AN INTRODUCTION TO THE MODULARITY THEOREM

DESMOND SAUNDERS

Abstract. The aim of this paper is to state one of the most important results of modern number theory: the modularity theorem. In fact, we will state it three times, all following [1]. Our particular aim is to make this presentation accessible with minimal background knowledge. No prior knowledge of modular forms, elliptic curves, or algebraic curves is necessary.

Contents

1. Introduction

2025 marks the 30th anniversary of the publication of the remarkable paper "Modular elliptic curves and Fermat's Last Theorem" by Andrew Wiles, which finally proved Fermat's Last Theorem, over 300 years after it was conjectured. He did

Date: August 28, 2024.

so by proving a very different and much less accessible result: the modularity theorem, also known as the Taniyama–Shimura–Weil conjecture before it was proven. This result connects two structures of utmost importance to modern number theory: elliptic curves and modular forms. Our goal is to provide an accessible but rigorous introduction to the machinery necessary to state three different versions of the modularity theorem, with increasing structure and specificity:

Theorem (Modularity Theorem, Version I)**.** *For every complex elliptic curve E with $j(E) \in \mathbb{Q}$, there is some $N \in \mathbb{N}$ such that there exists a surjective holomorphic map from $X_0(N)$ to E.*

Theorem (Modularity Theorem, Version II)**.** *For every complex elliptic curve E with $j(E) \in \mathbb{Q}$, there is some $N \in \mathbb{N}$ such that there exists a surjective holomorphic homomorphism of complex tori from $J_0(N)$ to E.*

Theorem (Modularity Theorem, Version III)**.** *For every complex elliptic curve E with $j(E) \in \mathbb{Q}$, there is some $N \in \mathbb{N}$ and a newform $f \in \mathcal{S}_2(\Gamma_0(N))$ such that there exists a surjective holomorphic homomorphism of complex tori from A_f to E.*

As stated, the goal of this paper is to be accessible, but the theory of modular forms does use results from many areas of mathematics. This paper should be understandable to anyone who has taken undergraduate courses in analysis (including differential forms), complex analysis, linear algebra, and abstract algebra. We attempt to prove everything necessary for the statements of the modularity theorems, however in certain places where the proofs would become too tedious or require too much background from topology or Riemann surface theory we may just sketch them in brief. Around the ends of Section 3 and Section 4 we allow ourselves to make some statements without proof in order to give the geometric structures $J_0(N)$ and A_f a complex analytic structure and show the equivalence of our three versions of the modularity theorem.

1.1. **Basic Ideas and Notation.** Let $\mathcal{H} = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$ denote the complex upper half plane. Of fundamental interest to us is the action of $\mathrm{GL}_2(\mathbb{R})$ (or more often the subgroup $\mathrm{SL}_2(\mathbb{Z})$) on $\mathcal{H}$:

Definition 1.1. Let $\gamma = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in M_{2\times 2}(\mathbb{R})$ be nonsingular and $\tau \in \mathbb{C} \cup \{\infty\}$. Then

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

If $c = 0$ then ∞ is fixed and otherwise ∞ is sent to $\frac{a}{c}$ and $\frac{-d}{c}$ is sent to ∞. Since γ is nonsingular we don't have to worry about both numerator and denominator being equal to zero.

One can check by computation that this action respects matrix multiplication. Note that $-I$ acts as the identity. For this reason some authors choose to quotient out by the subgroup $\{\pm I\}$ and consider the action of $\mathrm{PSL}_2(\mathbb{Z})$, but we will find it more convenient to use $\mathrm{SL}_2(\mathbb{Z})$.

Proposition 1.2. $\mathrm{Im}(\gamma(\tau)) = \det(\gamma)\dfrac{\mathrm{Im}(\tau)}{|c\tau + d|^2}$.

Proof. A simple manipulation shows

$$\begin{aligned}\operatorname{Im}\left(\frac{a\tau+b}{c\tau+d}\right) &= \operatorname{Im}\left(\frac{(a\tau+b)\overline{(c\tau+d)}}{(c\tau+d)\overline{(c\tau+d)}}\right)\\ &= \frac{\operatorname{Im}(ac|\tau|^2+bc\overline{\tau}+ad\tau+bd)}{|c\tau+d|^2}\\ &= \frac{0+(ad-bc)\operatorname{Im}(\tau)+0}{|c\tau+d|^2}\\ &= \det(\gamma)\frac{\operatorname{Im}(\tau)}{|c\tau+d|^2}\end{aligned}$$

as desired. □

Corollary 1.3. *The action of* $\mathrm{SL}_2(\mathbb{Z})$ *on* $\mathbb{C}$ *sends* $\mathcal{H}$ *to itself.*

So we have established the action of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathcal{H}$. It will also be useful for us to note that the derivative of $\gamma \in M_{2\times 2}(\mathbb{R})$ is

$$\gamma'(\tau) = \frac{\det(\gamma)}{(c\tau+d)^2} \tag{1.4}$$

or for $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ we have $\gamma'(\tau) = \frac{1}{(c\tau+d)^2}$.

For some intuition for what this action looks like, consider the following result from algebra:

Proposition 1.5. $\mathrm{SL}_2(\mathbb{Z})$ *is generated by the matrices*

$$T = \begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix} \quad \textit{and} \quad S = \begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix}$$

These matrices correspond to the functions

$$\tau \mapsto \tau + 1 \quad \text{and} \quad \tau \mapsto -\frac{1}{\tau}$$

so the action of $\mathrm{SL}_2(\mathbb{Z})$ on τ either shifts τ to the left or right by 1, or reflects τ across the unit circle and then the imaginary axis, or any combination of these two.

We will also provide the basic definition of a Riemann surface, both for the unfamiliar reader and to introduce notation:

Definition 1.6. Let X be a Hausdorff topological space. A *complex atlas* is an open cover $\{U_i\}_{i\in I}$ of X where each open set has an associated homeomorphism (called a *chart*) $\varphi_i : U_i \to V_i$ where V_i is an open subset of $\mathbb{C}$. These charts must be *compatible*, i.e. if we let

$$V_{i,j} = \varphi_i(U_i \cap U_j), \quad V_{j,i} = \varphi_j(U_i \cap U_j), \quad \text{and} \quad \varphi_{j,i} : V_{i,j} \to V_{j,i} = \varphi_j \circ \varphi_i^{-1},$$

then $\varphi_{j,i}$ is biholomorphic for all $i, j \in I$. The map $\varphi_{j,i}$ is called the *transition map.*

A connected Hausdorff topological space X with a complex atlas is called a *Riemann surface.*

Local properties related to the complex numbers can easily be translated to Riemann surfaces via the charts. For example, whether a function is holomorphic at a point is perfectly well-defined on Riemann surfaces.

Proposition 1.7. *Let* X *and* Y *be compact Riemann surfaces and let* $f : X \to Y$ *be holomorphic. Then* f *is either constant or surjective.*

Proof. We know $f(X)$ is compact, and so closed. By the open mapping theorem of complex analysis, if f is non-constant then $f(X)$ is also open. Since Y is connected, Y is the only non-empty subset of itself that is both closed and open. □

Finally, since we will be using $*$ for the pullback, we will use the slightly unusual notation $V^\wedge$ to denote the dual of a vector space V.

2. Version (I)

2.1. **Complex Elliptic Curves.** Although our main focus is on developing the modular forms part of the modularity theorem, we start with a brief discussion of elliptic curves, which will also serve as motivation for some of the later ideas.

Definition 2.1. A *lattice* Λ in $\mathbb{C}$ is given by $\Lambda = \omega_1\mathbb{Z} \oplus \omega_2\mathbb{Z}$ where ω_1, ω_2 form a basis for $\mathbb{C}$ over $\mathbb{R}$. A *complex elliptic* curve or *complex torus* E is the quotient $E = \mathbb{C}/\Lambda$ for some lattice Λ.

Different choices of ω_1 and ω_2 can create the same lattice and so the same torus. For example, by possibly switching one for its negative, we can choose ω_1 and ω_2 such that $\frac{\omega_1}{\omega_2} \in \mathcal{H}$. Assuming this convention, we can obtain a more complete characterization:

Lemma 2.2. *Let $\Lambda = \omega_1\mathbb{Z} \oplus \omega_2\mathbb{Z}$ and $\Lambda' = \omega_1'\mathbb{Z} \oplus \omega_2'\mathbb{Z}$ be lattices. Then $\Lambda = \Lambda'$ if and only if there exists a $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ such that*

$$\begin{bmatrix}\omega_1' \\ \omega_2'\end{bmatrix} = \gamma \begin{bmatrix}\omega_1 \\ \omega_2\end{bmatrix}$$

Proof. First suppose that $\Lambda = \Lambda'$, so that there exist $a, b, c, d \in \mathbb{Z}$ such that $\omega_1' = a\omega_1 + b\omega_2$ and $\omega_2' = c\omega_1 + d\omega_2$. Let $\gamma = \left[\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right]$ so that

$$\begin{bmatrix}\omega_1' \\ \omega_2'\end{bmatrix} = \gamma \begin{bmatrix}\omega_1 \\ \omega_2\end{bmatrix}.$$

Since the situation is entirely symmetrical, we could have exchanged the roles of Λ and Λ', resulting in some matrix γ', also with integer entries. Since $\{\omega_1, \omega_2\}$ and $\{\omega_1', \omega_2'\}$ are both bases for $\mathbb{C}$ over $\mathbb{R}$, we conclude that $\gamma' = \gamma^{-1}$. Thus γ and γ^{-1} both have integer entries and in particular $\det(\gamma)$ and $\det(\gamma^{-1})$ are integers. Therefore $\det(\gamma) = \pm 1$. Using the normalizing convention from above, we can further conclude that $\det(\gamma) = 1$, and so indeed $\gamma \in \mathrm{SL}_2(\mathbb{Z})$.

For the other direction, suppose

$$\begin{bmatrix}\omega_1' \\ \omega_2'\end{bmatrix} = \gamma \begin{bmatrix}\omega_1 \\ \omega_2\end{bmatrix}.$$

Then ω_1' and ω_2' can be expressed as integer sums of ω_1 and ω_2, and so $\Lambda' \subset \Lambda$. Multiplying both sides by γ^{-1} then shows that $\Lambda \subset \Lambda'$. □

Such elliptic curves are Riemann surfaces which also inherit an abelian group structure from addition in $\mathbb{C}$. We would like to categorize when two such curves are equivalent as Riemann surfaces and groups, i.e. when there exists a holomorphic group isomorphism between them.

Proposition 2.3. *Let $\mathbb{C}/\Lambda$ and $\mathbb{C}/\Lambda'$ be two complex elliptic curves. There exists a holomorphic group isomorphism between them if and only if there exists some $m \in \mathbb{C}$ such that $m\Lambda = \Lambda'$.*

Proof. If such an m exists, then it is easy to show $\varphi : \mathbb{C}/\Lambda \to \mathbb{C}/\Lambda'$ sending $z + \Lambda$ to $mz + \Lambda'$ is a holomorphic group isomorphism.

Conversely, suppose such a φ exists. Let $\pi : \mathbb{C} \to \mathbb{C}/\Lambda$ and $\pi' : \mathbb{C} \to \mathbb{C}/\Lambda'$ be projection maps. The key step is to construct a holomorphic function $\tilde{\varphi} : \mathbb{C} \to \mathbb{C}$ such that the diagram

$$\begin{array}{ccc} \mathbb{C} & \xrightarrow{\tilde{\varphi}} & \mathbb{C} \\ \downarrow{\scriptstyle \pi} & & \downarrow{\scriptstyle \pi'} \\ \mathbb{C}/\Lambda & \xrightarrow{\varphi} & \mathbb{C}/\Lambda' \end{array}$$

commutes. This uses some basic topology, but it is not hard to see how such a map is constructed. Let $\Lambda = \omega_1\mathbb{Z} \oplus \omega_2\mathbb{Z}$ and $\Lambda' = \omega_1'\mathbb{Z} \oplus \omega_2'\mathbb{Z}$. Let z be in the parallelogram defined by ω_1 and ω_2 and suppose $\varphi(z + \Lambda) = w + \Lambda'$. Then we send $\tilde{\varphi}(z)$ to the coset representative for $w + \Lambda'$ which lies in the parallelogram defined by ω_1' and ω_2'. Passing into adjacent parallelograms, we know $\pi' \circ \tilde{\varphi} = \varphi \circ \pi$ and so $\tilde{\varphi}$ is defined up to a constant in Λ'. We can choose this constant inductively, passing from parallelogram to parallelogram, so that $\tilde{\varphi}$ is holomorphic on $\mathbb{C}$.

By this construction, for any $\lambda \in \Lambda$, we have $\tilde{\varphi}(\lambda + z) = \tilde{\varphi}(z) + \lambda'$ for some $\lambda' \in \Lambda'$. Thus $\tilde{\varphi}(\lambda + z) - \tilde{\varphi}(z)$ is a continuous function which maps into the discrete set Λ', so it must be constant. Differentiating both sides we get

$$\tilde{\varphi}'(\lambda + z) = \tilde{\varphi}'(z).$$

In other words, $\tilde{\varphi}'$ is Λ periodic, so bounded, so constant by Liouville's theorem.

Therefore $\tilde{\varphi}(z) = mz + b$ and $\varphi(z + \Lambda) = mz + b + \Lambda'$. Since φ is a homomorphism, we know $\varphi(0) = 0$, which implies $b + \Lambda' = 0 + \Lambda'$ and $m\Lambda + \Lambda' = \Lambda' \implies m\Lambda \subset \Lambda'$. Since φ is invertible, we equally conclude that $\frac{1}{m}\Lambda' \subset \Lambda$ and so indeed $m\Lambda = \Lambda'$. □

If $m\Lambda = \Lambda'$, we say that Λ and Λ' are *homothetic*, with the map sending z to mz being a homothety. Thus we have shown that holomorphic isomorphism of complex elliptic curves is equivalent to homothety of their respective lattices. To characterize elliptic curves modulo isomorphism it suffices to characterize lattices modulo homothety:

Theorem 2.4. *The set of lattices modulo homothety is in bijection with $\mathcal{H}$ modulo the action of $\mathrm{SL}_2(\mathbb{Z})$. The bijection is given by*

$$\mathrm{SL}_2(\mathbb{Z})\tau \mapsto [\tau\mathbb{Z} \oplus \mathbb{Z}].$$

The lattice $\tau\mathbb{Z} \oplus \mathbb{Z}$ is denoted Λ_τ.

Proof. The action of $\mathrm{SL}_2(\mathbb{Z})$ splits $\mathcal{H}$ into orbits, and we denote the set of these orbits $\mathrm{SL}_2(\mathbb{Z})\backslash\mathcal{H}$. Let $\tau \in \mathcal{H}$ and let $\gamma(\tau)$ be another point in the same orbit. Then $\gamma(\tau)$ maps to

$$\frac{a\tau + b}{c\tau + d}\mathbb{Z} \oplus \mathbb{Z}$$

which is homothetic to $(a\tau + b)\mathbb{Z} \oplus (c\tau + d)\mathbb{Z}$, and by Lemma 2.2 this lattice is equivalent to $\tau\mathbb{Z} \oplus \mathbb{Z}$. This shows the map is well-defined, and the argument for injectivity follows similarly. By Proposition 2.3, this also characterizes elliptic curves up to isomorphism. □

If $f : \mathcal{H} \to \mathbb{C}$ is a function such that $f(\gamma(\tau)) = f(\tau)$ for all $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ (i.e. f is $\mathrm{SL}_2(\mathbb{Z})$ invariant), then we can consider f as a function on equivalence classes of elliptic curves by composing f with the inverse map from Theorem 2.4.

2.2. **The Curve $\mathrm{SL}_2(\mathbb{Z})\backslash\mathcal{H}$.** We have shown that the set of equivalence classes of complex elliptic curves is parametrized by the orbit space

$$\mathrm{SL}_2(\mathbb{Z})\backslash\mathcal{H}.$$

In other words, $\mathrm{SL}_2(\mathbb{Z})\backslash\mathcal{H}$ is the moduli space for complex elliptic curves. This is the origin of the word modular in modular curves and modular forms. We now turn our attention to studying these quotients and related objects.

The first goal is to understand $\mathrm{SL}_2(\mathbb{Z})\backslash\mathcal{H}$ as a geometric object. A natural starting point is to try and find a subset of $\mathcal{H}$ that contains exactly one representative of each $\mathrm{SL}_2(\mathbb{Z})$ orbit. It turns out that we can find such a set that is closed and connected, up to a certain boundary identification.

Definition 2.5. Let Γ be a group acting on $\mathcal{H}$. A *fundamental domain* for a quotient $\Gamma\backslash\mathcal{H}$ is a closed set $\mathcal{D} \subset \mathcal{H}$ such that no two points in the interior of $\mathcal{D}$ are Γ-equivalent and every point in $\mathcal{H}$ is Γ-equivalent to some point in $\mathcal{D}$.[1]

Proposition 2.6. *Let*

$$\mathcal{D} = \left\{ \tau \in \mathcal{H} : |\operatorname{Re}(\tau)| \leq \frac{1}{2} \text{ and } |\tau| \geq 1 \right\}.$$

Then $\mathcal{D}$ is a fundamental domain of $\mathrm{SL}_2(\mathbb{Z})\backslash\mathcal{H}$. Points on the boundary are identified by reflection across the imaginary axis.

Proof. First we show that every $\tau \in \mathcal{H}$ is $\mathrm{SL}_2(\mathbb{Z})$-equivalent to some point in $\mathcal{D}$. Consider the lattice $\Lambda_\tau = \tau\mathbb{Z} \oplus \mathbb{Z}$. This lattice must have some point $c\tau + d$ of minimal absolute value. Clearly c and d are co-prime, otherwise we could scale down by their common factor. Therefore, by the Euclidean algorithm, there exist $a, b \in \mathbb{Z}$ such that $ad - bc = 1$, and in particular the matrix $\gamma = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ is in $\mathrm{SL}_2(\mathbb{Z})$. By Proposition 1.2 we know that

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

and by the minimality of $|c\tau + d|$, we know that this is less than or equal to $\operatorname{Im}(\alpha(\tau))$ for all $\alpha \in \mathrm{SL}_2(\mathbb{Z})$. Using T from Proposition 1.5 some integer $n \in \mathbb{Z}$ times, we can shift $\gamma(\tau)$ such that its real part is between $\pm\frac{1}{2}$. Let $\gamma' = T^n\gamma$ so that $\gamma'(\tau) = \gamma(\tau) + n$. Since we only changed the real part, the imaginary part is still maximal. Applying S we get

$$\operatorname{Im}(\gamma'(\tau)) \geq \operatorname{Im}(S\gamma'(\tau)) = \frac{\operatorname{Im}(\gamma'(\tau))}{|\gamma'(\tau)|^2}$$

and so $|\gamma'(\tau)| \geq 1$. Since $|\operatorname{Re}(\gamma'(\tau))| \leq \frac{1}{2}$ indeed we have $\gamma'(\tau) \in \mathcal{D}$.

To show the other direction, suppose that τ_1 and τ_2 are distinct points of $\mathcal{D}$ and there exists some $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ such that $\tau_1 = \gamma\tau_2$. We want to show that either $|\operatorname{Re}(\tau_1)| = \frac{1}{2}$ and $\tau_2 = \tau_1 \pm 1$ (the identification of the two half-lines of the boundary) or $|\tau_1| = |\tau_2| = 1$ and they have opposite real part (the identification of the two halves of the circular arc of the boundary). Let $\gamma = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$. Without loss

[1] Precise definitions of fundamental domain vary, but this definition gives the general idea.

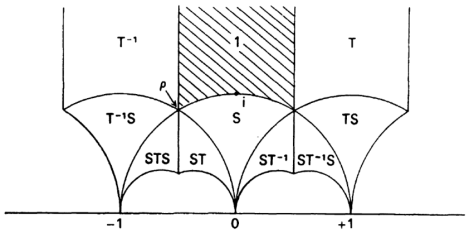


FIGURE 1. $\mathcal{D}$ with some $\mathrm{SL}_2(\mathbb{Z})$ translates. From [5].

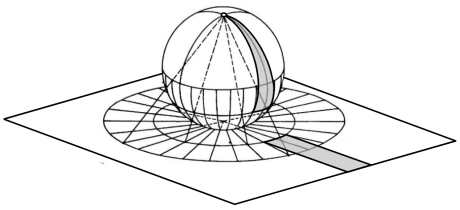


FIGURE 2. Stereographic projection of $\mathcal{D}$. From [4].

of generality, suppose $\mathrm{Im}(\tau_2) \geq \mathrm{Im}(\tau_1)$ and so applying Proposition 1.2 it must be that $|c\tau_1 + d|^2 \leq 1$. Since all points of $\mathcal{D}$ have imaginary part greater than $\frac{\sqrt{3}}{2}$,

$$\frac{\sqrt{3}}{2}|c| \leq |c|\,\mathrm{Im}(\tau_1) = \mathrm{Im}(c\tau_1) = \mathrm{Im}(c\tau_1 + d) \leq |c\tau_1 + d| \leq 1$$

and so $|c| \leq \frac{2}{\sqrt{3}}$, and since $c \in \mathbb{Z}$, we know $|c| \in \{0, 1\}$. If $c = 0$ then in order for $\det(\gamma) = 1$ to hold, we know that $a = d = \pm 1$ so $\gamma = \pm\left[\begin{smallmatrix} 1 & b \\ 0 & 1 \end{smallmatrix}\right]$ and $\tau_2 = \gamma(\tau_1) = \tau_1 + b$. It must be that $b = \pm 1$ and we are in the first case.

Now suppose that $c = 1$. Then we can show a similar constraint on $|d|$. We have

$$1 \geq |\tau_1 + d|^2 = |(\mathrm{Re}(\tau_1) + d) + \mathrm{Im}(\tau_1)i|^2 = (\mathrm{Re}(\tau_1) + d)^2 + \mathrm{Im}(\tau_1)^2$$

which implies

$$(\mathrm{Re}(\tau_1) + d)^2 \leq 1 - \mathrm{Im}(\tau_1)^2 \leq 1 - \frac{3}{4} \implies |\,\mathrm{Re}(\tau_1) + d| \leq \frac{1}{2}$$

and so $|d| \leq 1$. The same argument also works if $c = -1$.

If $|d| = 1$, then the only way for the above inequalities to be satisfied is if they are equalities, i.e. $|\,\mathrm{Re}(\tau_1)| = \frac{1}{2}$ and $\mathrm{Im}(\tau_1) = \frac{\sqrt{3}}{2}$, so τ_1 is the point ρ or $\rho + 1$ in Figure 1. Thus

$$\frac{\sqrt{3}}{2} \leq \mathrm{Im}(\tau_2) = \frac{\mathrm{Im}(\tau_1)}{|\pm\tau_1 \pm 1|^2} \leq \frac{\sqrt{3}}{2}$$

so τ_2 must be the only other point of $\mathcal{D}$ with this minimal imaginary part, namely $\tau_1 \pm 1$.

Finally if $|c| = 1$ and $d = 0$, then the condition $|c\tau_1 + d| \leq 1$ becomes $|\tau_1| \leq 1$ and indeed $|\tau_1| = 1$. By our formula from Proposition 1.2, $\mathrm{Im}(\tau_1) = \mathrm{Im}(\tau_2)$. Therefore we could have taken $\mathrm{Im}(\tau_2) \leq \mathrm{Im}(\tau_1)$, flipped all of these calculations, and shown that either one of the above cases holds or $|\tau_2| = 1$ as well. Then τ_1 and τ_2 have magnitude 1 and the same imaginary part, so they have opposite real parts. □

Acting on $\mathcal{D}$ by any $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ gives an equally valid fundamental domain (although it may not be closed), some of which are pictured in Figure 1. While it is nice that $\mathcal{D}$ is closed and connected, we would really like it to be compact. Indeed, the stereographic projection of $\mathcal{D}$ in Figure 2 suggests that, under the right topology, $\mathcal{D}$ could be compactified by adding a single point at infinity. To make this precise, note that the $\mathrm{SL}_2(\mathbb{Z})$ orbit of ∞ is $\mathbb{Q} \cup \{\infty\}$. So let $\mathcal{H}^* = \mathcal{H} \cup \mathbb{Q} \cup \{\infty\}$, which is sent to itself by the action of $\mathrm{SL}_2(\mathbb{Z})$, and let $\mathcal{D}^* = \mathcal{D} \cup \{\infty\}$ be the

fundamental domain for $\mathrm{SL}_2(\mathbb{Z})\backslash\mathcal{H}^*$. To show that $\mathcal{D}^*$ is compact, we need to define an appropriate topology for $\mathcal{H}^*$:

Definition 2.7. The topology on $\mathcal{H}^*$ is is the smallest topology which contains:

(1) The standard Euclidean topology on $\mathcal{H}$,
(2) The sets
$$\mathcal{N}_M = \{\tau \in \mathcal{H} : \mathrm{Im}(\tau) > M\} \cup \{\infty\}$$
for all $M \in \mathbb{R}^+$,
(3) All $\mathrm{SL}_2(\mathbb{Z})$ images of these sets $\mathcal{N}_M$ (which are either sets of the same form or circles tangent to the real axis containing a single rational number).

Proposition 2.8. *$\mathcal{D}^*$ is compact under the above topology.*

Proof. Let $\mathscr{G}$ be an arbitrary open cover of $\mathcal{D}^*$. Then $\mathscr{G}$ must contain a neighborhood of ∞. The only open sets which contain ∞ are of the form $\mathcal{N}_M \cup G$ where $M \in \mathbb{R}^+$ and G is an open set in the Euclidean topology (plus maybe some rational points). The points in $\mathcal{D}$ not covered by this open set are some closed subset of $\{\tau \in \mathcal{D} : \mathrm{Im}(\tau) \leq M\}$, which is compact under the Euclidean topology and so compact under our topology as well. Thus adding $\mathcal{N}_M \cup G$ to some finite subcover of this set, we get a finite subcover for all of $\mathcal{D}$. □

2.3. **Congruence Subgroups and Modular Curves.** We do not want to only study the curve $\mathrm{SL}_2(\mathbb{Z})\backslash\mathcal{H}$, but a family of related curves. In particular, by quotienting out by smaller groups or subgroups of $\mathrm{SL}_2(\mathbb{Z})$, we get a curve which is in some sense 'larger' and can encode additional information. Indeed, although we will not show it, the curves we will define below encode information such as an elliptic curve plus a particular cyclic subgroup or a point of interest.

Definition 2.9. The *principle congruence subgroup of level N* is the subgroup
$$\Gamma(N) = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2(\mathbb{Z}) : a \equiv d \equiv 1 \text{ and } b \equiv c \equiv 0 \pmod{N} \right\}.$$
and we define $\Gamma(1) = \mathrm{SL}_2(\mathbb{Z})$.

Proposition 2.10. *$\Gamma(N)$ is normal and of finite index in $\mathrm{SL}_2(\mathbb{Z})$.*

Proof. Consider the natural homomorphism from $\mathrm{SL}_2(\mathbb{Z})$ to $\mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$ which reduces each entry mod N. Then $\Gamma(N)$ is the kernel of the homomorphism, so it is normal. Furthermore, $\mathrm{SL}_2(\mathbb{Z})/\Gamma(N)$ is isomorphic to the image of this homomorphism, some subgroup of $\mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$.[2] Since $\mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$ is finite (it has at most N^4 elements), so is this subgroup. Thus indeed $[\mathrm{SL}_2(\mathbb{Z}) : \Gamma(N)]$ is finite. □

Definition 2.11. A *congruence subgroup of level N* is a subgroup $\Gamma \leq \mathrm{SL}_2(\mathbb{Z})$ such that $\Gamma(N) \subset \Gamma$. In particular let
$$\Gamma_0(N) = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2(\mathbb{Z}) : c \equiv 0 \pmod{N} \right\} \quad \text{and}$$
$$\Gamma_1(N) = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2(\mathbb{Z}) : a \equiv d \equiv 1 \text{ and } c \equiv 0 \pmod{N} \right\}.$$

[2] In fact, the homomorphism is surjective, so this quotient is isomorphic to all of $\mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$.

By Proposition 2.10, we immediately get that all congruence subgroups have finite index in $\mathrm{SL}_2(\mathbb{Z})$. It is not hard to describe the fundamental domains of these congruence subgroups if we allow them to be disconnected, although the boundary identification becomes more convoluted and so the picture less helpful:

Proposition 2.12. *Let Γ be a congruence subgroup and let $\{\alpha_j\}_{j=1}^d$ be coset representatives for Γ in $\mathrm{SL}_2(\mathbb{Z})$. Then a fundamental domain for Γ is*

$$\bigcup_{j=1}^{d} \alpha_j(\mathcal{D}).$$

Definition 2.13. Let Γ be a congruence subgroup. Then

$$Y(\Gamma) = \Gamma\backslash\mathcal{H} \quad \text{and} \quad X(\Gamma) = \Gamma\backslash\mathcal{H}^*$$

are called *modular curves*. The points of $X(\Gamma)$ not in $Y(\Gamma)$ (i.e. the Γ orbits in $\mathbb{Q} \cup \{\infty\}$) are called the *cusps* of $X(\Gamma)$.

We will write $Y(N)$ or $X(N)$ to mean $Y(\Gamma(N))$ or $X(\Gamma(N))$, and similarly for $Y_0(N), X_0(N), Y_1(N)$, and $X_1(N)$. We want to consider these curves as topological spaces in their own right (and ultimately as Riemann surfaces). The natural topology to use is the quotient topology: if π is the projection map from $\mathcal{H}$ to $Y(\Gamma)$, then $U \subset Y(\Gamma)$ is open if $\pi^{-1}(U)$ is open in $\mathcal{H}$, and similarly with $\mathcal{H}^*$ and $X(\Gamma)$.

Proposition 2.14. *For any congruence subgroup Γ, $Y(\Gamma)$ and $X(\Gamma)$ are Hausdorff and connected.*

Proof. It is fairly easy to see from the description of the fundamental domain that $Y(1)$ and $X(1)$ are Hausdorff (Figure 3 may be helpful to see this). Showing the more general case requires a proof not dissimilar from that of Proposition 2.6, which we omit.

The fact that these curves are connected comes immediately from the fact that $\mathcal{H}$ and $\mathcal{H}^*$ are connected in their respective topologies, and the projection maps are continuous by definition of the quotient topology. □

As with the full modular group, adding the cusps is sufficient to make $X(\Gamma)$ compact:

Proposition 2.15. *For any congruence subgroup Γ, $X(\Gamma)$ is compact.*

Proof. Let π be the projection map from $\mathcal{H}^*$ to $X(\Gamma)$. Let $\{\alpha_j\}_{j=1}^d$ be the coset representatives for Γ in $\mathrm{SL}_2(\mathbb{Z})$, which there are finitely many of since Γ has finite index. By Proposition 2.12 we have

$$X(\Gamma) = \pi\left(\bigcup_{j=1}^{d} \alpha_j(\mathcal{D})\right) = \bigcup_{j=1}^{d} \pi(\alpha_j(\mathcal{D})).$$

The maps α_j are continuous from $\mathcal{H}^*$ to $\mathcal{H}^*$, and by the definition of the quotient topology π is continuous to $X(\Gamma)$. The continuous image of a compact set is compact, so we have written $X(\Gamma)$ as a finite union of compact sets. □

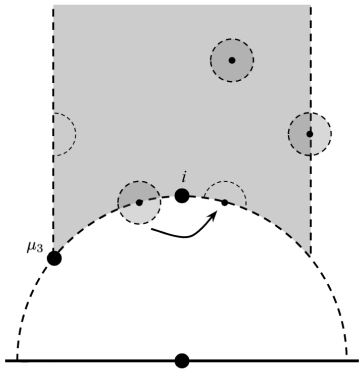


FIGURE 3. $\mathcal{D}$ with some charts and elliptic points.

2.4. **Modular Curves as Riemann Surfaces.** To show that these modular curves are Riemann surfaces, we need to find the charts as in Definition 1.6. In this section we will give formulas and some motivation for these charts but will not prove their compatibility or that they are in fact homeomorphisms. Understanding these formulas is not necessary for our purposes, although they will be referenced later.

Figure 3 shows that around most points of $X(1)$, the projection map restricted to a small enough open neighborhood is injective, so a local inverse can serve as the necessary chart. See Proposition 2.17. However, this injectivity fails at i and μ_3. This motivates the following definition:

Definition 2.16. Let Γ be a congruence subgroup. For any $\tau \in \mathcal{H}$, let Γ_τ denote the stabilizer of τ in Γ. Then τ is an *elliptic point* for Γ if Γ_τ is non-trivial, i.e.

$$\Gamma_\tau \not\subset \{\pm I\}.$$

The *period* of τ is the order of this stabilizer, up to the equivalent actions of $\pm I$:

$$h_\tau = |\{\pm I\}\Gamma_\tau / \{\pm I\}|.$$

Proposition 2.17. *Let τ be a non-elliptic point of Γ. Then for some open neighborhood U of τ the restriction of the projection map $\pi|_U$ is a homeomorphism. The local inverse is the necessary chart around τ.*[3]

Proposition 2.18. *Let τ be an elliptic point for Γ. Let $\rho_\tau(z) = z^{h_\tau}$ and let*

$$\delta_\tau = \begin{bmatrix} 1 & -\tau \\ 1 & -\overline{\tau} \end{bmatrix} \in GL_2(\mathbb{Z}).$$

For a sufficiently small open neighborhood U of τ, let $\psi : U \to V = (\rho_\tau \circ \delta_\tau)|_U$. Then the local chart $\varphi : \pi(U) \to V$ is defined by $\varphi \circ \pi = \psi$.

Proof (idea). The idea is that ψ mimics the identification of π around τ, but going from $\mathbb{C}$ to $\mathbb{C}$. In particular, δ_τ takes τ to 0 and does some straightening, then ρ_τ wraps around 0 in the same manner as π. See Figure 4. □

[3]Sometimes it is convenient to have the local coordinates centered at 0, in which case we can compose with δ_τ from Proposition 2.18.

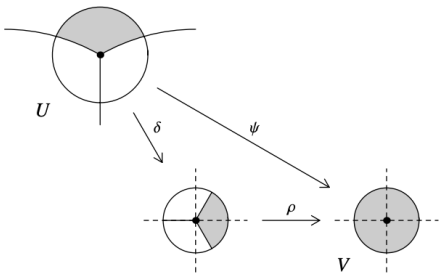


FIGURE 4.
Elliptic point charts. From [1].

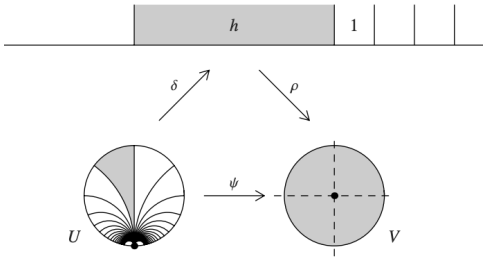


FIGURE 5.
Cusp charts. From [1].

Finally, for $X(\Gamma)$ we need to put charts around the cusps. The process is similar to the elliptic points.

Definition 2.19. Let $s \in \mathbb{Q} \cup \{\infty\}$ and let $\delta_s \in \mathrm{SL}_2(\mathbb{Z})$ take s to ∞. Then the *width of* s is

$$h_s = |\mathrm{SL}_2(\mathbb{Z})_\infty / (\{\pm I\}\delta\Gamma\delta^{-1})_\infty|$$

Since δ_s takes s to ∞ not 0, we use the exponential map rather than the power map to take it to a neighborhood of 0:

Proposition 2.20. *Let $s \in \mathbb{Q} \cup \{\infty\}$ and let $\rho_s = e^{2\pi i/h_s}$. For a sufficiently small neighborhood U of s, let $\psi : U \to V = (\rho_s \circ \delta_s)|_U$. Then the local chart $\varphi : \pi(U) \to V$ is defined by $\varphi \circ \pi = \psi$.*

Proof (idea). See Figure 5. □

2.5. **Modular Forms, Automorphic Forms, and Cusp Forms.** The next natural objects to study are functions on these modular curves. Since modular curves are Riemann surfaces, we are particularly interested in holomorphic functions, but these do not turn out to be interesting objects to study:

Proposition 2.21. *Let Γ be a congruence subgroup and let $f : X(\Gamma) \to \mathbb{C}$ be holomorphic. Then f is constant.*

Proof (sketch). Since $X(\Gamma)$ is compact, f must be bounded. It is not hard to show that if f is holomorphic on $X(\Gamma)$, then $f \circ \pi$ is holomorphic on $\mathcal{H}$. By Liouville's theorem we get that $f \circ \pi$ is constant, and therefore so is f. □

In other words, functions on $\mathcal{H}$ that are invariant under Γ and sufficiently well-behaved as they approach the cusps are overly constrained. We will choose to loosen the invariance constraint by introducing a new operator that does slightly more than just composition:

Definition 2.22. Let $k \in \mathbb{Z}$. Let $f : \mathcal{H} \to \mathbb{C}$ and $\gamma = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \mathrm{GL}_2^+(\mathbb{Q})$. Then $f[\gamma]_k : \mathcal{H} \to \mathbb{C}$ is defined by

$$f[\gamma]_k(\tau) = \det(\gamma)^{k/2}(c\tau + d)^{-k} f(\gamma(\tau))$$

Definition 2.23. Given a set of matrices Γ, a meromorphic function $f : \mathcal{H} \to \mathbb{C}$ is *weakly modular of weight k with respect to* Γ if

$$f[\gamma]_k = f \quad \forall \gamma \in \Gamma.$$

If $\alpha, \beta \in \mathrm{GL}_2^+(\mathbb{Q})$ then $[\alpha]_k[\beta]_k = [\alpha\beta]_k$ as operators. On $\mathcal{H}$, the function $c\tau + d$ has no poles or zeroes, so if f is holomorphic so too is $f[\gamma]_k$, and similarly for meromorphic. Since we will almost always be dealing with $\gamma \in \mathrm{SL}_2(\mathbb{Z})$, the factor of $\det(\gamma)$ can usually be ignored.[4]

By way of motivation, for $k = 2$ we have $f[\gamma]_k = \gamma'(f \circ \gamma)$ by (1.4), a familiar formulation from the chain rule (Theorem 3.2 will make this connection precise). Taking products of weight 2 weakly modular functions, we get weakly modular functions of all even weights.

With this piece of notation, we can also say what it means for a function to be sufficiently well-behaved as it approaches ∞. Note that the function $e^{2\pi i\tau/h}$ takes $\mathcal{H}$ to the open unit disk minus 0, which we denote D'.

Definition 2.24. Let $\gamma = \left[\begin{smallmatrix} 1 & h \\ 0 & 1 \end{smallmatrix}\right] \in \mathrm{SL}_2(\mathbb{Z})$ and suppose $f : \mathcal{H} \to \mathbb{C}$ is a function such that $f = f[\gamma]_k$ for some $k \in \mathbb{Z}$. Then $f(\tau) = f(\tau + h)$ so (even though the complex logarithm is only defined up to $2\pi i\mathbb{Z}$) let $g : D' \to \mathbb{C}$ be given by

$$g(q) = f\left(\log(q)\frac{h}{2\pi i}\right) \implies f(\tau) = g(q_h) \text{ where } q_h = e^{2\pi i\tau/h}.$$

We say f is *holomorphic at* ∞ if g has a holomorphic extension to the point 0. Similarly for meromorphic.

If we know a function f is holomorphic on $\mathcal{H}$ (and so g is holomorphic on D'), to show that f is holomorphic at ∞ it suffices to show that $g(q)$ is bounded as $q \to 0$, or in other words $f(\tau)$ is bounded as $\mathrm{Im}(\tau) \to \infty$.

Of course, we want to be holomorphic at all the cusps, not just ∞. For all $s \in \mathbb{Q}$ there is some $\alpha \in \mathrm{SL}_2(\mathbb{Z})$ that takes ∞ to s, so rather than make a new definition we use the operator $[\alpha]_k$. Thus we can finally define modular forms:

Definition 2.25. Let Γ be a congruence subgroup and let $k \in \mathbb{Z}$. Then $f : \mathcal{H} \to \mathbb{C}$ is a *modular form of weight k with respect to* Γ if

(i) f is weakly modular of weight k with respect to Γ, and
(ii) f is holomorphic on $\mathcal{H}$ and $f[\alpha]_k$ is holomorphic at ∞ for all $\alpha \in \mathrm{SL}_2(\mathbb{Z})$.

The set of such forms is denoted $\mathcal{M}_k(\Gamma)$.

Note that the matrix

$$T^N = \begin{bmatrix} 1 & N \\ 0 & 1 \end{bmatrix}$$

is in any congruence subgroup of level N. Furthermore, it is easy to check that $f[\alpha]_k$ is weakly modular of weight k with respect to $\alpha^{-1}\Gamma\alpha$, which also contains T^N (since $\Gamma(N)$ is normal). Thus condition (i) ensures the requirements are met for condition (ii) to be well-defined.

Although modular forms will be our main focus, we can easily replace holomorphy with meromorphy to get:

[4]Sometimes it can be more convenient to raise the determinant to the power of $k - 1$, but for $\mathrm{SL}_2(\mathbb{Z})$ or $k = 2$ it does not matter.

Definition 2.26. Let Γ be a congruence subgroup and let $k \in \mathbb{Z}$. Then $f : \mathcal{H} \to \mathbb{C}$ is a *automorphic form of weight k with respect to* Γ if

(i) f is weakly modular of weight k with respect to Γ, and
(ii) f is meromorphic on $\mathcal{H}$ and $f[\alpha]_k$ is meromorphic at ∞ for all $\alpha \in \mathrm{SL}_2(\mathbb{Z})$.

The set of such forms is denoted $\mathcal{A}_k(\Gamma)$.

If $f(\tau)$ is a modular form, then $g(q_h)$ has a holomorphic continuation to 0, so we can write a Taylor series expansion of g around $q_h = 0$:

$$g(q_h) = \sum_{n=0}^{\infty} a_n q_h^n.$$

Recalling that $q_h = e^{2\pi i\tau/h}$, we get

$$f(\tau) = \sum_{n=0}^{\infty} a_n e^{2\pi n i\tau/h}$$

which is called the *Fourier development of f*. Having a Fourier development and being holomorphic at ∞ are equivalent.

Definition 2.27. Let $f \in \mathcal{M}_k(\Gamma)$. Then we say f is a *cusp form* if it goes to zero at the cusps, i.e. $a_0 = 0$ in the Fourier development of $f[\alpha]_k$ for all $\alpha \in \mathrm{SL}_2(\mathbb{Z})$.

The set of cusp forms of weight k with respect to Γ is denoted $\mathcal{S}_k(\Gamma)$.

The sets $\mathcal{M}_k(\Gamma), \mathcal{A}_k(\Gamma)$, and $\mathcal{S}_k(\Gamma)$ are all vector spaces over $\mathbb{C}$. A crucial feature of the theory of modular forms (which we will not discuss very much) is that these spaces are all finite dimensional.

We add one final observation about the $[\gamma]_k$ operator which will be useful later:

Proposition 2.28. *Let Γ_1 and Γ_2 be congruence subgroups such that $\gamma\Gamma_1\gamma^{-1} \subset \Gamma_2$ for some $\gamma \in \mathrm{GL}_2^+(\mathbb{Q})$. Then $[\gamma]_k$ takes $\mathcal{M}_k(\Gamma_2)$ to $\mathcal{M}_k(\Gamma_1)$ and $\mathcal{S}_k(\Gamma_2)$ to $\mathcal{S}_k(\Gamma_1)$.*

Proof. Let $f \in \mathcal{M}_k(\Gamma_2)$. As we noted above, it is fairly straightforward to show that $f[\gamma]_k$ is weakly modular of weight k with respect to $\gamma^{-1}\Gamma_2\gamma$, which contains Γ_1. Thus we focus on showing that condition (ii) from Definition 2.25 holds for $f[\gamma]_k$. Given $\alpha \in \mathrm{SL}_2(\mathbb{Z})$, $\gamma\alpha$ is in $\mathrm{GL}_2^+(\mathbb{Q})$, so to show that $f[\gamma]_k[\alpha]_k = f[\gamma\alpha]_k$ is holomorphic at ∞, it suffices to show that $f[\gamma]_k$ is holomorphic at ∞ for a generic $\gamma \in \mathrm{GL}_2^+(\mathbb{Q})$.

Let $\gamma = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$. We would like to find a matrix $\alpha \in \mathrm{SL}_2(\mathbb{Z})$ such that $\alpha\gamma$ has lower left entry 0. If $c = 0$ then $\alpha = I$. Otherwise let $\frac{p}{q}$ be $\frac{a}{c}$ written in lowest terms, such that p and q are coprime. Then there exist $s, t \in \mathbb{Z}$ such that $ps - qt = 1$. Let

$$\alpha = \begin{bmatrix} -s & t \\ q & -p \end{bmatrix} \in \mathrm{SL}_2(\mathbb{Z}).$$

Then computation shows that $\alpha\gamma$ has lower entry 0. Let

$$\gamma' = \alpha\gamma = r \begin{bmatrix} a' & b' \\ 0 & c' \end{bmatrix}$$

with $r \in \mathbb{Q}^+$ and $a', b', d' \in \mathbb{Z}$ with g.c.d. 1. Changing α for $\pm\alpha$ we can ensure $a', d' > 0$. Then $f[\gamma]_k = (f[\alpha^{-1}]_k)[\gamma']_k$ and since f is a modular form we know that $f[\alpha^{-1}]_k$ has a Fourier development:

$$(f[\alpha^{-1}]_k)(\tau) = \sum_{n=0}^{\infty} a_n e^{2\pi i n\tau/h}$$

for some period h. Thus

$$(f[\gamma]_k)(\tau) = (f[\alpha^{-1}]_k)[\gamma']_k(\tau) = \frac{\det(\gamma')^{k-1}}{(rd')^k}\sum_{n=0}^{\infty} a_n e^{2\pi i n(\frac{a'}{d'}\tau+\frac{b'}{d'})/h}$$
$$\propto \sum_{n=0}^{\infty}\left(a_n e^{2\pi i n\frac{b'}{d'h}}\right)e^{2\pi i(na')\tau/(d'h)}$$

(where the constant multiplier was dropped for simplicity). This is the necessary Fourier development of $f[\gamma]_k$ with period $d'h$, so indeed $f[\gamma]_k$ is holomorphic at ∞.

This completes the proof that $[\gamma]_k$ takes $\mathcal{M}_k(\Gamma_2)$ to $\mathcal{M}_k(\Gamma_1)$. To see that cusp forms get taken to cusp forms, note that in the proof if $a_0 = 0$ in the initial Fourier development, that carries through to the Fourier development of $f[\gamma]_k$. □

We now provide some examples of modular forms. The zero function is a modular form of every weight with respect to any Γ. Constant functions are the only modular forms of weight 0, also with respect to any Γ. The following more interesting examples also happen to be necessary to defining the function j that appears in the statement of the modularity theorem.

Definition 2.29. Let $k \geq 4$ be even, and let Λ be a lattice. Then

$$G_k(\Lambda) = \sum_{\omega\in\Lambda}{}' \frac{1}{\omega^k}.$$

where primed summation means without the point $(0, 0)$. In particular, the *Eisenstein series of weight k* is the function

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

Proposition 2.30. *$G_k(\tau)$ converges on all of $\mathcal{H}$ and is a modular form of weight k with respect to $\mathrm{SL}_2(\mathbb{Z})$.*

Proof. Let $\gamma = \left[\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right] \in \mathrm{SL}_2(\mathbb{Z})$. We want to show that $G_k[\gamma]_k = G_k$, i.e.

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

Examining the proof of Theorem 2.4 we see that $\Lambda_{\gamma(\tau)} = (c\tau+d)^{-1}\Lambda_\tau$. Thus

$$G_k(\gamma(\tau)) = G_k(\Lambda_{\gamma(\tau)}) = \sum_{\omega\in(c\tau+d)^{-1}\Lambda_\tau}{}' \frac{1}{\omega^k} = \sum_{\omega\in\Lambda_\tau}{}' \left(\frac{c\tau+d}{\omega}\right)^k = (c\tau+d)^k G_k(\tau)$$

as desired. Now we will show that $G_k(\tau)$ converges absolutely and is bounded as $\mathrm{Im}(\tau) \to \infty$, and so is holomorphic at ∞. Let

$$D = \left\{\tau \in \mathcal{H} : |\,\mathrm{Re}(\tau)| \leq \frac{1}{2} \text{ and } \mathrm{Im}(\tau) \geq \frac{\sqrt{3}}{2}\right\}$$

so that $\mathcal{D} \subset D$. Since we know G_k is weakly modular, showing convergence on $\mathcal{D}$ shows convergence on all of $\mathcal{H}$. Furthermore, if we can show that G_k is bounded on D then $G_k(\tau) = (G_k[T]_k)(\tau) = G_k(\tau+1)$ shows that it is bounded for all $\mathrm{Im}(\tau) \geq \frac{\sqrt{3}}{2}$.

Our first step is to show that $|\tau+\delta| \geq \frac{1}{3}\sup\{1,|\delta|\}$ for all $\tau \in D$ and $\delta \in \mathbb{R}$. If $|\delta| < \frac{3}{2}$ then

$$|\tau+\delta| \geq \operatorname{Im}(\tau) > \frac{1}{2} \geq \frac{1}{3}\sup\{1,|\delta|\}.$$

If $|\delta| \geq \frac{3}{2}$ then

$$|\tau+\delta| \geq |\operatorname{Re}(\tau)+\delta| \geq |\delta| - \frac{1}{2} \geq \frac{2}{3}|\delta| \geq \frac{1}{3}\sup\{1,|\delta|\}.$$

By taking partial sums over expanding squares of radius n, we also find that the sum

$$\sum_{(c,d)\in\mathbb{Z}^2}{}' \frac{1}{\sup\{|c|,|d|\}^k} = \sum_{n=1}^{\infty}(4n+4)\frac{1}{n^k}$$

converges absolutely for $k \geq 4$. Finally, the Riemann zeta function $\zeta(k) = \sum_{n=1}^{\infty}\frac{1}{n^k}$ converges for $k \geq 2$. Thus for $\tau \in D$ we have

$$\begin{aligned}
\sum_{(c,d)\in\mathbb{Z}^2}{}' \frac{1}{|c\tau+d|^k} &= 2\zeta(k) + \sum_{c\neq 0, d\in\mathbb{Z}} \frac{1}{(|c||\tau+\frac{d}{c}|)^k} \\
&\leq 2\zeta(k) + \sum_{c\neq 0, d\in\mathbb{Z}} \frac{1}{(|c|\frac{1}{3}\sup\{1,\frac{|d|}{|c|}\})^k} \\
&= 2\zeta(k) + 3^k \sum_{c\neq 0, d\in\mathbb{Z}} \frac{1}{\sup\{|c|,|d|\}^k} \\
&\leq 2\zeta(k) + 3^k \sum_{(c,d)\in\mathbb{Z}^2}{}' \frac{1}{\sup\{|c|,|d|\}^k}
\end{aligned}$$

which we established converges absolutely. Furthermore, this bound is independent of τ, showing that $G_k(\tau)$ is bounded on D and so holomorphic at ∞ as desired.[5] $\square$

Of particular importance are the functions

$$g_2(\tau) = 60G_4(\tau) \quad \text{and} \quad g_3(\tau) = 140G_6(\tau).$$

since they give the connection between complex tori and curves defined by equations of the form $y^2 = 4x^3 + ax + b$, which the more recognizable definition of elliptic curves. In particular, a complex torus $\mathbb{C}/\Lambda$ is isomorphic as a group and a Riemann surface to the curve

$$y^2 = 4x^3 + g_2(\Lambda)x + g_3(\Lambda).$$

We will not prove this, but it also motivates looking at the discriminant function,

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

which is a modular form of weight 12 with respect to $\mathrm{SL}_2(\mathbb{Z})$ (in fact, it is a cusp form). Comparing powers also shows:

[5]The fact that G_k is holomorphic on $\mathcal{H}$ follows quickly from the fact that it converges absolutely on $\mathcal{H}$, although the proof does have to be slightly modified.

Corollary 2.31. *The* modular function $j : \mathcal{H} \to \mathbb{C}$ *given by*

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

is $\mathrm{SL}_2(\mathbb{Z})$ *invariant.*

Since j is not constant, by Proposition 2.21 we know that j cannot be a weight 0 modular form. It does happen to be a weight 0 automorphic form, however the above corollary is sufficient for defining $j(E)$ for any elliptic curve E, which is what we need to state the modularity theorem.

2.6. **The Modularity Theorem.**

Theorem 2.32 (Modularity Theorem, Version I)**.** *For every complex elliptic curve E with $j(E) \in \mathbb{Q}$, there is some $N \in \mathbb{N}$ such that there exists a surjective holomorphic map from $X_0(N)$ to E.*

3. Version (II)

3.1. **Holomorphic and Meromorphic Differentials.**

Riemann surfaces are 1-dimensional complex manifolds, so we should be able to write differential 1-forms on them.

Definition 3.1. Let X be a Riemann surface with charts $\varphi_i : U_i \to V_i$ for i in some index set I. A *meromorphic differential on* X (of degree 1) is a collection of differential 1-forms $\omega_i = f_i(z)dz$ on each V_i, with f_i meromorphic, such that these forms are *compatible*, i.e.

$$\varphi_{j,i}^*(\omega_j|_{V_{j,i}}) = \omega_i|_{V_{i,j}}$$

where the asterisk denotes the pullback and the other notation comes from Definition 1.6.

The set of such differential forms, a vector space over $\mathbb{C}$, is denoted $\Omega^1(X)$. We can also require the functions f_i to be holomorphic, in which case we get the vector space $\Omega^1_{\mathrm{hol}}(X)$.

Essentially differential forms are still of the form $f(z)dz$, just defined locally with the coordinate maps. It turns out that these differentials on the modular curves $X(\Gamma)$ provide a good way of studying modular forms, due to the following connection:

Theorem 3.2. *Let Γ be a congruence subgroup of $\mathrm{SL}_2(\mathbb{Z})$. Then $\mathcal{A}_2(\Gamma)$ and $\Omega^1(X(\Gamma))$ are isomorphic as complex vector spaces.*

Proof (sketch). Given any two Riemann surfaces X and Y and a holomorphic map $h : X \to Y$ there is a corresponding pullback map $h^* : \Omega^1(Y) \to \Omega^1(X)$ which suitably defines $h_i : \mathbb{C} \to \mathbb{C}$ for each coordinate patch V_i and then sends $f_i(z)dz$ to $f_i(h_i(z))h_i'(z)dz$. In particular we can consider the pullback of the projection map $\pi : \mathcal{H} \to X(\Gamma)$ which sends each $\omega \in \Omega^1(X(\Gamma))$ to a differential $f(\tau)d\tau$ on $\mathcal{H}$, omitting the cusps. Here we will show that for all $\omega \in \Omega^1(X(\Gamma))$, this $f(\tau)$ is weakly modular of weight 2, and then construct (but not justify) an explicit isomorphism

$$\omega : \mathcal{A}_2(\Gamma) \to \Omega^1(X(\Gamma))$$

such that $\pi^*(\omega(f)) = f(\tau)d\tau$.[6]

Let $\gamma \in \Gamma$ and let $\omega \in \Omega^1(X(\Gamma))$ with $\pi^*(\omega) = f(\tau)d\tau$. Then $\pi \circ \gamma = \pi$ and so

$$f(\tau)d\tau = \pi^*(\omega) = (\pi \circ \gamma)^*(\omega) = \gamma^*(\pi^*(\omega)) = \gamma^*(f(\tau)d\tau) = f(\gamma(\tau))\gamma'(\tau)d\tau$$

and this last expression is equal to $(f[\gamma]_2)(\tau)d\tau$ by (1.4). Thus indeed $f = f[\gamma]_2$, so f is weakly modular of weight 2.

Now let $f \in \mathcal{A}_2(\Gamma)$. We need to define a differential $f_i(z)dz$ on each V_i and let $\omega(f) = \{f_i(z)dz\}$. As with the charts in Section 2.4, we define $\omega(f)$ in three steps: at generic points of $X(\Gamma)$, at elliptic points, and at cusps. Let $\varphi_i : U_i \to V_i$ be a chart around a generic point τ. At such points, we simply change coordinates to be centered around 0. Recall δ_τ from Proposition 2.18. Then we define $f_i(z)dz = (f[\delta_\tau^{-1}])(z)dz$ on V_i.

If τ is an elliptic point, recall the matrix δ_τ and integer h_τ from Proposition 2.18. Then

$$f_i(z)dz = \frac{z^{1/h_\tau}\ (f[\delta_\tau^{-1}]_2)(z^{1/h_\tau})}{h_\tau z}dz \quad \text{on } V_i. \tag{3.3}$$

It is not immediately obvious that this is well-defined given the ambiguity of z^{1/h_τ} over the complex numbers, however it can be shown that the function $z(f[\delta_\tau^{-1}])(z)$ is invariant under the transformation $z \mapsto \mu_{h_\tau} z$, making this well-defined. Similarly if s is a cusp recall the matrix δ_s and integer h_s from Definition 2.19. Then

$$f_i(z)dz = \frac{h_s}{2\pi i z}(f[\delta_s^{-1}])\left(\log(z)\frac{h_s}{2\pi i}\right)\ dz \quad \text{on } V_i \tag{3.4}$$

and this is also well-defined. The local versions of π from $\mathbb{C}$ to $\mathbb{C}$ around elliptic points and cusps are precisely the ψ maps from Proposition 2.18 and Proposition 2.20, so one can check that pulling back the formulas in (3.3) and (3.4) by the corresponding ψ gives $f(\tau)d\tau$. □

Corollary 3.5. *$\mathcal{S}_2(\Gamma)$ is isomorphic to $\Omega^1_{hol}(X(\Gamma))$.*

Proof. Given a meromorphic form $f : \mathcal{H} \to \mathbb{C}$ and $\tau \in \mathcal{H}$, let $\nu_\tau(f)$ denote the order of vanishing of f at τ, i.e. if $f(z) = \sum_{n=-\infty}^{\infty} a_n(z-\tau)^n$ is the Laurent expansion of f, then $a_n = 0$ for all $n < \nu_\tau(f)$. If s is a cusp then let $\nu_s(f)$ denote the order of the first non-zero coefficient in the Fourier development of $f[\delta_s^{-1}]$. Since $(c\tau + d)^{-2}$ has no poles or zeroes on $\mathcal{H}$, given $\tau \in \mathcal{H}$ and $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ we have

$$\nu_\tau(f[\gamma]) = \nu_{\gamma(\tau)}(f).$$

An automorphic form f is a cusp form if and only if it is holomorphic on $\mathcal{H}$, so $\nu_\tau(f) \geq 0$ for all $\tau \in \mathcal{H}$, and zero at the cusps, so $\nu_s(f) \geq 1$ for all $s \in \mathbb{Q} \cup \{\infty\}$.

Let $\omega(f) = \{f_i(z)dz\}$. If $f_i(z)$ is centered around a generic point τ then

$$\nu_0(f_i) = \nu_0(f[\delta_\tau^{-1}]) = \nu_\tau(f)$$

so f_i is holomorphic if and only if f is holomorphic at τ. If τ is an elliptic point we analyze (3.3) to get

$$\nu_0(f_i) = \frac{1}{h_\tau}\nu_0(z(f[\delta_\tau^{-1}])(z)) - 1.$$

[6]We use ω both as a variable to represent elements of $\Omega^1(X(\Gamma))$ and as the name of the isomorphism, however we do this in two different steps so these two uses should not overlap.

Our assertion that the function $z^{1/h_\tau}(f[\delta_\tau^{-1}])(z^{1/h_\tau})$ is well-defined and meromorphic also implies that $\nu_0(z(f[\delta_\tau^{-1}])(z)) = 1 + \nu_\tau(f)$ is a multiple of h_τ. Thus $\nu_0(f_i)$ is greater than or equal to zero if and only if $1 + \nu_\tau(f)$ is strictly greater than zero and so at least h_τ. This is true if $\nu_\tau(f) \geq 0$. Thus f_i is holomorphic around all non-cusp points if and only if f is holomorphic on $\mathcal{H}$.

If $f_i(z)$ is centered around a cusp, then the Laurent expansion of

$$(f[\delta_s^{-1}])\left(\log(z)\frac{h_s}{2\pi i}\right)$$

is precisely the Fourier development of $f[\delta_s^{-1}]$ (with a change of variables) and so

$$\nu_0(f_i) = \nu_0\left(\frac{h_s}{2\pi i z}(f[\delta_s^{-1}])\left(\log(z)\frac{h_s}{2\pi i}\right)\right) = -1 + \nu_s(f).$$

Thus $\nu_0(f_i) \geq 0$ if and only if $\nu_s(f) \geq 1$, i.e. if f is a cusp form. □

3.2. **The Jacobian.** Once we have differential forms, it is natural to do what differential forms were made for: integrate them. In particular, there is a well-defined notion of path integration.

Definition 3.6. Let X be a Riemann surface and $\omega = (\omega_i)_{i\in I}$ be a differential form on X. Let $\gamma : [0,1] \to X$. Suppose the image of γ lies entirely within one coordinate patch U_i with chart φ_i. Then we define

$$\int_\gamma \omega = \int_{\varphi_i\circ\gamma} \omega_i.$$

If the image of γ goes between coordinate patches, we split it up and take the sum.

To show that this is well defined, we need to show that if the image of γ lies in more than one coordinate patch, say in the intersection $U_i \cap U_j$, then the value of the integral is independent of which patch we choose. This follows from the compatibility as we defined it in Definition 3.1:

$$\int_{\varphi_j\circ\gamma} \omega_j|_{V_{j,i}} = \int_{\varphi_{j,i}\circ\varphi_i\circ\gamma} \omega_j|_{V_{j,i}} = \int_{\varphi_i\circ\gamma} \varphi_{j,i}^*(\omega_j|_{V_{j,i}}) = \int_{\varphi_i\circ\gamma} \omega_i|_{V_{i,j}}.$$

In standard complex analysis, path integrals of holomorphic functions are completely determined by the endpoints of the path, i.e. integration around loops is always zero. However, on a general Riemann surface this is not the case. Therefore it would be helpful to (in some sense) quotient away by integration over loops.

These path integrals are linear operators on $\Omega^1_{\text{hol}}(X)$ so they are elements of the dual space $\Omega^1_{\text{hol}}(X)^\wedge$. Thus we define the following subgroup:

Definition 3.7. Let X be a compact Riemann surface. The *(first) homology group of* X, denoted $H_1(X,\mathbb{Z})$, is the subgroup of $\Omega^1_{\text{hol}}(X)^\wedge$ generated by integrals over loops. In other words

$$H_1(X,\mathbb{Z}) = \left\{\sum_{i=1}^n k_i \int_{\alpha_i} : n \in \mathbb{N}, k_i \in \mathbb{Z}, \text{ and } \alpha_i : [0,1] \to X \text{ is a loop}\right\}$$

Definition 3.8. The Jacobian of X is

$$\text{Jac}(X) = \Omega^1_{\text{hol}}(X)^\wedge / H_1(X,\mathbb{Z}).$$

So the elements of $\mathrm{Jac}(X)$ are essentially integrals of the form $\int_{x_1}^{x_2}$, since such integrals are defined up to integration around loops (see Theorem 3.11).

If $X = X(\Gamma)$ is a modular curve, by the dual of the map ω from Corollary 3.5, we know that

$$\Omega^1_{\mathrm{hol}}(X(\Gamma))^\wedge \cong \mathcal{S}_2(\Gamma)^\wedge.$$

Let $H_1(X(\Gamma), \mathbb{Z})$ denote both the homology group of $X(\Gamma)$ and its image under the map $\omega^\wedge$. Then we can write

$$\mathrm{Jac}(X(\Gamma)) = \mathcal{S}_2(\Gamma)^\wedge / H_1(X, \mathbb{Z}).$$

So we will more often think of the elements of the Jacobian as equivalence classes of linear maps on the space of weight 2 cusp forms. For simplicity, we introduce the notation

$$\mathrm{Jac}(X_0(N)) = J_0(N) \quad \text{and} \quad \mathrm{Jac}(X_1(N)) = J_1(N).$$

The definition of the Jacobian is essentially sufficient for our second statement of the modularity theorem, however the theorem refers to a holomorphism from a Jacobian, which requires it to have a complex analytic structure. For the result here we refer to the reader to a text on Riemann surface theory such as [2]:

Proposition 3.9. *Let X be a Riemann surface. It is a well-known result from topology that X is a sphere with g tori stuck to it for some $g \in \mathbb{N}$. For each of these tori, let $\alpha_i : [0,1] \to X$ be a loop around the inside like an equator, and let β_i be a perpendicular loop like a band around the torus. Then*

$$\Omega^1_{hol}(X)^\wedge = \mathbb{R}\int_{\alpha_1} \oplus\, \mathbb{R}\int_{\beta_1} \oplus \cdots \oplus \mathbb{R}\int_{\alpha_g} \oplus\, \mathbb{R}\int_{\beta_g}$$

and

$$H_1(X, \mathbb{Z}) = \mathbb{Z}\int_{\alpha_1} \oplus\, \mathbb{Z}\int_{\beta_1} \oplus \cdots \oplus \mathbb{Z}\int_{\alpha_g} \oplus\, \mathbb{Z}\int_{\beta_g}.$$

In other words, $\Omega_1(X)^\wedge$ is a finite dimensional vector space over $\mathbb{C}$ and $H_1(X, \mathbb{Z})$ is a lattice, so $\mathrm{Jac}(X)$ is complex torus (specifically a g-dimensional complex torus).

3.3. **The Modularity Theorem.**

Theorem 3.10 (Modularity Theorem, Version II)**.** *For every complex elliptic curve E with $j(E) \in \mathbb{Q}$, there is some $N \in \mathbb{N}$ such that there exists a surjective holomorphic homomorphism of complex tori from $J_0(N)$ to E.*

We will now briefly explain why versions (I) and (II) of the modularity theorem are equivalent, which will require some statements without proof. In particular, one of the most important results about the Jacobian is Abel's Theorem:

Theorem 3.11 (Abel's Theorem)**.** *Let X be a Riemann surface and fix a base point $x_0 \in X$. Let $\sum_x n_x x$ be a* degree-0 divisor *on X, i.e. a finite formal sum over points in X with each $n_x \in \mathbb{Z}$ such that $\sum_x n_x = 0$. Then the map into the Jacobian*

$$\sum_x n_x x \mapsto \sum_x n_x \int_{x_0}^{x}$$

is well-defined and surjects. Furthermore, the map descends to an isomorphism between the Jacobian and the degree-0 Picard group *(which we do not define here).*

We state this theorem without defining the Picard group because it shows that the Jacobian does just consist of $\mathbb{Z}$-linear sums of integrals of the form $\int_{x_0}^{x}$. Note that by adding any integer multiple of $\int_{x_0}^{x_0}$ we can essentially ignore the requirement that the n_x sum to 0. Furthermore, it is used for the following result:

Proposition 3.12. *If a Riemann surface X has genus greater than 0, it embeds in its Jacobian by*

$$X \to \mathrm{Jac}(X), \qquad x \mapsto \int_{x_0}^{x}.$$

If X is a complex elliptic curve, then this embedding is an isomorphism.

Proof (sketch). That this map is well-defined is an immediate consequence of Abel's theorem, as outlined above. The fact that this map is injective requires a more careful application of Abel's theorem (first injecting into the Picard group and then using the isomorphism from Abel's theorem) which we omit.

Let $\mathbb{C}/\Lambda$ be a complex elliptic curve. Then holomorphic differentials on $\mathbb{C}/\Lambda$ pull back to holomorphic Λ-periodic functions on $\mathbb{C}$. Such functions are bounded, and so are constant. Thus $\Omega^1_{\mathrm{hol}}(\mathbb{C}/\Lambda)$ only consists of constant functions, and so the integrals in $\Omega^1_{\mathrm{hol}}(\mathbb{C}/\Lambda)^\wedge$ are translation invariant. In particular, letting $x_0 = 0 + \Lambda$ and computing in the Jacobian (i.e. modulo loops), we have

$$\int_{0+\Lambda}^{x_1+\Lambda} + \int_{0+\Lambda}^{x_2+\Lambda} = \int_{0+\Lambda}^{x_1+\Lambda} + \int_{x_1+\Lambda}^{x_1+x_2+\Lambda} = \int_{0+\Lambda}^{x_1+x_2+\Lambda}.$$

This calculation shows that the map both surjects and is a group homomorphism when X is an elliptic curve. □

We also need a generalization of Proposition 2.3:

Lemma 3.13. *Let $g, h \in \mathbb{N}$, let $\mathbb{C}^g/\Lambda_g$ and $\mathbb{C}^h/\Lambda_h$ be complex tori, and let $\varphi : \mathbb{C}^g/\Lambda_g \to \mathbb{C}^h/\Lambda_h$ be a holomorphic homomorphism. Then*

$$\varphi(z + \Lambda_g) = Mz + b + \Lambda_h$$

for some $b \in \mathbb{C}^h$ and $M \in M_{h\times g}(\mathbb{C})$.

Proposition 3.14. *Versions (I) and (II) of the modularity theorem are equivalent.*

Proof. First we show that (II) implies (I). Let $\varphi : J_0(N) \to E$ be the surjective holomorphic homomorphism from the theorem. This surjection shows that $J_0(N)$ is non-trivial, and so $X_0(N)$ has genus greater than 0 (see Proposition 3.9). Thus we have an embedding $f : X_0(N) \to J_0(N)$ and so $\varphi \circ f$ is a map from $X_0(N)$ to E. This composite map inherits being a holomorphic homomorphism, but it remains to show it is surjective. By Proposition 1.7, it suffices to show that it is non-constant. Since $f(x_0)$ is the zero integral, $\varphi(f(x_0)) = 0_E$. To show that $\varphi \circ f$ is non-constant, it thus suffices to show that the image of f is not contained in the kernel of φ. As noted above, Abel's theorem shows that the $\mathbb{C}$-span (in fact, the $\mathbb{Z}$-span) of the image of f is all of $J_0(N)$. However, Lemma 3.13 shows that the kernel of φ is a subspace of $J_0(N)$ of strictly lower dimension as a $\mathbb{C}$ vector space, and so $\mathrm{span}(\ker(\varphi)) \neq J_0(N)$. Thus the image of f cannot be a subset of the kernel of φ, and this completes the proof.

To show that (I) implies (II), let $h : X_0(N) \to E$ be the holomorphic surjection given by version (I). The dual of the pullback of h, $(h^*)^\wedge$, sends homology to homology, since if α is a loop in $X_0(N)$ then

$$(h^*)^\wedge\left(\int_\alpha\right) = \int_\alpha h^* = \int_{h(\alpha)}$$

and $h(\alpha)$ is a loop in E. Thus $(h^*)^\wedge$ descends to the jacobians, giving a surjective holomorphic homomorphism

$$J_0(N) \to \mathrm{Jac}(E)$$

and by Proposition 3.12 we have $\mathrm{Jac}(E) \cong E$, so this is the map conjectured by version (II). □

4. Version (III)

4.1. **The Double Coset and Hecke Operators.** Let Γ_1 and Γ_2 be congruence subgroups of $\mathrm{SL}_2(\mathbb{Z})$. It is natural to consider maps between the spaces $\mathcal{M}_k(\Gamma_1)$ and $\mathcal{M}_k(\Gamma_2)$. The most important such maps come from the following family:

Definition 4.1. Let $\alpha \in \mathrm{GL}_2^+(\mathbb{Q})$. Then the *weight-k* $\Gamma_1\alpha\Gamma_2$ *operator* (or in general a *double coset operator*)

$$[\Gamma_1\alpha\Gamma_2]_k : \mathcal{M}_k(\Gamma_1) \to \mathcal{M}_k(\Gamma_2)$$

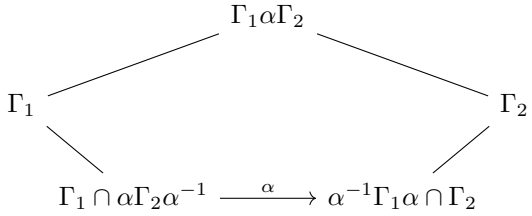
is given by

$$f \mapsto \sum f[\beta_j]_k$$

where the β_j are representatives of the orbits for Γ_1 in the set

$$\Gamma_1\alpha\Gamma_2 = \{\gamma_1\alpha\gamma_2 : \gamma_1 \in \Gamma_1 \text{ and } \gamma_2 \in \Gamma_2\}.$$

We can also think of this map as descending into subgroups rather than passing through the larger set $\Gamma_1\alpha\Gamma_2$, as in the diagram:



We will use this equivalence to show that the double coset operator is well-defined, and also introduce notation that will be useful later.

Definition 4.2. Let Γ_2 and Γ_3 be congruence subgroups of $\mathrm{SL}_2(\mathbb{Z})$ with $\Gamma_3 \leq \Gamma_2$. Let π_2 be the projection map from $X(\Gamma_3)$ to $X(\Gamma_2)$. Then the *trace* of π_2 is the map $\mathrm{tr}_{\pi_2} : \mathcal{M}_k(\Gamma_3) \to \mathcal{M}_k(\Gamma_2)$ given by

$$f \mapsto \sum f[\gamma_j]_k$$

where the γ_j are coset representatives for $\Gamma_3\backslash\Gamma_2$.

Proposition 4.3. *The trace is well-defined. That is, it is independent of the choice of coset representatives, and the resulting function is indeed in* $\mathcal{M}_k(\Gamma_2)$.

Proof. Suppose γ_j and $\{\gamma_j'\}$ are two different representatives for the same coset of Γ_3 in Γ_2, so that each $\gamma_j = \gamma_{3,j}\gamma_j'$ for some $\gamma_{3,j} \in \Gamma_3$. Since $f \in \mathcal{M}_k(\Gamma_3)$, we have

$$f[\gamma_j'] = f[\gamma_{3,j}\gamma_j] = (f[\gamma_{3,j}])[\gamma_j] = f[\gamma_j].$$

Let $\gamma_2 \in \Gamma_2$. It is a fact of group theory that multiplication by γ_2 permutes the coset representatives for any subgroup of Γ_2. In particular, $\{\gamma_j\gamma_2\}$ is a new set of equally valid coset representatives. Since we are taking the sum over all cosets, and using the previous result that it doesn't matter which coset representatives we choose, we also get

$$(\mathrm{tr}_{\pi_2} f)[\gamma_2]_k = \sum_j f[\gamma_j]_k[\gamma_2]_k = \sum_j f[\gamma_j\gamma_2]_k = \mathrm{tr}_{\pi_2} f.$$

Proposition 2.28 shows that the resulting function is also holomorphic at the cusps, so indeed the resulting function is in $\mathcal{M}_k(\Gamma_2)$. □

Proposition 4.4. *Let $[I]_k$ denote the inclusion map from $\mathcal{M}_k(\Gamma_1)$ to $\mathcal{M}_k(\Gamma_1 \cap \alpha\Gamma_2\alpha^{-1})$ and let π_2 be the projection map from $X(\alpha^{-1}\Gamma_1\alpha \cap \Gamma_2)$ to $X(\Gamma_2)$. Then*

$$[\Gamma_1\alpha\Gamma_2] = \mathrm{tr}_{\pi_2} \circ [\alpha]_k \circ [I]_k.$$

Proof. To show that tr_{π_2} is well-defined, we have to show that $\Gamma_3 = \alpha^{-1}\Gamma_1\alpha \cap \Gamma_2$ is a congruence subgroup. Suppose Γ_1 is a congruence subgroup of level N_1 and Γ_2 of level N_2. Let $K \in \mathbb{Z}$ such that $K\alpha$ and $K\alpha^{-1}$ have integer entries. We want to show that $\Gamma(K^2N_1N_2) \subset \Gamma_3$. Since N_2 divides $K^2N_1N_2$, we have $\Gamma(K^2N_1N_2) \subset \Gamma_2$. Furthermore

$$\begin{aligned}\alpha\Gamma(K^2N_1N_2)\alpha^{-1} &\subset \alpha(I + K^2N_1N_2\,\mathrm{M}_2(\mathbb{Z}))\alpha^{-1} \\ &= I + N_1N_2 \cdot K\alpha \cdot \mathrm{M}_2(\mathbb{Z}) \cdot K\alpha^{-1} \\ &\subset I + N_1N_2\,\mathrm{M}_2(\mathbb{Z})\end{aligned}$$

and since $\alpha\Gamma(K^2N_1N_2)\alpha^{-1}$ only consists of matrices with determinant 1, it is in fact a subset of $\Gamma(N_1N_2) \subset \Gamma(N_1) \subset \Gamma_1$. Thus

$$\alpha\Gamma(K^2N_1N_2)\alpha^{-1} \subset \Gamma_1 \implies \Gamma(K^2N_1N_2) \subset \alpha^{-1}\Gamma_1\alpha$$

so Γ_3 is indeed a congruence subgroup.

Thus we know that this composition is indeed a map from $\mathcal{M}_k(\Gamma_1)$ to $\mathcal{M}_k(\Gamma_2)$ by Proposition 2.28 and Proposition 4.3. All that remains is to show it is equivalent to $[\Gamma_1\alpha\Gamma_1]$. We have

$$\mathrm{tr}_{\pi_2} \circ [\alpha]_k \circ [I]_k = \sum_j (f[\alpha]_k)[\gamma_j]_k = \sum_j f[\alpha\gamma_j]_k$$

where $\{\gamma_j\}$ are the coset representatives for Γ_3 in Γ_2. It suffices to show that $\{\alpha\gamma_j\}$ are the orbit representatives for Γ_1 in $\Gamma_1\alpha\Gamma_2$. In particular, the map

$$\alpha : \Gamma_2 \to \Gamma_1\alpha\Gamma_2, \quad \gamma_2 \mapsto \alpha\gamma_2$$

induces a bijection between $\Gamma_3\backslash\Gamma_2$ and $\Gamma_1\backslash\Gamma_1\alpha\Gamma_2$. The map from Γ_2 to $\Gamma_1\backslash\Gamma_1\alpha\Gamma_2$ taking γ_2 to $\Gamma_1\alpha\gamma_2$ surjects. The kernel is comprised of matrices $\gamma_2 \in \Gamma_2$ such that $\Gamma_1\alpha\gamma_2 = \Gamma_1\alpha$. In other words $\gamma_2 \in \alpha^{-1}\Gamma_1\alpha$, so the kernel is precisely Γ_3. Quotienting out by the kernel, this map becomes a bijection from $\Gamma_3\backslash\Gamma_2$ to $\Gamma_1\backslash\Gamma_1\alpha\Gamma_2$, as desired. □

Corollary 4.5. *The double coset operators take cusp forms to cusp forms.*

Proof. This follows immediately from the definition and Proposition 2.28. □

Certain special cases of the double coset operator are of particular interest, the *Hecke operators.*

Definition 4.6. Let p be a prime. Then T_p is an operator which takes $\mathcal{M}_k(\Gamma_0(N))$ to itself given by

$$T_p = \left[\Gamma_0(N)\left[\begin{smallmatrix}1 & 0\\ 0 & p\end{smallmatrix}\right]\Gamma_0(N)\right]_k$$

or defined equivalently for $\Gamma_1(N)$.

There is another class of Hecke operators called diamond operators. However, these operators act trivially on $\mathcal{M}_k(\Gamma_0(N))$, and so will be ultimately unnecessary for our statement of the modularity theorem.[7] Similarly, one can extend the definition to all n in $\mathbb{N}$ in a non-trivial manner, but we will implicitly account for this by using the Hecke algebra as defined in Definition 4.20.

Proposition 4.7. *Let q and p be primes. Then $T_pT_q = T_qT_p$. In other words, the Hecke operators commute.*

Proof. We will prove this by explicitly finding the coset representatives β_j for $\Gamma_0(N)$ in $\Gamma_0(N)\alpha\Gamma_0(N)$ (where $\alpha = \left[\begin{smallmatrix}1 & 0\\ 0 & p\end{smallmatrix}\right]$), and then computing. Using the result from in Proposition 4.4, we first find the representatives for $\Gamma_3\backslash\Gamma_0(N)$ where $\Gamma_3 = \alpha^{-1}\Gamma_0(N)\alpha \cap \Gamma_0(N)$. By conjugating a generic matrix in $\Gamma_0(N)$ by α and then requiring the resulting matrix have integer entries, we find that

$$\Gamma_3 = \left\{\begin{bmatrix} a & pb\\ Nc & d\end{bmatrix} : a, b, c, d \in \mathbb{Z} \text{ and } ad - (pb)(Nc) = 1\right\}.$$

We will show that the coset representatives are

$$\gamma_j = \begin{bmatrix}1 & j\\ 0 & 1\end{bmatrix} \quad \text{for} \quad 0 \le j < p$$

if $p|N$. If p does not divide N, then we add the additional coset representative

$$\gamma_\infty = \begin{bmatrix}p & m\\ N & n\end{bmatrix}$$

where m, n are such that $pn - mN = 1$ (if $p \nmid N$ they are coprime, so such n, m must exist). Two such matrices are Γ_3-equivalent if and only if $\gamma_j\gamma_k^{-1} \in \Gamma_3$, which requires the upper right entry to be divisible by p. However, if both $j, k < \infty$ then the upper right entry of $\gamma_j\gamma_k^{-1}$ is $j - k$, so either $j = k$ and they are trivially in the same orbit or $j - k < p$, and so the upper right entry is not divisible by p. Additionally

$$\gamma_j\gamma_\infty^{-1} = \begin{bmatrix} n - jN & -m + jp\\ -N & p\end{bmatrix}$$

and m is not a multiple of p, since otherwise we could not have $pn - mN = 1$. Thus neither is $-m + jp$ for any j. This shows these matrices are in distinct orbits. Let

$$\delta = \begin{bmatrix} a_0 & b_0\\ Nc_0 & d_0\end{bmatrix} \in \Gamma_0(N).$$

[7] The general theory is more often developed with the smaller $\Gamma_1(N)$ in mind, rather than $\Gamma_0(N)$. However, the Modularity Theorem is more precise when referencing $\Gamma_0(N)$, so going forward we develop the theory for this subgroup.

We want to find a matrix $\gamma \in \Gamma_3$ such that $\gamma\gamma_j = \delta$ for some $0 \leq j < p$ or $j = \infty$. If $p|N$, then it cannot be that $p|a_0$, since $\det(\delta) = 1$. Thus there is some j between 0 and $p-1$ such that $b_0 - ja_0$ is divisible by p. Then let

$$\gamma = \begin{bmatrix} a_0 & b_0 - ja_0 \\ Nc_0 & d_0 - jNc_0 \end{bmatrix}.$$

Which is indeed in Γ_3 and $\gamma\gamma_j = \delta$. If $p \nmid N$, we have the additional possibility that $p|a$. In this case

$$\delta\gamma_\infty^{-1} = \begin{bmatrix} na_0 - Nb_0 & -ma_0 + pb_0 \\ Nnc_0 - Nd_0 & -mNc_0 + pd_0 \end{bmatrix}.$$

Since $p|a_0$, we indeed have that p divides the upper right entry, and so $\delta\gamma_\infty^{-1} \in \Gamma_3$. This shows that we have successfully identified the coset representatives. To find orbit representatives for $\Gamma_0(N)$ in $\Gamma_0(N)\alpha\Gamma_0(N)$ we left multiply by α to get

$$\beta_j = \begin{bmatrix} 1 & j \\ 0 & p \end{bmatrix} \quad \text{for} \quad 0 \leq j < p$$

plus the additional representative when $p \nmid N$:

$$\beta_\infty = \begin{bmatrix} p & m \\ Np & np \end{bmatrix} = \begin{bmatrix} 1 & m \\ N & pn \end{bmatrix} \begin{bmatrix} p & 0 \\ 0 & 1 \end{bmatrix}$$

and since this first matrix is in $\Gamma_0(N)$, we let $\beta_\infty = \left[\begin{smallmatrix} p & 0 \\ 0 & 1 \end{smallmatrix}\right]$.

Now let $\beta_{p,j}$ be the representatives for T_p and $\beta_{q,j}$ the representatives for T_q. Suppose that both p and q divide N. Then

$$T_pT_q(f) = \sum_{j=0}^{p-1}\sum_{i=0}^{q-1} f[\beta_{q,i}]_k[\beta_{p,j}]_k = \sum_{j=0}^{p-1}\sum_{i=0}^{q-1} f\left[\begin{smallmatrix} 1 & pi+j \\ 0 & pq \end{smallmatrix}\right]_k = \sum_{j=1}^{pq-1} f\left[\begin{smallmatrix} 1 & j \\ 0 & pq \end{smallmatrix}\right]_k$$

which is symmetric in p and q. Now suppose only $q \nmid N$. Then the above computation still applies for all $\beta_{q,i}$ except $i = \infty$. Thus we only need to show that the expressions

$$\sum_{j=0}^{p-1} f[\beta_{q,\infty}]_k[\beta_{p,j}]_k = \sum_{j=0}^{p-1} f\left[\begin{smallmatrix} q & qj \\ 0 & p \end{smallmatrix}\right]_k \quad \text{and} \quad \sum_{j=0}^{p-1} f[\beta_{p,j}]_k[\beta_{q,\infty}]_k = \sum_{j=0}^{p-1} f\left[\begin{smallmatrix} q & j \\ 0 & p \end{smallmatrix}\right]_k$$

are equivalent. By left multiplying by $\left[\begin{smallmatrix} 1 & k \\ 0 & 1 \end{smallmatrix}\right]$, a matrix $\left[\begin{smallmatrix} q & j \\ 0 & p \end{smallmatrix}\right]$ can be made $\Gamma_0(N)$-equivalent to any matrix $\left[\begin{smallmatrix} q & i \\ 0 & p \end{smallmatrix}\right]$ with $i \equiv j \pmod{p}$. Thus it suffices to show that $\{qj\}_{j=1}^{p-1}$ contains all remainders mod p, which is immediate since p and q are distinct primes and so have greatest common divisor 1.

If both p and q don't divide N, then we are only left to check $f[\beta_{q,\infty}]_k[\beta_{p,\infty}]_k = f[\beta_{p,\infty}]_k[\beta_{q,\infty}]_k$, and indeed $\beta_{q,\infty}\beta_{p,\infty} = \left[\begin{smallmatrix} pq & 0 \\ 0 & 1 \end{smallmatrix}\right] = \beta_{p,\infty}\beta_{q,\infty}$. □

Definition 4.8. Let $f \in \mathcal{S}_k(\Gamma_0(N))$ be non-zero. Then f is an *eigenform* if it is an eigenvector of T_p for all primes p, i.e. for all primes p there exists $\lambda \in \mathbb{C}$ such that

$$T_p(f) = \lambda f$$

If the first coefficient a_1 in the Fourier development of f is 1, then f is *normalized*.

4.2. **Hecke Operators and Jacobians.** The double coset operator maps $\mathcal{S}_2(\Gamma_1)$ to $\mathcal{S}_2(\Gamma_2)$, and therefore it has a dual map

$$[\Gamma_1 \alpha \Gamma_2]_2^\wedge : \mathcal{S}_2(\Gamma_2)^\wedge \to \mathcal{S}_2(\Gamma_1)^\wedge$$

acting by composition. The goal of this section is to show that this action descends to the Jacobian. Since

$$[\Gamma_1 \alpha \Gamma_2]_2 = \mathrm{tr}_{\pi_2} \circ [\alpha]_2 \circ [I]_2$$

it suffices to show that the dual of each of these functions (translated into the language of differential forms) preserves the homology group.

Proposition 4.9. *Let $\Gamma_1 \leq \Gamma_2$ be congruence subgroups and let π be the projection map from $X(\Gamma_1)$ to $X(\Gamma_2)$. The map from $\Omega^1_{hol}(X(\Gamma_2))^\wedge$ to $\Omega^1_{hol}(X(\Gamma_1))^\wedge$ induced by $\mathrm{tr}_\pi^\wedge$ takes homology to homology.*

Proof. Let π_1 and π_2 be the projection maps from $\mathcal{H}^*$ to $X(\Gamma_1)$ or $X(\Gamma_2)$ respectively, and let $\omega_1 : \mathcal{S}_2(\Gamma_1) \to \Omega^1_{\mathrm{hol}}(X(\Gamma_1))$ and $\omega_2 : \mathcal{S}_2(\Gamma_2) \to \Omega^1_{\mathrm{hol}}(X(\Gamma_2))$ be the isomorphisms given by Corollary 3.5, which were defined such that given $f \in \mathcal{S}_2(\Gamma_i)$

$$\pi_i^*(\omega_i(f)) = f(\tau)d\tau \quad i = 1, 2.$$

We would like to show that the map

$$\omega_2 \circ \mathrm{tr}_\pi \circ \omega_1^{-1} : \Omega^1_{\mathrm{hol}}(X(\Gamma_1)) \to \Omega^1_{\mathrm{hol}}(X(\Gamma_2))$$

which we will denote by the same symbol tr_π, dualizes to a map sending homology to homology. We have

$$\mathrm{tr}_\pi(\omega_1(f)) = \omega_2 \left(\sum_j f[\gamma_j]_2 \right)$$

where γ_j are the coset representatives for $\Gamma_1 \backslash \Gamma_2$. Let $\delta : [0,1] \to X(\Gamma_2)$ be a loop so that $\int_\delta$ is an element of $H_1(X(\Gamma_2), \mathbb{Z})$. For any $\omega_1(f) \in \Omega^1_{\mathrm{hol}}(X(\Gamma_1))$ we have

$$\left(\mathrm{tr}_\pi^\wedge \int_\delta \right) (\omega_1(f)) = \int_\delta \mathrm{tr}_\pi(\omega_1(f)) = \int_\delta \omega_2 \left(\sum_j f[\gamma_j]_2 \right)$$

Now let $\tilde{\delta}$ be a lift of δ to $\mathcal{H}^*$, which is to say a continuous function $\tilde{\delta} : [0,1] \to \mathcal{H}^*$ (although we can't guarantee it is a loop) such that $\pi_2 \circ \tilde{\delta} = \delta$.[8] Then

$$\begin{aligned}
\int_\delta \omega_2 \left(\sum_j f[\gamma_j]_2 \right) = \int_{\tilde{\delta}} \pi_2^* \left(\omega_2 \left(\sum_j f[\gamma_j]_2 \right) \right) &= \sum_j \int_{\tilde{\delta}} f[\gamma_j]_2(\tau)\, d\tau \\
&= \sum_j \int_{\tilde{\delta}} \gamma_j'(\tau) f(\gamma_j(\tau))\, d\tau \\
&= \sum_j \int_{\gamma_j \circ \tilde{\delta}} f(\tau)\, d\tau \\
&= \sum_j \int_{\pi_1 \circ \gamma_j \circ \tilde{\delta}} \omega_1(f).
\end{aligned}$$

[8] See [3] for why such a lift always exists.

So we have to show that, when taken together, the paths $\pi_1 \circ \gamma_j \circ \tilde{\delta}$ form a loop or integer sum of loops in $X(\Gamma_1)$. Let $\tilde{\delta}(0) = \tau_0$ so that $\tilde{\delta}(1) = \gamma_2(\tau_0)$ for some $\gamma_2 \in \Gamma_2$ (recall that $\tilde{\delta}$ projects to a loop in $X(\Gamma_2)$). Then for each j, the start and end points of $\pi_1 \circ \gamma_j \circ \tilde{\delta}$ are

$$\Gamma_1\gamma_j(\tau_0) \quad \text{to} \quad \Gamma_1\gamma_j\gamma_2(\tau_0).$$

The set $\{\Gamma_1\gamma_j(\tau_0)\}$ is precisely the (finite, discrete) set of points in $X(\Gamma_1)$ which π maps to $\Gamma_2\tau_0$. As we have mentioned before, multiplication by γ_2 permutes coset representatives, so the function sending the initial point of each path $\pi_1 \circ \gamma_j \circ \tilde{\delta}$ to its final point is a permutation on the finite set $\pi^{-1}(\Gamma_2\tau_0)$. Therefore, concatenating these paths must indeed give some integer sum of loops, given by the cyclic structure of this permutation.

Finally, since tr_π is $\mathbb{Z}$-linear, showing this result for operators of the form $\int_\delta$ immediately extends to all of $H_1(X(\Gamma_2), \mathbb{Z})$. □

Proposition 4.10. *Let Γ_1 and Γ_2 be congruence subgroups and let $\alpha \in \mathrm{GL}_2^+(\mathbb{Q})$ such that $\alpha\Gamma_1\alpha^{-1} \subset \Gamma_2$. The map from $\Omega^1_{hol}(X(\Gamma_1))^\wedge$ to $\Omega^1_{hol}(X(\Gamma_2))^\wedge$ induced by $[\alpha]_2$ takes homology to homology.*

Proof. As before, let ω_1 and ω_2 be the isomorphisms from Corollary 3.5, and let $\omega_2(f)$ be an arbitrary element of $\Omega^1_{\mathrm{hol}}(X(\Gamma_2))$. Then

$$(\omega_1 \circ [\alpha]_2 \circ \omega_2^{-1})(\omega_2(f)) = \omega_1(f[\alpha]_2)$$

We will denote the map $\omega_1 \circ [\alpha]_2 \circ \omega_2^{-1}$ by α^* (one can show it is the pullback of the map $\Gamma_1\tau \mapsto \Gamma_2\alpha(\tau)$). Let $\delta : [0,1] \to X(\Gamma_1)$ be a loop. Using the same manipulations as before we get

$$\left((\alpha^*)^\wedge \int_\delta\right)(\omega_2(f)) = \int_\delta \alpha^*(\omega_2(f)) = \int_\delta \omega_1(f[\alpha]_2) = \int_{\pi_2\circ\alpha\circ\tilde{\delta}} \omega_2(f).$$

Where $\tilde{\delta}$ is a lift of δ from $X(\Gamma_1)$ to $\mathcal{H}^*$. Again let $\tau_0 = \tilde{\delta}(0)$ so $\tilde{\delta}(1) = \gamma_1(\tau_0)$ for some $\gamma_1 \in \Gamma_1$. Then

$$(\pi_2 \circ \alpha \circ \tilde{\delta})(0) = \Gamma_2\alpha(\tau_0) \quad \text{and} \quad (\pi_2 \circ \alpha \circ \tilde{\delta})(1) = \Gamma_2\alpha\gamma_1(\tau_0).$$

Since $\alpha\Gamma_1\alpha^{-1} \subset \Gamma_2$, we have $\alpha\gamma_1\alpha^{-1} = \gamma_2$ for some $\gamma_2 \in \Gamma_2$ and so

$$\Gamma_2\alpha\gamma_1(\tau_0) = \Gamma_2\gamma_2\alpha(\tau_0) = \Gamma_2\alpha(\tau_0).$$

Thus $\pi_2 \circ \alpha \circ \tilde{\delta}$ has the same start and end point, so indeed α^* takes integration over loops to integration over loops. Since it is $\mathbb{Z}$-linear, it therefore takes $H_1(X(\Gamma_1), \mathbb{Z})$ to $H_1(X(\Gamma_2), \mathbb{Z})$. □

Corollary 4.11. *Let p be a prime. Then the action of the Hecke operator T_p on $\mathcal{S}_k(\Gamma_0(N))^\wedge$ descends to the Jacobian $J_0(N)$.*

4.3. **The Petersson Inner Product.** We have introduced linear transformations between our spaces of modular forms, so in continuing to add linear-algebraic structure, in this section we introduce an inner product. This inner product will be very similar to the inner product familiar from Fourier theory:

$$\langle f, g\rangle = \int f(z)\overline{g(z)}\, dz$$

with the integral appropriately defined over $X(\Gamma)$ and certain factors added to ensure everything works.

Definition 4.12. Let Γ be a congruence subgroup of $\mathrm{SL}_2(\mathbb{Z})$ and let $\{\alpha_j\}$ be the coset representatives for $\{\pm I\}\Gamma\backslash \mathrm{SL}_2(\mathbb{Z})$. Let μ denote the *hyperbolic measure* on $\mathcal{H}$, such that if $\tau = x + yi \in \mathcal{H}$ then

$$d\mu(\tau) = \frac{dx\ dy}{y^2}.$$

If φ is Γ invariant then we define

$$\int_{X(\Gamma)} \varphi(\tau)\ d\mu(\tau) = \int_{\bigcup_j \alpha_j(\mathcal{D})} \varphi(\tau)\ d\mu(\tau)$$

where $\mathcal{D}$ is the fundamental domain from Proposition 2.6.

In other words, we can make sense of integrating a Γ invariant function over $X(\Gamma)$ by integrating over the fundamental domain from Proposition 2.12.

Proposition 4.13. *This is independent of our choice of α_j. Furthermore, if φ is continuous and bounded, this integral converges.*

Proof. If α_j' is another coset representative, then $\alpha_j' = \gamma\alpha_j$ for some $\gamma \in \Gamma$. Thus

$$\int_{\alpha_j'(\mathcal{D})} \varphi(\tau)\ d\mu(\tau) = \int_{(\gamma\circ\alpha_j)(\mathcal{D})} \varphi(\tau)\ d\mu(\tau) = \int_{\alpha_j(\mathcal{D})} \varphi(\gamma(\tau))\ d\mu(\gamma(\tau)).$$

We asserted that φ is Γ invariant, so it suffices to show that $d\mu(\tau)$ is as well. In fact, $d\mu(\tau)$ is $\mathrm{SL}_2(\mathbb{Z})$ invariant. To see this, we first do some manipulations with differential forms:

$$d\mu(\tau) = \frac{dx\ dy}{y^2} = \frac{d\left(\frac{\tau+\overline{\tau}}{2}\right)\ d\left(\frac{\tau-\overline{\tau}}{2i}\right)}{\left(\frac{\tau-\overline{\tau}}{2i}\right)^2} = \frac{(d\tau)^2 + d\overline{\tau}d\tau - d\tau d\overline{\tau} - (d\overline{\tau})^2}{(-i)(\tau-\overline{\tau})^2} = \frac{2i\ d\overline{\tau}d\tau}{(\tau-\overline{\tau})^2}.$$

And so we have

$$d\mu(\alpha(\tau)) = \frac{2i\ d\alpha(\overline{\tau})d\alpha(\tau)}{(\alpha(\tau)-\alpha(\overline{\tau}))^2} = \alpha'(\tau)^2\alpha'(\overline{\tau})^2\frac{2i\ d\overline{\tau}d\tau}{(\alpha(\tau)-\alpha(\overline{\tau}))^2}.$$

Let $\alpha = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ with $ad - bc = 1$. Then using (1.4) and analyzing the denominator and we have

$$\begin{aligned}(c\tau+d)^2(c\overline{\tau}+d)^2 &\left(\left(\frac{a\tau+b}{c\tau+d}\right)^2 + \left(\frac{a\overline{\tau}+b}{c\overline{\tau}+d}\right)^2 - 2\left(\frac{a\tau+b}{c\tau+d}\right)\left(\frac{a\overline{\tau}+b}{c\overline{\tau}+d}\right)\right) \\ &= ((a\tau+b)(c\overline{\tau}+d) - (a\overline{\tau}+b)(c\tau+d))^2 \\ &= ((ad-bc)\tau - (ad-bc)\overline{\tau})^2 \\ &= (\tau-\overline{\tau})^2\end{aligned}$$

so that

$$d\mu(\alpha(\tau)) = \alpha'(\tau)^2\alpha'(\overline{\tau})^2\frac{2i\ d\overline{\tau}d\tau}{(\alpha(\tau)-\alpha(\overline{\tau}))^2} = \frac{2i\ d\overline{\tau}d\tau}{(\tau-\overline{\tau})^2} = d\mu(\tau)$$

as desired. To show convergence, we note that

$$\int_{\bigcup_j \alpha_j(\mathcal{D})} \varphi(\tau)\ d\mu(\tau) = \sum_j \int_{\mathcal{D}} \varphi(\alpha_j(\tau))d\mu(\tau).$$

Thus it suffices to show that the integral of $\varphi(\alpha_j(\tau))$ (which is itself a bounded continuous function) converges on $\mathcal{D}$. Letting M be the bound for $\varphi \circ \alpha_j$, this is immediate since

$$\int_{\mathcal{D}} |\varphi(\alpha_j(\tau))| d\mu(\tau) \leq \iint_{\mathcal{D}} \frac{M}{y^2}\, dx\, dy$$

and this integral converges. □

Definition 4.14. Given a congruence subgroup Γ, the *volume* of $X(\Gamma)$ is

$$V_\Gamma = \int_{X(\Gamma)} d\mu(\tau).$$

Given $f, g \in \mathcal{S}_k(\Gamma)$, we want to use this integral to define an inner product. However, the function $f\overline{g}$ is not Γ invariant as is required by our definition.

Lemma 4.15. *Given $f, g \in \mathcal{S}_k(\Gamma)$, the function*

$$\varphi(\tau) = f(\tau)\overline{g(\tau)}\,\mathrm{Im}(\tau)^k$$

is Γ invariant and bounded.

Proof. Let $\gamma = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \Gamma$. Then, using Proposition 1.2, we get

$$\begin{aligned} \varphi(\gamma(\tau)) &= f(\gamma(\tau))\overline{g(\gamma(\tau))}\,\mathrm{Im}(\gamma(\tau))^k \\ &= f[\gamma]_k(\tau)(c\tau+d)^k \overline{g[\gamma]_k(\tau)}\,\overline{(c\tau+d)^k}\,\mathrm{Im}(\tau)^k |c\tau+d|^{-2k} \\ &= f(\tau)\overline{g(\tau)}\,\mathrm{Im}(\tau)^k \end{aligned}$$

as desired. To show it is bounded, we note that Γ invariance means it suffices to show that $\varphi \circ \alpha_i$ is bounded on $\mathcal{D}$ for α_i the coset representatives of Γ in $\mathrm{SL}_2(\mathbb{Z})$. Since it is continuous, $\varphi \circ \alpha_i$ is bounded on any compact subset of $\mathcal{D}$, namely below some sufficiently large cutoff on the imaginary part.

For sufficiently large τ, the magnitude of a modular form goes by the first non-zero term in its Fourier development, and for cusp forms this is the first non-constant term. Thus for $\mathrm{Im}(\tau)$ sufficiently large, $f[\alpha_i]_k$ and $\overline{g[\alpha_i]_k}$ are both at most of the order

$$|q_h| = \left|e^{2\pi i\tau/h}\right| = e^{-2\pi\,\mathrm{Im}(\tau)/h}$$

and this exponential decay is much faster than the growth of $\mathrm{Im}(\tau)^k$. □

Definition 4.16. Let Γ be a congruence subgroup of $\mathrm{SL}_2(\mathbb{Z})$. Then the *Petersson inner product* is

$$\langle\ ,\ \rangle_\Gamma : \mathcal{S}_2(\Gamma) \times \mathcal{S}_2(\Gamma) \to \mathbb{C},$$

where if $f, g \in \mathcal{S}_2(\Gamma)$ then

$$\langle f, g\rangle_\Gamma = \frac{1}{V_\Gamma} \int_{X(\Gamma)} f(\tau)\overline{g(\tau)}\,\mathrm{Im}(\tau)^k\, d\mu(\tau).$$

Proposition 4.13 and Lemma 4.15 combine to show that this integral converges. Furthermore this function is immediately linear in f, conjugate-symmetric, and positive definite, so it is an inner product. The factor of V_Γ is only really useful when comparing inner products on different curves.

4.4. **Oldforms and Newforms.** Let M divide N. Then $\Gamma_0(N)$ is a subgroup of $\Gamma_0(M)$, and so $\mathcal{S}_k(M)$ is a subset of $\mathcal{S}_k(N)$. Furthermore, for any $f \in \mathcal{S}_k(M)$, let d be a divisor of N/M and let

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

Then $\alpha_d \Gamma_0(M) \alpha_d^{-1} \subset \Gamma_1(N)$, so by Proposition 2.28 $f[\alpha_d]_k$ is an element of $\mathcal{S}_k(N)$ for any $f \in \mathcal{S}_k(M)$. This motivates the following definition:

Definition 4.17. The subspace of *oldforms at level* N is the subspace of $\mathcal{S}_k(\Gamma_0(N))$ generated over $\mathbb{C}$ by the set

$$\bigcup_{M|N} \bigcup_{d|\frac{N}{M}} \{f[\alpha_d] : f \in \mathcal{S}_k(\Gamma_0(M))\}$$

and is denoted $\mathcal{S}_k(\Gamma_0(N))^{\text{old}}$.

Naturally, once we have identified these oldforms, we want to focus on those forms whose behavior is new to the level N:

Definition 4.18. The *subspace of newforms of level* N is the orthogonal complement of the space of oldforms with respect to the Petersson inner product:

$$\mathcal{S}_k(\Gamma_0(N))^{\text{new}} = \left(\mathcal{S}_k(\Gamma_0(N))^{\text{old}}\right)^{\perp}.$$

We will however distinguish the subspace of newforms from the functions we will call newforms:

Definition 4.19. A *newform* of level N is a normalized eigenform (see Definition 4.8) that is in the space of newforms of level N.

This terminology is consistent because such newforms form a basis for the subspace of newforms (which we will not show). Thus we could call the subspace of newforms the subspace *generated* by newforms, although this would imply a definition of newform which was independent of the subspace.

4.5. **The Abelian Variety Associated to a Newform.** Our two statements of the modularity theorem so far associate every rational elliptic curve with some geometric or algebraic structure based on a modular *curve*. In this section, we introduce such a structure based on a modular *form*, which we will use in the final version of the modularity theorem, giving it more specificity.

Definition 4.20. Given $N \in \mathbb{N}$, the *Hecke algebra* is the ring of transformations of $\mathcal{S}_k(\Gamma_0(N))$ generated over $\mathbb{Z}$ by the Hecke operators,

$$\mathbb{T}_{\mathbb{Z}} = \mathbb{Z}[\{T_p : p \text{ prime}\}].$$

If f is an eigenform, then f is an eigenvector for all transformations in $\mathbb{T}_{\mathbb{Z}}$.

Definition 4.21. Let f be a newform of level N_f. Then the function $\lambda_f : \mathbb{T}_{\mathbb{Z}} \to \mathbb{C}$ that sends each operator T to its eigenvalue for f, i.e.

$$Tf = \lambda_f(T) f,$$

is a homomorphism from $\mathbb{T}_{\mathbb{Z}}$ to $\mathbb{C}$. Let I_f denote the kernel of this homomorphism.

By Corollary 4.11, we can consider the action of I_f on $J_0(N_f)$. In particular, $J_0(N_f)$ is a module over $\mathbb{T}_\mathbb{Z}$ and I_f an ideal of $\mathbb{T}_\mathbb{Z}$, so $I_f J_0(N_f)$ is a subgroup of $J_0(N)$. This enables us to make the following definition:

Definition 4.22. Let f be a newform of level N_f. The *Abelian variety associated to* f is

$$A_f = J_0(N_f)/I_f J_0(N_f).$$

As with the Jacobian, this new structure is essentially all we need to state another version of the modularity theorem. However, in order to state that there exists a holomorphism from A_f we want to give it a complex structure. As with the Jacobian, this will require us to state some preliminary results without proof. The missing proofs can be found in [1].

Proposition 4.23. *Let $f = \sum_{n=1}^\infty a_n(f)q^n$ be a normalized eigenform. For all $n \in \mathbb{N}$ there is some $T_n \in \mathbb{T}_\mathbb{Z}$ such that $T_n f = a_n(f)f$. If $n = p$ is prime, this is just T_p.*

Proposition 4.24. *Let $f = \sum_{n=1}^\infty a_n q^n$ be a newform of level N_f, and let $\mathbb{K}_f = \mathbb{Q}(\{a_n\})$. For any embedding $\sigma : \mathbb{K}_f \to \mathbb{C}$ the function*

$$f^\sigma = \sum_{n=1}^\infty \sigma(a_n)q^n$$

is also a newform of level N_f.

This allows us to prove the following lemma:

Lemma 4.25. *Let $\mathbb{T}_\mathbb{C}$ be defined the same way as $\mathbb{T}_\mathbb{Z}$, just over $\mathbb{C}$. Then $\mathbb{T}_\mathbb{C}^\wedge \cong \mathcal{S}_k(\Gamma_0(N))$.*

Proof. Consider the map $F : \mathbb{T}_\mathbb{C} \times \mathcal{S}_k(\Gamma_0(N)) \to \mathbb{C}$ which sends (T, f) to $a_1(Tf)$, where $a_1 : \mathcal{S}_k(\Gamma_0(N)) \to \mathbb{C}$ sends a cusp form to the coefficient on the first term in its Fourier development. Then F is bilinear. To turn this map into the desired isomorphism, we also need to show that it is non-degenerate in both terms.

Let $T \in \mathbb{T}_\mathbb{C}$ and suppose $F(T, f) = 0$ for all $f \in \mathcal{S}_k(\Gamma_0(N))$. Applying this to $T_n f$ and using Proposition 4.23 we have

$$0 = a_1(TT_n f) = a_1(T_n(Tf)) = a_n(Tf) \implies Tf = 0$$

and since this is true for all $f \in \mathcal{S}_k(\Gamma_0(N))$, indeed $T = 0$. Conversely let $f \in \mathcal{S}_k(\Gamma_0(N)$ and suppose $F(T, f) = 0$ for all $T \in \mathbb{T}_\mathbb{C}$. In particular, $T_n \in \mathbb{T}_\mathbb{C}$ so $a_n(f) = a_1(T_n f) = 0$ and indeed $f = 0$.

Therefore the maps from $\mathcal{S}_k(\Gamma_0(N))$ to $\mathbb{T}_\mathbb{C}^\wedge$ sending f to $(T \mapsto F(T, f))$ and from $\mathbb{T}_\mathbb{C}$ to $\mathcal{S}_k(\Gamma_0(N))^\wedge$ sending T to $(f \mapsto F(T, f))$ are both linear and injective. From Corollary 3.5 and Proposition 3.9 we know that $\dim(\mathcal{S}_k(\Gamma_0(N))^\wedge) = \dim(\mathcal{S}_k(\Gamma_0(N)))$ are both finite, so the existence of both of these maps combines to show that

$$\dim(\mathbb{T}_\mathbb{C}) = \dim(\mathbb{T}_\mathbb{C}^\wedge) = \dim(\mathcal{S}(\Gamma_0(N))) = \dim(\mathcal{S}(\Gamma_0(N))^\wedge).$$

In particular, since all dimensions are the same, these two injections become bijections, and so $\mathbb{T}_\mathbb{C}^\wedge \cong \mathcal{S}_k(\Gamma_0(N))$. □

We will also need a few unproven facts from algebra more generally, in particular the tensor product (for more details on the tensor product than are given here, see [8]). Let R be a ring and let N and M be R-modules. Recall that $N \otimes_R M$ is the unique R-module equipped with a R-bilinear map $b : N \times M \to N \otimes_R M$ such that, for any R-module Q and any R-bilinear map f from $M \times N$ to Q, f can be written as the composite of b and a linear map from $M \otimes_R N$ to Q.

We will only consider the case where N and M are $\mathbb{Z}$-modules, i.e. Abelian groups, and so we will drop the $\mathbb{Z}$ subscript. If either N or M is also a module for some other ring A, then $N \otimes_R M$ is naturally an A-module as well, by the rule

$$a\left(\sum_i n_i \otimes m_i\right) = \sum_i (an_i) \otimes m_i$$

or similarly if M is the A-module.

In particular, let $\mathbf{k}$ be a field and G be a finitely generated Abelian group. Then $\mathbf{k}$ is certainly a $\mathbf{k}$-module, and thus so too is $G \otimes \mathbf{k}$, or in other words a vector space over $\mathbf{k}$. We will need the following properties of the tensor product in this scenario:

Lemma 4.26. *Let $\mathbf{k}$ be a field with characteristic 0 and let G be a finitely generated Abelian group. Then*

1. *$G \otimes \mathbf{k} \cong \mathbf{k}^{\mathrm{rank}(G)}$.*
2. *For any subgroup $K \leq G$ we have $(G/K) \otimes \mathbf{k} \cong (G \otimes \mathbf{k})/(K \otimes \mathbf{k})$.*
3. *Let A be a ring with an ideal J and suppose G is also an A-module. Then $JG \otimes \mathbf{k} \cong J(G \otimes \mathbf{k})$.*

Lemma 4.27. *Let A be a ring, J an ideal in A, and let M be both an A-module and vector space over some field $\mathbf{k}$. Let $M[J]$ denote the elements of M annihilated by J. The dual space of M is is naturally an A-module as well and*

$$M^\wedge / JM^\wedge \cong M[J]^\wedge$$

as A-modules. The isomorphism is given by the restriction map $\varphi + JM^\wedge \mapsto \varphi|_{M[J]}$.

All of this allows us to prove the following two results:

Proposition 4.28. *Let f be a newform and let*

$$V_f = \mathit{span}(\{f^\sigma : \sigma \text{ is an embedding of } \mathbb{K}_f \text{ into } \mathbb{C}\}) \subset \mathcal{S}_k(\Gamma_0(N_f))$$

and Λ_f be the restriction of $H_1(X_0(N_f), \mathbb{Z})$ to V_f. Then

$$A_f \cong V_f^\wedge / \Lambda_f.$$

Proof. Here and through the rest of this section let $\mathcal{S}_2 = \mathcal{S}_2(\Gamma_0(N_f)$ and let $H_1 = H_1(X(\Gamma_0(N), \mathbb{Z})$. Let π be the projection map from $\mathcal{S}_2^\wedge$ to $\mathcal{S}_2^\wedge / I_f \mathcal{S}_2^\wedge$. Then

$$A_f = (\mathcal{S}_2^\wedge / H_1)/I_f(S_2^\wedge / H_1) \cong (\mathcal{S}_2^\wedge / I_f \mathcal{S}_2^\wedge)/\pi(H_1)$$

(for this last equivalence one can check that $(\varphi + H_1) + I_f(S_2^\wedge / H_1) \mapsto (\varphi + I_f \mathcal{S}_2^\wedge) + \pi(H_1)$ is an isomorphism). We then apply the isomorphism from Lemma 4.27 to take $\mathcal{S}_2^\wedge / I_f \mathcal{S}_2^\wedge$ to $\mathcal{S}_2[I_f]^\wedge$. Each $\pi(\varphi) = \varphi + I_f \mathcal{S}_2^\wedge \in \pi(H_1)$ gets mapped to $\varphi|_{\mathcal{S}_2[I_f]}$, so the combination of π and this isomorphism simply restricts each function in H_1, sending H_1 to $H_1|_{\mathcal{S}_2[I_f]}$. Thus

$$A_f \cong \mathcal{S}_2[I_f]^\wedge / H_1|_{\mathcal{S}_2[I_f]}$$

so it suffices to show that $\mathcal{S}_2[I_f] = V_f$. In other words, the forms annihilated by I_f are exactly f and its conjugates (and their sums).

For any embedding $\sigma : \mathbb{K}_f \to \mathbb{C}$ we want to show that f^σ is annihilated by I_f. Let $T \in I_f$, i.e. $\lambda_f(T) = 0$. Then T is some $\mathbb{Z}$-linear sum of products of the operators T_n which sends f to 0. In other words, by Proposition 4.23, it is a $\mathbb{Z}$-linear sum of products of Fourier coefficients of f which equals 0. Then $T(f^\sigma)$ is this same $\mathbb{Z}$-linear sum of products of the coefficients $\sigma(a_n)$. Since σ is an embedding, it factors through products, sums, and multiplication by integers, so we get

$$T(f^\sigma) = \sigma(\lambda_f(T))f^\sigma = \sigma(0)f^\sigma = 0.$$

This shows that $V_f \subset \mathcal{S}_2[I_f]$. To show equality, it now suffices to show that $\dim(\mathcal{S}_2[I_f]) \leq \dim(V_f)$.

By Proposition 4.23, we know that the image of $\lambda_f : \mathbb{T}_\mathbb{Z} \to \mathbb{C}$ is precisely $\mathbb{Z}[\{a_n\}]$ and so

$$\mathbb{Z}[\{a_n\}] \cong \mathbb{T}_\mathbb{Z}/I_f.$$

By Corollary 4.11, we can consider $\mathbb{T}_\mathbb{Z}$ as a ring of endomorphisms of H_1. By Proposition 3.9 we know that H_1 has finite rank over $\mathbb{Z}$, and therefore so does its ring of endomorphisms, of which $\mathbb{T}_\mathbb{Z}$ is a subring. Thus $\mathbb{Z}[\{a_n\}]$ has finite rank, and so $\mathbb{K}_f$ has finite degree over $\mathbb{Q}$. In the language of algebraic number theory, it is a number field. It is a fact from algebraic number theory that the number of embeddings of a number field $\mathbb{K}_f$ into $\mathbb{C}$ is $[\mathbb{K}_f : \mathbb{Q}]$. The newforms f^σ are all linearly independent so

$$\dim(V_f) = [\mathbb{K}_f : \mathbb{Q}] = \operatorname{rank}(\mathbb{T}_\mathbb{Z}/I_f).$$

Now consider the surjection from $\mathbb{T}_\mathbb{Z} \otimes \mathbb{C}$ to $\mathbb{T}_\mathbb{C}$ given by

$$\sum_i U_i \otimes z_i \mapsto \sum_i z_i U_i$$

which is well-defined by the basic properties of the tensor product. The image of $I_f \otimes \mathbb{C}$ are elements of the form $\sum_i z_i U_i$ for $U_i \in I_f$, which is

$$\sum_i z_i U_i = \sum_i (z_i T_1) U_i = \sum_i U_i (z_i T_1) \in I_f \mathbb{T}_\mathbb{C}.$$

Thus the image of $I_f \otimes \mathbb{C}$ is a subgroup of $I_f \mathbb{T}_\mathbb{C}$, so the induced map from $(\mathbb{T}_\mathbb{Z} \otimes \mathbb{C})/(I_f \otimes \mathbb{C})$ to $\mathbb{T}_\mathbb{C}/I_f\mathbb{T}_\mathbb{C}$ is also a surjection. By Lemma 4.25,

$$\dim(\mathcal{S}_2[I_f]) = \dim(\mathcal{S}_2^\wedge/I_f\mathcal{S}_2^\wedge) = \dim(\mathbb{T}_\mathbb{C}/I_f\mathbb{T}_\mathbb{C})$$

and then using the above surjection and the properties from Lemma 4.26,

$$\begin{aligned}\dim(\mathbb{T}_\mathbb{C}/I_f\mathbb{T}_\mathbb{C}) &\leq \dim((\mathbb{T}_\mathbb{Z} \otimes \mathbb{C})/(I_f \otimes \mathbb{C})) \\ &= \dim(\mathbb{T}_\mathbb{Z}/I_f \otimes \mathbb{C}) \\ &= \operatorname{rank}(\mathbb{T}_\mathbb{Z}/I_f) \\ &= [\mathbb{K}_f : \mathbb{Q}]\end{aligned}$$

and so indeed $\dim(\mathcal{S}_2[I_f]) \leq \dim(V_f)$, and this completes the proof. □

Proposition 4.29. *Λ_f is a lattice in $V_f^\wedge$.*

Proof. Since $V_f^\wedge$ is a finite dimensional vector space over $\mathbb{R}$ and Λ_f is a finitely generated ring over $\mathbb{Z}$, to show that Λ_f is a lattice, it suffices to show that a minimal set of generators for Λ_f is also a basis for $V_f^\wedge$. It in turn suffices to show that the $\mathbb{R}$-span of Λ_f is $V_f^\wedge$ and that $\operatorname{rank}(\Lambda_f) \leq \dim_{\mathbb{R}}(V_f^\wedge)$.

Since $V_f \subset \mathcal{S}_2$, the restriction map $\rho : \mathcal{S}_2^\wedge \to V_f^\wedge$ is a surjection. Using the fact that we already know H_1 is a lattice, we have

$$\mathbb{R}\Lambda_f = \mathbb{R}H_1|_{V_f} = \mathbb{R}\rho(H_1) = \rho(\mathbb{R}H_1) = \rho(\mathcal{S}_2^\wedge) = V_f^\wedge.$$

To show that $\operatorname{rank}(H_1) \leq \dim(V_f^\wedge)$, first consider the linear map from H_1/I_fH_1 to Λ_f sending $\varphi + I_fH_1$ to $\varphi|_{V_f}$. To show this map is well-defined, we have to show that for any $\varphi \in I_fH_1$, $\varphi|_{V_f} = \rho(\varphi)$ is zero on V_f. Given $T \in I_f$ and $\varphi \in H_1$, let $T\varphi$ be a basis element of I_fH_1. Then for every $f \in V_f = \mathcal{S}_2[I_f]$ (see previous proof) we have $T(f) = 0$ and so $T\varphi(f) = (\varphi \circ T)(f) = \varphi(0) = 0$, as desired. Furthermore

$$\Lambda_f = \rho(H_1) \implies \Lambda_f \cong H_1/\ker(\rho|_{H_1}) = H_1/(\ker(\rho) \cap H_1).$$

We just showed that $I_fH_1 \subset \ker(\rho)$, and by definition $I_fH_1 \subset H_1$, so the map from H_1/I_fH_1 to $H_1/(\ker(\rho) \cap H_1) \cong \Lambda_f$ is a surjection. Using this surjection and Lemma 4.26 we have

$$\begin{aligned}\operatorname{rank}(\Lambda_f) &\leq \operatorname{rank}(H_1/I_fH_1)\\ &= \dim_{\mathbb{R}}(H_1/I_fH_1 \otimes \mathbb{R})\\ &= \dim_{\mathbb{R}}((H_1 \otimes \mathbb{R})/I_f(H_1 \otimes \mathbb{R})).\end{aligned}$$

Since H_1 is a free and finitely generated $\mathbb{Z}$-module (by Proposition 3.9), the surjection from $H_1 \otimes \mathbb{R}$ to $\mathcal{S}_2^\wedge$ sending $\sum_i \varphi_i \otimes x_i$ to $\sum_i x_i\varphi_i$ is an isomorphism. To see this, let $\varphi_1, \ldots, \varphi_d$ be minimal set of generators for H_1, which is also a basis for $\mathcal{S}_2^\wedge$. By collecting like terms we can write any element of $H_1 \otimes \mathbb{R}$ as $\sum_{i=1}^d \varphi_i \otimes x_i$, which is in bijection with arbitrary elements of $\mathcal{S}_2^\wedge$ which can be written $\sum_{i=1}^d x_i\varphi_i$. Thus

$$\dim_{\mathbb{R}}((H_1 \otimes \mathbb{R})/I_f(H_1 \otimes \mathbb{R})) = \dim_{\mathbb{R}}(\mathcal{S}_2^\wedge/I_f\mathcal{S}_2^\wedge) = \dim_{\mathbb{R}}(\mathcal{S}_2[I_f]^\wedge) = \dim_{\mathbb{R}}(V_f^\wedge)$$

and this completes the proof. □

The above two results combine to show that A_f is a complex torus, as desired.

4.6. **The Modularity Theorem.**

Theorem 4.30 (Modularity Theorem, Version III)**.** *For every complex elliptic curve E with $j(E) \in \mathbb{Q}$, there is some $N \in \mathbb{N}$ and a newform $f \in \mathcal{S}_2(\Gamma_0(N))$ such that there exists a surjective holomorphic homomorphism of complex tori from A_f to E.*

To show that version (III) is equivalent to version (II) (and so also version (I)), we need some preliminary results:

Proposition 4.31. *Let $\mathbb{C}^g/\Lambda_g$ and $\mathbb{C}^h/\Lambda_h$ be complex tori. A surjective holomorphic homomorphism $\varphi : \mathbb{C}^g/\Lambda_g \to \mathbb{C}^h/\Lambda_h$ is called an* isogeny *if it has finite kernel. If such an isogeny exists, then there also exists an isogeny in the other direction, $\hat{\varphi} : \mathbb{C}^h/\Lambda_h \to \mathbb{C}^g/\Lambda_g$.*

Proof (sketch). Using Lemma 3.13 we can show that such an isogeny can only exist if $h = g$, and then

$$\varphi(z + \Lambda_g) = Mz + \Lambda_h$$

where M is an invertible $g \times g$ matrix such that $M\Lambda_g \subset \Lambda_h$. This means that there exists some basis $\{\omega_1, \ldots, \omega_{2g}\}$ of Λ_h and non-zero integers $n_1, \ldots, n_{2g}$ such that $\{n_1\omega_1, \ldots, n_{2g}\omega_{2g}\}$ is a basis for $M\Lambda_g$. Thus $(n_1 n_2 \ldots n_{2g})\Lambda_h \subset M\Lambda_g$ and in particular

$$(n_1 n_2 \ldots n_{2g})M^{-1}\Lambda_h \subset \Lambda_g$$

so the matrix $(n_1 n_2 \ldots n_{2g})M^{-1}$ gives the isogeny in the reverse direction. □

Proposition 4.32. *Let f, g be newforms of the same level. If there exists an embedding $\sigma : \mathbb{K}_f \hookrightarrow \mathbb{C}$ such that $f^\sigma = g$, then there exists an embedding $\sigma' : \mathbb{K}_g \hookrightarrow \mathbb{C}$ such that $f = g^{\sigma'}$, and so conjugation by embeddings in an equivalence relation on newforms. Furthermore, in this case $A_f = A_g$.*

Proof. To find σ', we restrict the codomain of σ to its image and take the inverse. In Proposition 4.28 we showed that I_f annihilates $f^\sigma = g$, so $I_f \subset I_g$. Applying the same logic to g and σ' shows the other inclusion, so $I_f = I_g$ and indeed $A_f = A_g$. □

Recall our unproven assertion that the newforms of level N form a basis for the subspace of newforms of level N. This result can be further extended to all of $\mathcal{S}_k(\Gamma_0(N))$:

Proposition 4.33. *Let $N \in \mathbb{N}$. For all $n \in \mathbb{N}$ let $\alpha_n = \left[\begin{smallmatrix} n & 0 \\ 0 & 1 \end{smallmatrix}\right]$. Then*

$$\{f[\alpha_n]_k : f \text{ is a newform of level } M \text{ and } nM|N\}$$

is a basis of $\mathcal{S}_k(\Gamma_0(N))$.

Proof (sketch). Taking as given that the newforms of level M form a basis for $\mathcal{S}_k(\Gamma_0(M))^{\text{new}}$, the fact that this set spans $\mathcal{S}_k(\Gamma_0(N))$ follows almost immediately from the definition of oldforms in Definition 4.17. We omit the proof that these forms are linearly independent, although the fact that they span and so contain a basis is sufficient to prove a weakened version of the following theorem which still shows the equivalence of versions (II) and (III) of the modularity theorem. □

Theorem 4.34. *Let $[f]$ denote an equivalence class of newforms from the previous proposition, and let $d : \mathbb{N} \to \mathbb{N}$ be the number of divisors function. Then there is an isogeny*

$$J_0(N) \to \bigoplus_{M|N} \bigoplus_{[f]} A_f^{d(N/M)}.$$

where the second direct sum is over equivalence classes of newforms of level M.

Proof. Throughout this proof we will write $\bigoplus$ to mean $\bigoplus_{M|N} \bigoplus_{[f]}$.

We can rewrite the basis from Proposition 4.33 as

$$\bigcup_{M|N} \bigcup_{\substack{\text{newforms } f \\ \text{of level } M}} \bigcup_{n|N/M} f[\alpha_n]_2 = \bigcup_{M|N} \bigcup_{[f]} \bigcup_{n} \bigcup_{\sigma} f^\sigma[\alpha_n]_2 \tag{4.35}$$

where the second union is over equivalence classes of newforms of level M, the third over divisors of N/M, and the last over embeddings of $\mathbb{K}_f$ into $\mathbb{C}$. We now define an explicit isomorphism

$$\Psi : \bigoplus V_f^{d(N/M)} \to \mathcal{S}_2 \tag{4.36}$$

The space V_f has basis $\{f^\sigma : \sigma \text{ is an embedding from } \mathbb{K}_f \text{ to } \mathbb{C}\}$. On the left hand side of (4.36), if $[f]$ is an equivalence class of newforms of level M then we have one copy of V_f for each divisor n of N/M. Thus we send f^σ in the copy of V_f corresponding to n to $f^\sigma[\alpha_n]_2$, a basis element from (4.35). Linearly extending this one to one mapping of bases, we get an isomorphism.

Restricting Ψ to a particular copy of V_f, we get $\Psi|_{V_f} = [\alpha_n]_2$ where n is some divisor of N/M for M the level of f. Thus the dual of Ψ is the direct product of the dual of these operators:

$$\Psi^\wedge = \prod_{M|N} \prod_{n|N/M} [\alpha_n]_2^\wedge : \mathcal{S}_2^\wedge \to \bigoplus (V_f^\wedge)^{d(N/M)}$$

We showed in Proposition 4.10 that the maps $[\alpha_n]_2^\wedge$ take integration over loops to integration over loops, so each map takes H_1 into H_1 restricted to V_f, i.e. Λ_f. Piecing these functions together, we have

$$\Psi^\wedge(H_1) \subset \bigoplus \Lambda_f^{d(M/N)}.$$

Thus $\Psi^\wedge$ descends to a surjection

$$\mathcal{S}_2^\wedge/H_1 \to \left(\bigoplus (V_f^\wedge)^{d(M/N)}\right) / \left(\bigoplus \Lambda_f^{d(M/N)}\right) \cong \bigoplus (V_f^\wedge/\Lambda_f)^{d(M/N)}. \tag{4.37}$$

To show this is an isogeny, we have to show that it has finite kernel, which is equivalent to saying $\Psi^\wedge(H_1)$ has finite index in $\bigoplus \Lambda_f^{d(M/N)}$. Since they are both finitely generated Abelian groups, it suffices to show that they have the same rank. By Proposition 4.29 we know that $\operatorname{rank}(\Lambda_f) = \dim_{\mathbb{R}}(V_f^\wedge)$. Since $\Psi^\wedge$ is an isomorphism it preserves dimension and rank, so

$$\begin{aligned}
\operatorname{rank}(\Psi^\wedge(H_1)) = \operatorname{rank}(H_1) = \dim_{\mathbb{R}}(\mathcal{S}_2^\wedge) &= \dim_{\mathbb{R}}\left(\bigoplus (V_f^\wedge)^{d(N/M)}\right) \\
&= \sum_{M|N} \sum_{[f]} d(N/M) \dim_{\mathbb{R}}(V_f^\wedge) \\
&= \sum_{M|N} \sum_{[f]} d(N/M) \operatorname{rank}(\Lambda_f) \\
&= \operatorname{rank}\left(\bigoplus \Lambda_f^{d(M/N)}\right)
\end{aligned}$$

Thus the surjection in (4.37) is indeed an isogeny, and applying the definition of the Jacobian and Proposition 4.28 we immediately get an isogeny

$$J_0(N) \to \bigoplus A_f^{d(N/M)}$$

as desired. □

Proposition 4.38. *Versions (II) and (III) of the modularity theorem are equivalent.*

Proof. Suppose version (III), so there is some $N \in \mathbb{N}$ and a newform $f \in \mathcal{S}_2(\Gamma_0(N))$ such that there exists a surjective holomorphic homomorphism from A_f to E. Composing with the map from Theorem 4.34, and sending all varieties other than A_f to zero, we get a surjective holomorphic homomorphism

$$J_0(N) \to \bigoplus_{M|N} \bigoplus_{[f]} A_f^{d(N/M)} \to E$$

(or we can replace $J_0(N)$ with $J_0(K)$ for any K such that $N|K$). This is the map conjectured by version (II).

Now suppose version (II), so there is some $N \in \mathbb{N}$ such that there exists a surjective holomorphic homomorphism from $J_0(N)$ to E. Let

$$\varphi : \bigoplus_{M|N} \bigoplus_{[f]} A_f^{d(N/M)} \to J_0(N)$$

be the isogeny in the reverse direction (which we know exists from Proposition 4.32). For each A_f in the direct sum, $\varphi|_{A_f}$ is a holomorphic homomorphism of compact Riemann surfaces, so it must either be surjective or constant. If all these restrictions were constant, then φ could not be surjective, a contradiction. So there must exist some newform f of level M (where $M|N$) such that $\varphi|_{A_f}$ surjects A_f onto $J_0(N)$, and then the map from version (II) surjects $J_0(N)$ onto E. □

Acknowledgements

I would like to of course thank my mentor Minchan Kang for suggesting this topic and giving me invaluable help throughout. I would also like to thank Peter May for organizing the REU, Jeff Harvey for giving a lecture related to modular forms which showed me just how fascinating they are (and pointing me towards additional resources), and Matt Emerton for giving me advice before the REU even got started. I am of course always thankful to my friends and family, other than anyone who has participated in robbing me of a pencil.

References

[1] Fred Diamond and Jerry Shurman. A First Course in Modular Forms. Springer. 2005.
[2] Hershel M. Farkas and Irwin Kra. Riemann Surfaces. Springer. 1992.
[3] Jonathan Evans. 7.02 Path-lifting, monodromy. https://www.homepages.ucl.ac.uk/~ucahjde/tg/html/cov-02.html.
[4] Tom Apostol. Mathematical Analysis. Pearson. 1973.
[5] Shiyue Li. Lecture Notes on Modular Forms. https://www.shiyue.li/mathcamp/shiyueli-mathcamp-modular-forms.pdf.
[6] Corrin Clarkson. Riemann Surfaces. https://www.math.uchicago.edu/~may/VIGRE/VIGRE2007/REUPapers/FINALFULL/Clarkson.pdf.
[7] Benjamin Church Elliptic Curves, Complex Tori, Modular Forms, and ℓ-adic Galois Representations. https://web.stanford.edu/~bvchurch/assets/files/ell_curves/Notes.pdf.
[8] Keith Conrad. Tensor Products. https://kconrad.math.uconn.edu/blurbs/linmultialg/tensorprod.pdf.
[9] The 1-2-3 of Modular Forms. Jan Hendrik Bruinier, Gerard van der Geer, Günter Harder, and Don Zagier. Springer. 2008.