at $\mathrm{Frob}_{\mathfrak{p}}$ to agree. As in the theorem, this case occurs when $f \in S_2(\Gamma_0(N))$. The Shimura-Taniyama Conjecture will guarantee that such newforms suffice to give a correspondence, so let's finally get on to stating STC.

6 The Shimura-Taniyama Conjecture

In its many forms, the Shimura-Taniyama Conjecture (STC) states that every elliptic curve is "modular". In this paper we have developed theory for understanding the Galois theoretic statement of STC, the version which Wiles proved (with some help from Taylor) in 1995:

Theorem 6.1. *(Galois Theoretic STC for Semi-Stable Elliptic Curves) Let E be a semi-stable elliptic curve over $\mathbb{Q}$ with conductor N. Then there exists a newform $f \in S_2(\Gamma_0(N))$ with number field $K_f = \mathbb{Q}$ such that $\rho_{f,\ell} \sim \rho_{E,\ell}$ for all ℓ.*

Thus, a semi-stable elliptic curve $E/\mathbb{Q}$ is called "modular" because it corresponds — through its Galois representations — to some modular form f. We of course have no hope of going through the proof of STC here. Instead, we go through an example which illustrates the incredible power of this correspondence.

Consider the congruence subgroup $\Gamma_0(11)$ and let $X_0(11)$ denote $\Gamma_0(11)\backslash\mathbb{H}^*$ as in definition 2.12. The expository article [Wes99] goes through and deduces that a fundamental domain for $X_0(11)$ is the blue region in figure 6. Note that, although the fundamental domain has five limits on the real line, $X_0(11)$ only has two distinct cusps. Indeed, figure 7 shows the fundamental domain wrapped up with two sides identified; the two cusps are the cusp at infinity in the center of the square and the single cusp at all four vertices of the square. From topology, we recognise the square with these side identifications as a torus, so we can realise $X_0(11)$ as a compact surface with genus one.

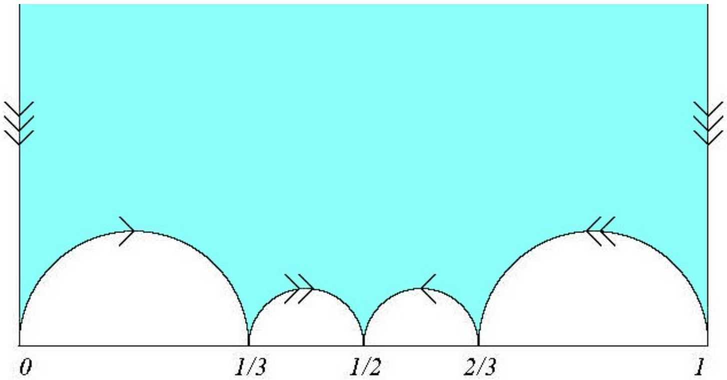


Figure 6: A fundamental domain for $\Gamma_0(11)$. The arrows denote sides which are identified under the action of some matrix of $\Gamma_0(11)$. Notice that this fundamental domain nevertheless contains a fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$, but contains one different from that in figure 1. Image adapted from [Wes99].

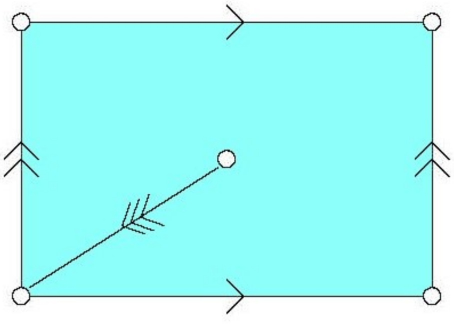


Figure 7: The fundamental domain for $\Gamma_0(11)$ in figure 6 with the triple-arrow boundaries identified and the other boundaries straightened. Image adapted from [Wes99].

Alternatively, an algebraic argument in section 3.1 and exercise 3.1.4 of [DS05] deduces that the genus of $X_0(11)$ is $\lfloor \frac{11+1}{12} \rfloor = 1$. Either way, recalling that an elliptic curve is a smooth projective curve of genus 1 allows us to regard $X_0(11)$ itself as an elliptic curve! (We ignore the subtle connections between the genus of a surface and the genus of a curve here; [Wes99] explicitly goes through the process of transferring $X_0(11)$ from a genus one surface to a genus one curve over $\mathbb{Q}$.)

Because $X_0(11)$ defines an elliptic curve, it has an associated Weierstrass equation (6). Once again, refer to [Wes99] for the details of how to discern that $X_0(11)$ has (minimal) Weierstrass equation

$$E : y^2 + y = x^3 - x^2 - 10x - 20$$

as an elliptic curve E over $\mathbb{Q}$. Equation in hand, let's see what the Shimura-Taniyama Conjecture, theorem 5.7, and theorem 5.13 say about E and its corresponding newform f.

By STC, we can determine the level N of f by determining the conductor of E. Recalling the notation from definition 4.7, we have that

$$\begin{aligned} b_2 &= -4, \\ b_4 &= -20, \\ b_6 &= -79, \text{ and} \\ b_8 &= -21. \end{aligned}$$

so that E has discriminant

$$\Delta = -161,051 = -1 \cdot 11^5$$

and c_4 value

$$c_4 = 496 = 2^4 \cdot 31.$$

The discriminant shows that E has bad reduction at 11 and good reduction at all other primes. Because 11 does not divide the c_4 value, E has semi-stable (multiplicative) reduction at 11 so definition 4.23 yields that that E has conductor $N := 11$. So by STC we seek a newform $f \in S_2(11)$.

Comparing theorems 5.7 and 5.13 and the images of our Galois representations at Frobenius elements, we see that $\rho_{E,\ell} \sim \rho_{f,\ell}$ forces

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

for all $p \nmid 11 \cdot \ell$. Varying ℓ, we obtain $p + 1 - \#\tilde{E}(\mathbb{F}_p) = a_p(E) = a_p(f)$ for all $p \neq 11$. We thus compute a couple values of $\#\tilde{E}(\mathbb{F}_p)$ to get the first couple Fourier coefficients of f.

- For $p = 2$, we have an equation $\tilde{E} : y^2 + y = x^3 + x^2$. By Fermat's Little Theorem mod two, this reduces to $y + y = x + x$ which holds for all points $(x, y) \in \mathbb{F}_2^2$. Together with the point at infinity, we have $\#\tilde{E}(\mathbb{F}_2) = 5$ so that $a_2(E) = 2 + 1 - 5 = -2$.

- For $p = 3$, reduction yields $\tilde{E} : y^2 + y = 2x^2 + 1$ (again using Fermat's Little Theorem). If $y = 0$, then $x = \pm 1$ and if $x = 0$, there are no solutions with $y \in \mathbb{F}_3$. So if neither x nor y equals zero, we may use $a^2 = 1 \mod 3$ for $a \neq 0$ to obtain $y + 1 = 0$. This gives the points $y = -1$ and $x = \pm 1$. Together with the point at infinity, we count $\#\tilde{E}(\mathbb{F}_3) = 5$ so that $a_3(E) = 3 + 1 - 5 = -1$.

- For $p = 5$, similar arguments become increasingly casewise (but no less interesting or useful); for brevity, we had a computer check all twenty-five to discover that four points lie on $\tilde{E}$: $(0, 0)$, $(0, 4)$, $(1, 0)$, and $(1, 4)$. So $\#\tilde{E}(\mathbb{F}_5) = 5$ so that $a_5(E) = 5 + 1 - 5 = 1$.

It follows by STC that

$$f(q) = q - 2q^2 - q^3 + a_4(f)q^4 + q^5 + \cdots$$

and, if we had further knowledge of the Dirichlet character χ for which $f \in S_2(11, \chi)$, we could determine $a_4(f)$ using $a_2(f)$. The prime coefficients suffice, however, in the following way.

Chapter 3 of [DS05] computes dimension formulas for the vector spaces (over $\mathbb{C}$) of modular forms $M_k(\Gamma_0(N))$ and of cusp forms $S_k(\Gamma_0(N))$. In particular, the dimension is always finite and (as we've seen previously in theorem 3.22) the set of newforms constitutes a basis for the subspace $S_k(\Gamma_0(N))^{new} \subset S_k(\Gamma_0(N))$. So in computing enough Fourier coefficients $a_p(f)$, we can uniquely determine which of the finitely-many newforms f must be. Referencing the dimension formulas in [DS05], we find that $\dim_{\mathbb{C}} S_2(\Gamma_0(N))$ is the genus of $X_0(N)$ as a Riemann surface. For $N = 11$, in particular, we have that the dimension of $S_2(\Gamma_0(11))$ equals 1 so there is but one possible newform: the unique newform of weight 2 and level 11, label 11.2.a.a in the LMFDB:

$$f(q) = q - 2q^2 - q^3 + 2q^4 + q^5 + 2q^6 - 2q^7 - 2q^9 - 2q^{10} + q^{11} - 2q^{12} + 4q^{13} + \cdots .$$

As they must, the Fourier coefficients we have so far computed agree with those given on LMFDB. Indeed, that $a_7(f) = -2$ in turn implies that $\tilde{E}(\mathbb{F}_7)$ has 10 points and that $a_{13}(f) = 4$ implies that $\tilde{E}(\mathbb{F}_{13})$ has 10 as well. Notice that we do not obtain information for $p = 11$ because the correspondence $a_p(f) = a_p(E)$ only holds for primes of good reduction (i.e. for primes not dividing the conductor N). Theorem 5.9 and the discussion which follows justify that the equality of images of Frobenius elements at good primes ensures the correspondence $\rho_{E,\ell}$ and $\rho_{f,\ell}$ of the Galois representations at all primes ℓ. At the bad primes ℓ dividing N, the behaviour of the Galois representations are more complicated — notice that theorems 5.7 and 5.13 only give information for good primes — but one can nevertheless make sense of the correspondence there. For details on the correspondence at primes of semi-stable and unstable reduction, refer to chapters 4 and 5 of [Sil94].

Let's compute one more example, so that that we can see a case when $S_2(\Gamma_0(N))$ has dimension greater than one. In particular, consider the elliptic curve E with (minimal) Weierstrass equation $y^2 + y = x^3 + x - 1$. In the same manner as before we can compute that E has discriminant $\Delta = -1 \cdot 307$ and c_4 value $c_4 = -48$. Thus, E has semi-stable bad reduction at 307 and no other primes, and thus has conductor $N = 307$. So STC guarantees that E corresponds to some newform $f \in S_k(\Gamma_0(307))$.

As before, we appeal to chapter 3 of [DS05] for the dimensions of the spaces of cusp forms; exercise 3.1.4 and section 3.5 together show that $\dim_{\mathbb{C}} S_2(\Gamma_0(307)) = 25$. Because $N = 307$ is prime, $S_2(\Gamma_0(307)) = S_2(\Gamma_0(307))^{new}$ and this means that there are twenty-five newforms of weight 2 and level 307. Of these, only four have number field $\mathbb{Q}$:

$$\begin{array}{lllllllllll}
f_1(q) & := q & & & -2q^4 & +4q^5 & & & & -3q^9 & + \cdots \\
f_2(q) & := q & +2q^2 & +2q^3 & +2q^4 & & +4q^6 & -3q^7 & & +q^9 & + \cdots \\
f_3(q) & := q & +2q^2 & & +2q^4 & +2q^5 & & +3q^7 & & -3q^9 & + \cdots \\
f_4(q) & := q & +q^2 & +2q^3 & -q^4 & & +2q^6 & +3q^7 & -3q^8 & +q^9 & + \cdots .
\end{array}$$

So STC requires that one of the four newforms f_i corresponds to E. Compute that $\#E(\mathbb{F}_2) = 1$ to narrow the choices to f_2 and f_3. Then compute that $\#E(\mathbb{F}_3) = 4$ to determine that E has corresponding newform f_3. It follows that $\#E(\mathbb{F}_7) = 5$ and — using LMFDB to see that $a_{97}(f_3) = 11$ — that $\#E(\mathbb{F}_{97}) = 87$.

7 Overview of the Proof of Fermat's Last Theorem

Historically, the proof of Fermat's Last Theorem (FLT) proceeded through a series of reductions. In 1995, Wiles (with some help from Taylor) added the final link in the chain by proving the Shimura-Taniyama Conjecture (STC) for semi-stable elliptic curves. The proof manifests as an enormous proof by contradiction, as it begins by assuming a hypothetical solution to $a^\ell + b^\ell = c^\ell$ with $abc \neq 0$ and $\ell > 7$ and ultimately contradicts the non-existence of modular forms of weight $k = 2$ and level $N = 2$:

1. In 1955, Shimura and Taniyama make their famous conjecture that all elliptic curves over $\mathbb{Q}$ (in particular, the semi-stable ones) are modular.

2. In 1975, Hellegouarch took a hypothetical non-trivial solution $a^\ell + b^\ell = c^\ell$ to the Fermat equation and considered the elliptic curve

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

   Subsequent work by Frey showed that E is semi-stable and that its associated Galois representation $\rho_{E,\ell}$ is irreducible for $\ell > 7$. By passing through STC, this linked FLT with modular forms.

3. In the late 80's, Serre devised an incredible series of powerful conjectures, which detail how various Galois representations "arise" from modular forms in a precise way. In particular, Serre conjectured that the Galois representations $\rho_{E,\ell}$ associated to the Frey-Hellegouarch curve arise from modular forms of weight $k = 2$ and level $N = 2$.

4. In 1990, Ribet proved that the Shimura-Taniyama conjecture implies Serre's conjecture for the representations $\rho_{E,\ell}$. Because there are no modular forms of weight $k = 2$ and level $N = 2$, the truth of STC would contradict the existence of the elliptic curve E, which in turn contradicts the existence of a solution $a^\ell + b^\ell = c^\ell$.

5. In 1995, Wiles (with some help from Taylor) furnished the final link in the chain by proving STC for semi-stable elliptic curves with irreducible representations $\rho_{E,\ell}$; by the properties of the Frey Curve E, Wiles's work sufficed to prove Fermat's Last Theorem.

And so, 358 years after Fermat's original note, the mathematical community completed the proof of Fermat's Last Theorem. While we have highlighted many mathematicians here, without a doubt there are hundreds of others who contributed.

We stated the Shimura-Taniyama Conjecture for semi-stable elliptic curves over $\mathbb{Q}$ as this case suffices to prove Fermat's Last Theorem. But the full conjecture in fact postulated that all elliptic curves over $\mathbb{Q}$ — not just the semi-stable ones — are modular. By 2001, work of Breuil, Conrad, Diamond, and Taylor had completed the general case so that we now know that all rational elliptic curves are modular; this result is sometimes called the Modularity Theorem, rather than the Shimura-Taniyama Conjecture. Even more recently, Robert Langlands has devised a series of conjectures which vastly generalise the Modularity Theorem. The so-called Langlands program thus seeks to correspond “automorphic forms” — natural generalisations of modular forms — to various arithmetic objects in algebraic geometry — such as elliptic curves. Already these conjectures guide the future of number theory, just as Shimura-Taniyama inspired twentieth century number theorists.

References

[DDT07] Henri Darmon, Fred Diamond, and Richard Taylor. Fermat's last theorem, September 2007.

[DS05] Fred Diamond and Jerry Shurman. *A First Course in Modular Forms*. Springer Science and Business Media, New York, 2005.

[Mil17] James S. Milne. Algebraic number theory, 2017. Available at `www.jmilne.org/math/`.

[Mil18] James S. Milne. Fields and galois theory, 2018. Available at `www.jmilne.org/math/`.

[Rid09] Larry Riddle. Sophie germain and fermat's last theorem, July 2009. Available at `https://www.agnesscott.edu/lriddle/women/women.htm`.

[Sch18] Berndt Schwerdtfeger. Modular fundamental domain, October 2018.

[Sil86] Joseph H. Silverman. *The Arithmetic of Elliptic Curves*. Springer Science and Business Media, New York, 1986.

[Sil94] Joseph H. Silverman. *Advanced Topics in the Arithmetic of Elliptic Curves*. Springer Science and Business Media, New York, 1994.

[Wes99] Tom Weston. The modular curves $X_0(11)$ and $X_0(11)$, 1999. Available at `https://people.math.umass.edu/~weston/ep.html`.