Modular Forms

Jacob Denson

October 29, 2018

Table Of Contents

Chapter 1

Modular Forms

Modular forms is an interesting field where one meets deep ideas, intwined with the concrete mathematics of combinatorics and number theory. At the core, the theory of modular forms the relation between two ideas:

- Functions on the Poincaré model of the hyperbolic plane, which are semi-invariant under a discrete group of isometries of the plane, describable as Möbius transformations.
- Certain infinite series, known as *q-expansions*, of the form

$$\sum_{k=-\infty}^{\infty} a_k q^k$$

  where $q = e^{2\pi i z}$ connects the convergence of the series in $\mathbf{D}$ to a function in the upper half plane.

The first viewpoint gives us deep geometric insights, whereas the second gives us very useful applications in number theory, combinatorics, and particle physics. For instance, these forms occur when counting curves in algebraic geometry, or when understanding the dimensions of finite simple groups. Because they occur so often in mathematics, the number theorist Joseph Eichler has been told to have said that there are five fundamental arithmetical operations – addition, subtraction, multiplication, division, and modular forms. We begin by examining in detail the symmetries which characterize modular forms, and then we explore how modular forms arise, and the most fundamental examples.

1.1 Geometry of the Upper Half Plane

The nicest holomorphic functions on the Riemann sphere $\mathbf{C}_\infty$ are the Möbius transformations, obtained via an action from $GL_2(\mathbf{C})$ by the map

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}(z) = \frac{az+b}{cz+d}$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}(\infty) = \lim_{z\to\infty} \frac{az+b}{cz+d} = \frac{a}{c}$$

The restriction that the matrix is invertible is required so that the action is invertible, and if the matrix wasn't invertible, the transformation would reduce to a constant. The kernel of this representation is the multiplicative group of complex numbers $\mathbf{C}^\times$. Indeed, the constraint that $(az+b)(cz+d)^{-1} = z$ for all z can be expressed in terms of the triviality of the polynomial $cz^2 + (d-a)z - b$, so that $c = b = d - a = 0$. Performing a quotient, we obtain the group $PGL_2(\mathbf{C}) = GL_2(\mathbf{C})/K$, which now acts faithfully on $\mathbf{C}_\infty$. It is a scale invariant version of the general linear group. Since every invertible matrix can be normalized to have determinant one, the Möbius transformations can also be described as $PSL_2(\mathbf{C}) = SL_2(\mathbf{C})/\{\pm I\}$.

The Möbius transformations preserving the projective real line are precisely those transformations $PGL_2(\mathbf{R})$ induced by an element of $GL_2(\mathbf{R})$. These transformations split into two connected components, those with positive determinant, which can be described as the transformations preserving the upper and lower half plane, and those with negative determinant, which map the upper half plane into the lower half plane, and vice versa. The elements $PSL_2(\mathbf{R}) = SL_2(\mathbf{R})/\{\pm I\}$ are those which are of most interest to us, since they form the family of hyperbolic isometries on the upper half plane, if this plane is given the Poincaré metric. The field of modular forms studies the actions of certain **Fuschian groups**, discrete subgroups of $PSL_2(\mathbf{R})$, relating to number theory. Most notably, we study the **modular group**

$$\Gamma = PSL_2(\mathbf{Z}) = SL_2(\mathbf{Z})/\{\pm I\}$$

The name comes from the theory of moduli spaces. Clearly Γ is discrete, because it is the quotient of a discrete group by a discrete subgroup. Indeed, in the sequel it will be important to note the following result.

Theorem 1.1. *Let G be a locally compact group acting transitively on a topological space X, such that the stabilizer of one, and hence every point $x \in X$ is a compact set K, and the induced map $G/K \to X$ is a homeomorphism. Then a subgroup Γ of G is discrete if and only if for every pair of compact sets A and B, there are finitely many elements $g \in \Gamma$ such that $gA \cap B$ is nonempty.*

Proof. Because of the homeomorphism, we may view X as being precisely G/K, with the canonical group action. Suppose that there exists infinitely many $g_\alpha \in \Gamma$, with corresponding $a_\alpha \in A$ such that $g_\alpha a_\alpha K = b_\alpha K$, where $b_\alpha \in B$. But this means that $b_\alpha^{-1} g_\alpha a_\alpha \in K$. Since A and B, and K are compact, we can choose a subsequence of elements such that $a_\alpha \to a \in A$, and $b_\alpha \to b \in B$, and $b_\alpha^{-1} g_\alpha a_\alpha \to k \in K$. Then $g_\alpha \to g = b^{-1}ka$. If Γ was discrete, then we would have $g_\alpha = g$ for sufficiently large α. But this is impossible since we assumed the g_α were distinct. Conversely, if for each compact A there are only finitely many $g \in \Gamma$ such that $gA \cap B$ is non-empty, then we can select a precompact neighbourhood U of the origin, and then there are only finitely many elements g such that $gU \cap U$ is non-empty. Selecting a neighbourhood V of the origin such that no non-trivial g satisfy $gV \cap V \neq \varnothing$, we find that $gV \cap hV$ are disjoint for distinct $g, h \in \Gamma$, so Γ is discrete. □

Corollary 1.2. *If G, K, and X are as in the last theorem, and Γ is a discrete subgroup of G, then for each $x \in X$, Γ_x is finite, there is a neighbourhood U of x such that if $U \cap gU \neq \varnothing$, then $gx = x$, and for any $x, y \in X$ which are not in the same orbit of Γ, there exists neighbourhoods U of x and V of y such that $gU \cap V = \varnothing$ for all $g \in \Gamma$.*

Corollary 1.3. *The space X/Γ is Hausdorff.*

Note that the conditions of the theorem also imply that the cosets induced from a discrete subgroup by a normal compact subgroup also form a discrete set in the quotient topology. The condition that G/K is isomorphic to X is trivial in most situations where the objects involved are suitably complete.

Theorem 1.4. *If G acts transtively on X, both spaces are locally compact, and G is first countable, then the natural map $G/K \to X$ is a homeomorphism.*

Proof. The natural map is obviously a continuous bijection, and it now suffices to verify it is open. Let U be an open subset of G containing some

$g \in G$. It will suffice to verify that gx is an interior point of gU. Consider a precompact symmetric neighbourhood V of the origin with $V^2 \subset U$, and consider a cover of G by the translates gV, and consider a countable basis $W_1, W_2, \ldots$ consisting only of sets contained in some gV. Fix g_n such that $W_n \in g_n V$. As $g_n V$ is precompact, so too is $g_n V x$ in X. Since X is Hausdorff, $g_n V x$ is closed. The Baire category theory implies that since $X = \bigcup g_n V x$, one of the sets $g_n V x$ has non-empty interior, which implies that Vx itself has non-empty interior. □

Example. *We know $PSL_2(\mathbf{R})$ acts transitively on* $\mathbf{H}$. *The stabilizer of i is $PSO_2(\mathbf{R})$, because if $(ai+b) = i(ci+d) = di - c$, then $a = d$ and $b = -c$, which means that the matrix induced by the coefficients is precisely a rotation matrix. Thus* $\mathbf{H}$ *is homeomorphic to $PSL_2(\mathbf{R})/PSO_2(\mathbf{R}) \cong SL_2(\mathbf{R})/SO_2(\mathbf{R})$*

Other important discrete groups studied in the theory of modular forms are the **principal congruence subgroups**, which are the kernels $\Gamma(N)$ of the natural maps from $PSL_2(\mathbf{Z})$ to $PSL_2(\mathbf{Z}_N)$, for an integer N, i.e.

$$\Gamma(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in PSL_2(\mathbf{Z}) : ad - bc = 1, b \equiv c \equiv 0 \pmod{N} \right\}$$

We actually have $\Gamma/\Gamma(N) \cong PSL_2(\mathbf{Z}_N)$, since the map $SL_2(\mathbf{Z}) \to SL_2(\mathbf{Z}_N)$ is surjective. A **congruence subgroup** is a subgroup of Γ containing $\Gamma(N)$ for some N. Two important examples are

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in PSL_2(\mathbf{Z}) : c \equiv 0 \bmod \mathrm{N} \right\}$$

$$\Gamma_1(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma : a \equiv d \equiv 1 \bmod \mathrm{N} \right\}$$

One of the most important properties of congruence subgroups is that they have finite index in Γ, which becomes important when we analyze the orbit spaces of $\mathbf{H}$ with respect to the actions of these groups.

1.2 Classification of Möbius Transformations

We begin by discussing the degrees of freedom the the group of Möbius transformations possess. We know from projective geometry that for any

three points in the complex projective line, there is a projective transformation mapping a triple of three points to another triple of three points. For the transformations preserving the upper half plane, we have slightly less freedom. The space of transformations preserving this plane has three real dimensions, so we can expect that we have enough degrees of freedom to map 'one and a half points' to any other pair of 'one and a half points'. This manifests itself as mapping any point in the upper half plane and any point on the boundary to any other point in the upper half plane and point on the boundary. This is done by forming the unique circle orthogonal to the real line passing through the two points, taking the third point on the real line, and then mapping the third point to the third point obtainable in the same manner in the other description. Since Möbius transformations preserve infinitismal angles, one verifies that the result Möbius transformation preserves the real line.

Now we classify the Möbius transformation. By the classification of Jordan normal forms, every Möbius transformation is representable by an element of $SL_2(\mathbf{C})$, and therefore equivalent to one of

$$\begin{pmatrix} 1/\lambda & 1 \\ 0 & 1/\lambda \end{pmatrix} \quad \begin{pmatrix} \lambda & 0 \\ 0 & 1/\lambda \end{pmatrix}$$

with $\lambda^2 = 1$ in the first case, and $\lambda \neq 1$ in the latter case. This means that by a projective change of coordinates, every projective transformation is equivalent either to the map $z \mapsto z + \lambda$, and in the second, a family of maps of the form $z \mapsto \alpha z$, with $\alpha = \lambda^2$. The translation transformations are known as **parabolic**, and the latter either **elliptic**, if $|\alpha| = 1$, **hyperbolic** if α is real and positive, and **loxadromic** afterwards.

One can quickly determine which class a transformation belongs to by looking at the trace of the matrix representation in $SL_2(\mathbf{C})$, which is determined up to multiplication by ± 1:

1. If $\mathrm{Tr}(M) = \pm 2$, then the transform is parabolic.
2. If $\mathrm{Tr}(M)$ is real and $|\mathrm{Tr}(M)| < 2$, then M is elliptic.
3. If $\mathrm{Tr}(M)$ is real and $|\mathrm{Tr}(M)| > 2$, then M is hyperbolic.
4. If $\mathrm{Tr}(M)$ is not real, then M is loxodromic.

One can also see which class a transformation belongs to by looking at geometric properties. The parabolic transformations are precisely the transformations which fix a single point, and can be viewed as a degenerate case of the transformation. In the case of an elliptic transformation, we have two fixed points, and the non fixed points 'orbit' about these fixed points. If the transformation is hyperbolic or loxodromic, then we have two fixed points x_0 and x_1, with x_0 an attractor and x_1 a repeller, and if we consider the one parameter semigroup induced by M, then points x flow to x_0 over time for all $x \neq x_1$. The number α represents both the speed to which the points are attracted, and the amount of rotation in the transformation. In the case of a hyperbolic transformation, there is no rotational factor, so points travel along circular arcs to x_0 rather than logarithmic spirals.

In the case where we take an element of $PSL_2(\mathbf{R})$, we can conclude more about the geometric properties of the transform. Firstly, suppose an element of $PSL_2(\mathbf{R})$ are parabolic. Then the eigenvalue of the matrix is real, so the fixed point of the transformation lies on the projective real line. If the fixed point is ∞, the transformation is of the form $z \mapsto z + t$ for some $t \in \mathbf{R}$. Otherwise, we have a fixed point on the real line, and points travel on circles tangent to the real line, converging to the fixed point if we consider the one parameter semigroup flow. In the case of an element of $PSL_2(\mathbf{R})$ with two fixed points, we know either we have two real eigenvalues, or the eigenvalues occur as complex conjugates of one another. Regardless, the trace is real, so no elements of $PSL_2(\mathbf{R})$ are loxodromic. A neat fact is that the characteristic polynomial of a matrix $M \in SL_2(\mathbf{R})$ is $X^2 - \mathrm{Tr}(M)X + 1$, and the discriminant of the polynomial is $\mathrm{Tr}(M)^2 - 4$. Thus we have two real eigenvalues, and thus two fixed points, precisely when the map is hyperbolic, and two complex conjugate eigenvalues, and thus two complex fixed points which occur as complex conjugates of one another, when the map is elliptic. In the former, hyperbolic case, points are attracted to and repelled from two different points on the projective real line, travelling along circles through these two points. In the latter, elliptic case, points travel on circles around these points.

1.3 The Fundamental Domain of the Modular Group

A **fundamental domain** is a geometric shape which represents the orbits of some group action, 'almost' uniquely. The 'almost uniqueness' is not

so precisely defined. For instance, in ergodic theory, we might want the region to be open except with a set of measure zero added. In the context of modular forms, we shall discuss the fundamental domain associated with a discrete subgroup of $SL_2(\mathbf{R})$. Here by a fundamental domain we shall mean a domain D contained in $\mathbf{H}$ such that $\overline{D}$ contains an element of each orbit class of the action, and there is at most one element from each orbit in D. Every discrete group Γ has a fundamental domain, but we shall prove this only for finite index subgroups of Γ. We use our geometric determination of the hyperbolic plane to determine what the fundamental domains of discrete subgroups of $SL_2(\mathbf{R})$ should look like. If Γ is a discrete subgroup of $SL_2(\mathbf{R})$, we say $z \in \mathbf{H}$ is an **elliptic point** if it is the fixed point of some elliptic $M \in \Gamma$, and an element s of the *projective real line* is called a **cusp** if there is a parabolic element of Γ fixing s.

Theorem 1.5. *If z is an elliptic point of Γ, then the stabilizer of z in Γ is a finite, cyclic group.*

Proof. Since $\mathbf{H}$ is homogenous under the action of $SL_2(\mathbf{R})$, we may assume that z is equal to i. Then the stabilizer of i is $SO_2(\mathbf{R}) \cap \Gamma$. A discrete subgroup of a compact group is finite, and every finite subgroup of $SO_2(\mathbf{R})$ is cyclic, completing the claim. □

We now have enough information to prove that the interior of the set

$$D = \{z \in \mathbf{H} : |z| \geqslant 1, |\mathfrak{Re}(z)| \leqslant 1/2\}$$

is a fundamental region for Γ. We shall find that the matrices

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \qquad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

are fundamental to this proof.

Theorem 1.6. *Let D be the region described above.*

1. *For any $z \in \mathbf{H}$, there is $M \in \Gamma$ for which $Mz \in \overline{D}$.*
2. *If $Mz = w$, and $z, w \in \overline{D}$, then z and w both occur on ∂D and are obtained from each other by reflection in the y axis.*
3. *For each z, the stabilizers Γ_z are trivial except that*

$$\Gamma_i = \{1, S\} \quad \Gamma_{e^{\pi i/3}} = \{1, ST, (ST)^2\} \quad \Gamma_{e^{2\pi i/3}} = \{1, TS, (TS)^2\}$$

Proof. Let $G = \langle S, T \rangle$. If D really was a fundamental region, then we could identify the point in D corresponding to each orbit Γx by taking the point in the strip $|\mathfrak{Re}(z)| < 1/2$ such that $|\mathfrak{Im}(Mz)|$ is maximized. Since

$$\mathfrak{Im}(Mz) = \frac{\mathfrak{Im}(z)}{|cz+d|^2}$$

Maximizing $\mathfrak{Im}(Mz)$ is equivalent to minimizing $|cz+d|$. For each M, there are only finitely many matrices with $|cz+d| < M$. We have

$$|cz+d| \geqslant \mathfrak{Im}(z)|c| \qquad |cz+d| \geqslant |c\mathfrak{Re}(z)+d| \geqslant |d| - |c\mathfrak{Re}(z)|$$

There are only finitely many integers c for which $\mathfrak{Im}(z)|c| < M$, and therefore only finitely many values of d for which $|d| - |c\mathfrak{Re}(z)| < M$. Let M maximize the value of $\mathfrak{Im}(z)$. Without loss of generality, we may assume $|\mathfrak{Re}(Mz)| < 1/2$, for otherwise we may apply T multiple times to transport Mz onto this strip, without changing the imaginary part. If $|Mz| < 1$, then

$$\mathfrak{Im}(SMz) = \frac{\mathfrak{Im}(Mz)}{|Mz|^2} > \mathfrak{Im}(Mz)$$

contradicting the maximality of M, so $Mz \in D$, and we have proved the first property.

Now suppose that for $z, w \in D$, $Mz = w$, where $M = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. We may assume $\mathfrak{Im}(w) \geqslant \mathfrak{Im}(z)$. Since

$$\mathfrak{Im}(w) = \frac{\mathfrak{Im}(z)}{|cz+d|^2}$$

$|cz+d| \leqslant 1$. Because $z \in D$, $\mathfrak{Im}(z) \geqslant \sqrt{1/2}$, so $|c| \leqslant \sqrt{2} < 2$, and we can split the proof into two cases ($c = \pm 1$ or $c = 0$):

- First assume $c = 0$. Then

  $$M = \begin{pmatrix} a & b \\ 0 & a^{-1} \end{pmatrix}$$

  since a and a^{-1} are integers, we must have $a = \pm 1$, and we might as well let $a = 1$, since $M = -M$ in $PSL_2(\mathbf{R})$. Then $w = z + b$. Now $|b| = |\mathfrak{Re}(z-w)| \leqslant 1$, so either $b = 0$ ($w = z$), $b = 1$ ($w = z+1$, so $\mathfrak{Re}(z) = -1/2$), or $b = -1$ ($w = z-1$, so $\mathfrak{Re}(z) = 1/2$).

- Assume $c = 1$ (if $c = -1$, replace M with $-M$). We must have

$$|z + d|^2 = \mathfrak{Im}(z)^2 + (\mathfrak{Re}(z) + d)^2 \leqslant 1$$

One possibility is that $d = 0$, so that $z \in \partial D \cap S^1$. Because

$$M = \begin{pmatrix} a & -1 \\ 1 & 0 \end{pmatrix}$$

$w = a - 1/z$. Now $-1/z \in D$ is also on the unit circle, so either $a = 0$, and $w = -1/z$, which is a just a reflection in the y axis, $a = 1$, $z = e^{i\pi/3}$, $w = z$, and $M = ST$, or $a = -1$, $z = e^{2i\pi/3}$, $w = z$, and $M = S^{-1}T = (TS)^2$.

If $d = 1$, then

$$M = \begin{pmatrix} a & a-1 \\ 1 & 1 \end{pmatrix}$$

So

$$w = \frac{az + (a-1)}{z+1} = a - \frac{1}{z+1}$$

Now $|z + 1| \geqslant 1$ if $z \in D$, so $|\frac{1}{z+1}| \leqslant 1$, and $a - \frac{1}{z+1}$ is in D if and only if $a = 0$, $z = e^{2\pi i/3}$, $w = z$, and $M = TS$. If $a = 1$, $-\frac{1}{z+1} = e^{2\pi i/3}$, so $z = e^{2\pi i/3}$, and $w = e^{\pi i/3}$. If $a = -1$, $-\frac{1}{z+1} = e^{\pi i/3}$, which implies $z = e^{2\pi i/3} - 1 \notin D$.

Finally, let $d = -1$. Then

$$M = \begin{pmatrix} a & -(1+a) \\ 1 & -1 \end{pmatrix}$$

so

$$w = \frac{az - (1+a)}{z-1} = a - \frac{1}{z-1}$$

As before, $|z-1| \geqslant 1$ if $z \in D$, so $|\frac{1}{z-1}| \leqslant 1$. Either $a = 0$, in which case $|z - 1| = 1$, so $z = w = e^{\pi i/3}$, $M = TS^{-1} = (ST)^2$, or $a = 1$, in which case $\frac{-1}{z-1} = e^{2\pi i/3}$, implying $z = e^{2\pi i/3}$, $w = e^{\pi i/3}$, or $a = -1$, in which case $\frac{-1}{z-1} = e^{\pi i/3}$, which would imply $z = e^{2\pi i/3} - 1 \notin D$.

We have addressed all cases, which shows that (ii) and (iii) hold. □

Corollary 1.7. *$PSL_2(\mathbf{Z})$ is generated by S and T.*

Proof. Given $M \in PSL_2(\mathbf{Z})$ pick z in the interior of D, and let $w = Mz$. We have verified in the above proof that there is $N \in \langle S, T \rangle$ for which $Nw \in D$, and since the orbit is unique on the interior of D, $MNz = z$. But then $MN \in \Gamma_z = \{I\}$, so $N = M^{-1}$, and so $M \in \langle S, T \rangle$. □

Remark. If we look at S and T as matrices in $SL_2(\mathbf{Z})$, we find these matrices still generate the group, because $(ST)^3 = -I$. If $M \in SL_2(\mathbf{Z})$ is arbitrary, then either M or $-M$ is in $\langle S, T \rangle$, and by multiplying by $(ST)^3$, we find $M \in \langle S, T \rangle$. Note that since the reduction of $SL_2(\mathbf{Z})$ to $SL_2(\mathbf{Z}_N)$ is surjective, $SL_2(\mathbf{Z}_N)$ is also generated by the induced elements S and T.

We now find the elliptic points of Γ. Let M be such an element. Then we know that M has finite order, hence it's two eigenvalues are roots of unity. But the two eigenvalues lie in a quadratic extension of $\mathbf{Q}$, and therefore have order divisible by 4 or 6.

1.4 The Fundamental Domain

It is an interesting combinatorial problem to count the number of elements in $GL_2(\mathbf{Z}_n)$. Since $SL_2(\mathbf{Z}_n)$ has index $\varphi(n)$ in $GL_n(\mathbf{Z}_n)$, the counting problems over the two groups are equivalent. Since $M_n(Q \times R) \cong M_n(Q) \times M_n(R)$, and if $n = p_1^{n_1} \dots p_m^{n_m}$, then $\mathbf{Z}_n \cong \mathbf{Z}_{p_1^{n_1}} \dots \mathbf{Z}_{p_m^{n_m}}$, it suffices to count $GL_2(\mathbf{Z}_{p^n})$ where p is a prime number. It is easiest to count $GL_2(\mathbf{Z}_p)$, since $\mathbf{Z}_p$ is a field. In this case, $GL_2(\mathbf{Z}_p)$ consists of ordered bases (v, w), with $v, w \in \mathbf{Z}_p^2$. In this case, there are $p^2 - 1$ choices for v (the number of nonzero vectors in $\mathbf{Z}_p^2$), and then $p^2 - p$ choices for w after choosing v (the number of vectors not in the span of v). Thus

$$\#[GL_2(\mathbf{Z}_p)] = (p^2 - 1)(p^2 - p) \qquad \#[SL_2(\mathbf{Z}_p)] = p(p^2 - 1)$$

Consider the reduction from $GL_2(\mathbf{Z}_{p^n})$ to $GL_2(\mathbf{Z}_p)$. The kernel consists of the matrices

$$\begin{pmatrix} 1 + pa & pb \\ pc & 1 + pd \end{pmatrix}$$

where a, b, c, d are arbitrary integers modulo p^n. There are p^{n-1} different choices for each integer, so that the kernel contains $p^{4(n-1)}$ different ma-

trices. Thus

$$\#[GL_2(\mathbf{Z}_{p^n})] = (p^2-1)(p^2-p)p^{4(n-1)} \quad \#[SL_2(\mathbf{Z}_{p^n})] = (p^2-1)p^{3(n-1)+1}$$

Thus the number of matrices in the general linear group $GL_2(\mathbf{Z}_n)$ grows only slightly faster than n^4.

1.5 Congruence Subgroups

An important family of subgroups of Γ are the **principal congruence subgroups**

$$\Gamma(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in PSL_2(\mathbf{Z}) : b \equiv c \equiv 0 \text{ modulo N} \right\}$$

The group is the kernel of the homomorphism obtained by reducing the integer coefficients of elements of Γ modulo N, giving us the exact sequence

$$0 \to \Gamma(N) \to \Gamma \to PSL_2(\mathbf{Z}_n) \to 0$$

A **congruence subgroup** of Γ is one which contains a principal congruence subgroup. Two important examples are

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : c \equiv 0 \text{ modulo } N \right\}$$

$$\Gamma_1(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N) : a \equiv d \equiv 1 \text{ modulo } N \right\}$$

The matrices of $\Gamma_1(N)$, when reduced modulo N, are written in the form

$$\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}$$

and the reduced $\Gamma_0(N)$ are the upper triangle matrices modulo N.

Using the combinatorics we developed over $SL_2(\mathbf{Z})$, we can count the indices of the various congruence subgroups in Γ. For instance, note that if $n \neq 2$, then $I \neq -I$, and so

$$[\Gamma : \Gamma(N)] = |PSL_2(\mathbf{Z}_n)| = \frac{1}{2}|SL_2(\mathbf{Z}_n)|$$

However, $|PSL_2(\mathbf{Z}_2)| = |SL_2(\mathbf{Z}_2)|$, so $[\Gamma : \Gamma(2)] = |SL_2(\mathbf{Z}_2)| = 6$. We have a formula to calculate $|SL_2(\mathbf{Z}_n)|$, so we can calculate $[\Gamma : \Gamma(N)]$ in essentially the same way. The indices here will tell us the number of 'cusps' in the fundamental domains of these congruences groups (For instance $[\Gamma, \Gamma] = 1$, so the standard fundamental domain has only one cusp at ∞).

As we work with fewer and fewer elements of Γ, the action of Γ on $\mathbf{H}$ gets weaker and weaker, and the fundamental domain gets larger and larger. As an example, let's compute the fundamental domain of $\Gamma(2)$. The index of Γ in $\Gamma(2)$ is easily computed to be

$$[\Gamma : \Gamma(2)] = |PSL_2(\mathbf{Z}_2)| = |SL_2(\mathbf{Z}_2)| = 6$$

Thus we may write $\Gamma = \bigcup_{i=1}^{6} M_i \Gamma(2)$ as the disjoint union of cosets, for some $M_i \in \Gamma$. We can choose these M_i to be

$$M_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad M_2 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad M_3 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

$$M_4 = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix} \quad M_5 = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix} \quad M_6 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

The elements must be in different cosets, since these elements differ modulo 2. It then follows that we have a fundamental domain

$$D_2 = \bigcup_{i=1}^{6} M_i^{-1} D$$

Certainly it contains a point in every orbit. It remains to check that repeated points occur only on the boundary, but I think it's pretty tedious to check this so I'll 'leave it for the reader'. The points 0, 1, and ∞ are known as cusps, for they are places where the fundamental domain can 'go off to infinity'. D only has one cusp, the one at ∞.

1.6 The Modular Curve

The fundamental domain of a congruence subgroup G of Γ allows us to understand the topology of $\mathbf{H}/G$ in greater detail. The space $\mathbf{H}/G$ is denoted $Y(G)$, and known as the **modular curve** of the group G. Our first statement should be obvious – since Γ acts on $\mathbf{H}$ discretely, any two points

in $z, w \in \mathbf{H}$ not identified in $Y(G)$ have neighbourhoods $z \in U$ and $w \in W$ such that MU and NW are disjoint, for any $M, N \in \Gamma$. This implies that $Y(G)$ is a Hausdorff topological space.

We would like to put an analytic structure on $Y(G)$, so that we can talk about holomorphic maps on the space. To do this, we need an atlas of mutually holomorphic coordinate maps. This is simple to do so, except at singularity points on $Y(G)$, which algebraically manifest themselves as points in $\mathbf{H}$ whose stabilizers with respect to the action of G are nontrivial. Let us address the trivial case first.

If the stabilizer is trivial at a point z, we may select a compact neighbourhood U of points all of whose stabilizer's are trivial, and then we notice that U intersects only finitely many of its translates $M_1 U, \ldots, M_n U$, because it is bounded a distance away from the real axis – where the space 'concentrates itself'. If we take a set V_i of disjoint neighbuorhoods around each $M_i z$, transport them back to z, and then take the intersection to get a neighbourhood V, then we obtain a neighbourhood whose translates are disjoint. It then follows that V projects injectively into $\mathbf{H}/G$, and this map is open so we have a local homeomorphism. They are mutually holomorphic, since in the inverse image the maps are just the maps $z \mapsto Mz$.

The problem with the points where the stabilizer is non-trivial is that the neighbourhoods are not locally injective here. Indeed, if $Mw = w$, then locally we identify points infinitely close to w because

$$M(w + M'(w)z) \approx w + M'(w)^{k+1} z$$

giving us a series of identifications around w which imply no neighbourhood around w can contained unique representatives of equivalence classes unless $M'(w) = 0$ (which only occurs when M is trivial, because M is a biholomorphic at every point on the Riemann sphere). In the case of the modular group, the map $z \mapsto -1/z$ fixes i, and locally acts like a rotation about 180 degrees, so every neighbourhood of i contains duplicates of representatives of points in $Y(G)$. The neighbourhoods of ω and ω^2 contain triplicates of points in the equivalence classes, and so locally the space around ω is 120°. We call such a point **elliptic**.

There is another way to determine the topological structure around an elliptic point. We note that the stabilizer subgroups of points identified under G are isomorphic, and in fact obtained from each other by conjugation. Provided that G is normal in Γ all the points identified by Γ have the

same stabilizer. Since the stabilizer subgroup with respect to G is always contained in the stabilizer subgroup with respect to Γ, the stabilizer is always finite and cyclic. The cardinality gives us a sense of the 'angle' of the quotient around the point.

This discussion allows us to place coordinate maps at elliptic points of $Y(G)$. Indeed, given a point z, with $Mz = z$, we can consider the Möbius transformation N which maps z to 0 and $\overline{z}$ to ∞. Since $M\overline{z} = \overline{z}$, NMN^{-1} fixes 0 and ∞, so it is a composition of a rotation and a dilation. Since NMN^{-1} has finite order, it must be a rotation by a commensurable angle. In fact, for the elliptic point i, NMN^{-1} is just a rotation by 180°, and for ω and ω^2, NMN^{-1} is just a rotation by 120°. Thus if we consider the power map $f(z) = z^k$, for $k = 2, 3$, $f \circ N$ is a continuous map which identifies points locally around the singularity if and only if they are identified in $Y(G)$. Thus $f \circ N$ decends to a map around the singularity point in $Y(G)$, which forms a coordinate chart. These maps remain mutually holomorphic with the maps we initially defined, because all in all transition maps will either be a power of a Möbius transform or a root of a Möbius transform. Thus $Y(G)$ is a Riemann surface.

The only deficiency of $Y(G)$ is that it is not compact – we left out the action at ∞. We may remedy this by taking the space $\mathbf{H}^*/G$, where $\mathbf{H}^* = \mathbf{H} \cup \{\infty\} \cup \mathbf{Q}$ is the half plane with 'cusps' added. These rational points are necessary for the action of G to be well defined on $\mathbf{H}^*$, because ∞ can be moved to every rational point. The stabilizer at ∞ is the same of all maps of the form $z \mapsto z + m$, where $m \in \mathbf{Z}$. The orbit space obtained is denoted $X(G)$, and is known as the **compactified modular curve**.

The points addended to $Y(G)$ to form the compactification $X(G)$ are known as cusps. For instance, we only have one cusp for $G = \Gamma$, the point at ∞, whereas for the congruence subgroup considered before we have 3 cusps, $\infty, 0, 1$. Each modular curve has only finitely many cusps, precisely because G has a finite index in Γ.

The only problem with this construction is that if we take the quotient topology on $X(G)$ from $\mathbf{H}^*/G$, there are simply not enough open sets in $\mathbf{H}^*/G$ to separate the rational points (which is obvious on the real line, since the rational points form a dense subset). Thus we define a new topological structure at the rational points and at ∞. First, we take the neighbourhoods of ∞ to be half spaces

$$\{\infty\} \cup \{z : \operatorname{Im}(z) > M\}$$

We then ascribe a topological structure to the rational points by making the Möbius transforms homeomorphisms of $\mathbf{H}^*$. The half spaces are mapped conformally onto circles tangent to the rational point. We can then take the quotient topology on $X(G)$ relative to this topology, and now $X(G)$ is compact and Hausdorff.

To complete our discussion, we need only define analytic coordinate maps at the cusps. First, if we are analyzing a rational cusp q, we first consider an $M \in \Gamma$ which maps q to ∞. We then define the width of q to be the index of the group MGM^{-1} in Γ. This is essentially dual to the period of an elliptic point. The period tells us the number of sectors of the disc surrounding an elliptic point that are identified in $Y(G)$. At a cusp, infinitely many sectors converge, but at ∞ these are straightened out to become vertical strips – the width tells us how many vertical strips are distinct in the identification. If q has width h, then we define the coordinate map $z \mapsto e^{2\pi i M(z)/h}$ to be the chart at q.

1.7 Modular Forms

A meromorphic function $f : \mathbf{H} \to \mathbf{C}$ is **weakly modular** of weight k (where k is an even integer) such that for any $M = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$,

$$f(Mz) = (cz+d)^k f(z)$$

We assume k is even, for if k is odd $(cz+d)^k \neq (-(cz+d))^k$, and so there are no non-zero weight k modular functions. Since

$$\begin{aligned} M'(z) &= \left(\frac{az+b}{cz+d}\right)\left(\frac{a}{az+b} - \frac{c}{cz+d}\right) \\ &= \left(\frac{az+b}{cz+d}\right)\frac{1}{(az+b)(cz+d)} \\ &= \frac{1}{(cz+d)^2} \end{aligned}$$

we can write

$$\frac{f(\phi(z))}{f(z)} = (cz+d)^k = \phi'(z)^{-k/2}$$

which we see as a relation between differential forms.

$$\phi_*(f\,dz^{k/2}) = (f \circ \phi)(\phi')^{k/2} dz^{k/2} = f\,dz^{k/2}$$

The structure theorem for $SL_2(\mathbf{Z})$ simplifies the verification of the transformation properties of weakly modular functions.

Theorem 1.8. *f is weakly modular of weight k if and only if*

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

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

Proof. Certainly the relations hold if f is weakly modular. Conversely, suppose f satisfies the relations, and define

$$G = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma : f\left(\frac{az+b}{cz+d}\right) = (cz+d)^k f(z) \right\}$$

Certainly S, T, and T^{-1} are in G. If $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in G$ and $N = \begin{pmatrix} m & n \\ o & p \end{pmatrix}$, then

$$MN = \begin{pmatrix} am+bo & an+bp \\ cm+do & cn+dp \end{pmatrix}$$

and we know

$$f(MN(z)) = (cNz+d)^k f(Nz) = (cNz+d)^k (oz+p)^k f(z)$$

since

$$(cNz+d)(oz+p) = d(oz+p) + c(mz+n) = (cm+do)z + (cn+dp)$$

so $MN \in G$. We conclude that all monomials in the variables T, T^{-1}, S are in G. But all elements of Γ are of this form, so $G = \Gamma$. □

A weakly modular function is periodic, so we may consider the map

$$g(z) = e^{2\pi i z}$$

which maps $\mathbf{H}$ holomorphically onto the punctured unit disk. We may then expand $f \circ g^{-1}$ at the origin in a Laurent series, leading to an expansion of f of the form

$$f(z) = \sum_{k=-\infty}^{\infty} a_k q^k$$

where $q = e^{2\pi iz}$. If f is meromorphic at ∞ (so that the coefficients a_k vanish for k small enough), it is a **modular function** of weight k. If f is actually holomorphic at ∞, so that $a_k = 0$ for $k < 0$, and is holomorphic on $\mathbf{H}$, then we call it a **modular form**. If, even further, we have $a_0 = 0$, then we say that f is a **cusp form**, for then f vanishes at ∞, and therefore at all cusps. We denote the set of modular forms, of weight k, by M_k, and the set of cusp forms by S_k (the German word for a cusp form is a 'spitzenform'). We note that holomorphicity at ∞ manifests as a growth condition $f(x+iy) = O(e^{Cy})$ for some $C > 0$, for then the corresponding function on $\mathbf{D}$ is, in polar coordinates, $O(r^{-C})$. If $f(x+iy) = O(e^{Cy})$ for all $C > 0$, then f is actually bounded near $i\infty$, and so f is a modular form.

To obtain a geometric picture, it is useful to add a point ∞ at infinity on $\mathbf{H}$. If we wish for the map $z \mapsto q$ to be a homeomorphism in $\mathbf{H} \cup \{\infty\}$, we put a topology on $\mathbf{H}$ by saying $z_i \to \infty$ if $\mathrm{Im}(z_i) \to \infty$. Equivalently, the neighbourhoods of ∞ are sets of the form

$$\{z \in \mathbf{H} : \mathrm{Im}(z) > M\}$$

as M ranges from 0 to ∞. This is certainly different to the relative topology induced by the Riemann sphere (if we need to explicitly identify the topology, we often say that $z \to i\infty$ if z converges to ∞ in the topology of the upper half plane. If we wish to analyze the effects of the Möbius group on this set, then we are forced to also consider all rational points, for these points comprise the orbit of ∞ under the modular group. If we now want to make the maps $\gamma \in \Gamma$ homeomorphisms, then we must let the neighbourhoods of each rational point be the images of the neighbourhoods of ∞. These are exactly the circles whose boundary cross through the rational number. These points are known as **cusps**, and are the minimal number of points needed to compactify the orbit space of Γ, and still maintain a nice theory of modular forms. We can then compactify the fundamental region to $\overline{D} = D \cup \{\infty\}$, which is now the fundamental region of Γ over $\overline{\mathbf{H}}$.

The sets of modular functions, forms, and cusp-forms form complex vector spaces. In addition, if f is a modular (function or form) of weight k, and g modular of weight k', then fg is modular of weight $k + k'$, and if $g \neq 0$, f/g modular of weight $k - k'$ (g needs to have no zeroes in $\mathbf{H}$ for f/g to be a modular form). In particular, we may consider the graded algebra M_* of all modular forms (which is also a field), and the set S_* of all cusp forms is an ideal.

1.8 Eisenstein Series

Here's a classical and very important family of modular forms.

Lemma 1.9. *For any $z, w \in \mathbf{C}$, the value*

$$G_s(z,w) = \sum_{(m,n)\neq 0} \frac{1}{|mz+nw|^s}$$

converges for $s > 2$.

Proof. Formally, we have

$$G_s(\lambda z, \lambda w) = \sum_{m,n=-\infty}^{\infty} \frac{1}{|\lambda mz + \lambda nw|^s} = \frac{G_s(L(z,w))}{|\lambda|^s}$$

so the convergence of the series for a particular value of s is invariant under complex multiplication. This also shows that G_s is a weakly modular of weight s, when s is an even integer. Thus we may assume that we are summing over $(1,z)$. If we count the number of elements in the k'th 'layer' of elements around the origin, we end up with $O(k)$ elements, and the value of $1/|m+nz|^s$ here is bounded above by $\min(1,|z|)^s/k^s = O(1/k^s)$. Thus

$$\sum_{m,n=\infty}^{\infty} \frac{1}{|m+nz|^s} \leqslant O(1) \sum_{k=1}^{\infty} \frac{1}{k^{s-1}}$$

which converges for $s-1 > 1$, so $s > 2$. We can actually check that $G_s(z,w)$ converges absolutely if and only if $s > 2$ by analyzing the coefficients in the sequence more carefully. □

This series gives rise to an important modular form. We define

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

which converges absolutely for $k > 2$, and locally uniformly convergent, so that G_k is holomorphic, and

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

$$G_k(-1/z) = \sum_{(n,m)\neq 0} \frac{1}{(n-m/z)^k} = \sum_{(n,m)\neq 0} \frac{z^k}{(n-mz)} = z^k G_k(z)$$

Thus G_k is weakly modular, of weight 12. These functions are known as Eisenstein series. Note that for $|z| > 1$, $|mz+n| \geqslant |m|$, so the series converges uniformly for $|z| > 2$. The symmetry condition defining modular forms then tells us that the series converges locally uniformly on $\mathbf{H}$, and therefore converges to a holomorphic function everywhere. Finally, by applying uniform convergence on $|z| > 2$, we find

$$\lim_{z\to i\infty} \sum_{m,n=-\infty}^{\infty} \frac{1}{(mz+n)^k} = \sum_{m,n=-\infty}^{\infty} \lim_{z\to i\infty} \frac{1}{(mz+n)^k} = \sum_{n=-\infty}^{\infty} \frac{1}{n^k} = 2\zeta(k)$$

So the function has a limit at $i\infty$, and is therefore holomorphic there. Thus we find

Theorem 1.10. *G_k is a modular form of weight k.*

It turns out that the coefficients of the G_k represent a certain arithmetic function of interest to number theorists. First, a proposition about the Riemann zeta function.

Lemma 1.11. *If k is a positive integer, then*

$$\zeta(2k) = (-1)^{n+1} \frac{(2\pi)^{2k}}{2(2k)!} B_{2k}$$

where B_k is the k'th Bernoulli number, defined by the expansion

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

Proof. Consider the locally uniformly convergent Fourier expansion

$$\cos(ay) = \frac{2a \sin \pi a}{\pi} \left(\frac{1}{2a^2} + \sum_{n=1}^{\infty} \frac{(-1)^n}{a^2 - n^2} \cos ny \right)$$

Taking $y = 1$, we find

$$\pi \cot \pi a = \frac{1}{a} + \sum_{n=1}^{\infty} \frac{2a}{a^2 - n^2} = \frac{1}{a} + \sum_{n=1}^{\infty} \frac{1}{a-n} + \frac{1}{a+n}$$

But

$$\pi \cot \pi a = \pi i \frac{e^{i\pi a} + e^{-i\pi a}}{e^{i\pi a} - e^{-i\pi a}} = \pi i + \frac{2\pi i}{e^{2\pi i a} - 1}$$

so if we write $2\pi i a = x$, we have two different expansions. The first is

$$\pi \cot(-ix/2) = \pi i + \frac{2\pi i}{x} \frac{x}{e^x - 1} = \frac{2\pi i B_0}{x} + (\pi i + 2\pi i B_1) + 2\pi i \sum_{k=1}^{\infty} \frac{B_{k+1}}{(k+1)!} x^k$$

and the second is

$$\begin{aligned}
\pi \cot(-ix/2) &= \frac{2\pi i}{x} + \sum_{n=0}^{\infty} \frac{1}{n} \left(\frac{1}{1 - (-a/n)} - \frac{1}{1 - (n/a)} \right) \\
&= \frac{2\pi i}{x} + \sum_{n=1}^{\infty} \frac{1}{n} \sum_{m=0}^{\infty} (-a/n)^m - (a/n)^m \\
&= \frac{2\pi i}{x} - \sum_{m=0}^{\infty} \left(\frac{2}{(2\pi i)^m} \sum_{n=1}^{\infty} \frac{1}{n^{2m+2}} \right) x^{2m+1} \\
&= \frac{2\pi i}{x} + \sum_{m=0}^{\infty} \frac{-2\zeta(2m+2)}{(2\pi i)^m} x^{2m+1}
\end{aligned}$$

This shows us that $\zeta(2m) = -\frac{(2\pi i)^{2m}}{2(2m)!} B_{2m}$, and also that B_k vanishes for odd numbers greater than or equal to 3. □

We use similar techniques to compute the q expansion of the Eisenstein series. Define the k'th divisor sum to be $\sigma_k(n) = \sum_{d \mid n} d^k$.

Theorem 1.12. *The Eisenstein series has the q-expansion*

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

Proof. Replace a with mz in the last proof to obtain the expansion

$$\sum_{n=-\infty}^{\infty} \frac{1}{mz + n} = \pi \cot(m\pi z) = \pi i + \frac{2\pi i}{q^m - 1} = \pi i - 2\pi i \sum_{k=1}^{\infty} e^{2\pi i k m z}$$

Differentiating in the variable mz $2n-1$ times, we find

$$\begin{aligned}\sum_{k=-\infty}^{\infty} \frac{1}{(mz+k)^{2n}} &= (-1)^n \frac{(2\pi)^{2n}}{(2n-1)!} \sum_{k=1}^{\infty} k^{2n-1} q^{km} \\ &= -\zeta(2n) \frac{2n}{B_{2n}} \sum_{k=1}^{\infty} k^{2n-1} q^{km}\end{aligned}$$

Thus

$$\begin{aligned}G_{2n}(z) &= 2\zeta(2n) - 2\zeta(2n) \frac{2n}{B_{2n}} \sum_{m=1}^{\infty} \sum_{k=1}^{\infty} k^{2n-1} q^{km} \\ &= 2\zeta(2n) \left[1 - \frac{2n}{B_{2n}} \sum_{m=1}^{\infty} \left(\sum_{k \mid m} k^{2n-1} \right) q^m \right]\end{aligned}$$

and this is the formula we wanted. □

We get rational coefficients to the Eisenstein series if we normalize by the zeta function, defining

$$E_{2n}(z) = \frac{G_{2n}(z)}{2\zeta(2n)} = 1 - \frac{2n}{B_{2n}} \sum_{m=1}^{\infty} \sigma_{2n-1}(m) q^m$$

This series actually comes about naturally in the theory. For instance, we could define

$$E_{2n}(z) = \frac{1}{2} \sum_{\gcd(m,n)=1} \frac{1}{(mz+n)^{2n}}$$

since it would then follow

$$G_{2n}(z) = \sum_{a=1}^{\infty} \sum_{\gcd(m,n)=a} \frac{1}{(mz+n)^{2n}} = 2 \sum_{a=1}^{\infty} \frac{1}{a^{2n}} E_{2n}(z) = 2\zeta(2n) E_{2n}(z)$$

In particular, we have a series of group actions on the set of meromorphic functions defined by

$$M_n f(z) = (cz+d)^{-n} f\left(\frac{az+bz}{cz+d} \right)$$

The weakly modular functions of weight k are exactly the functions stabilized by the action. In general, if we have a group action from a finite group G on a vector space V, and we want to find vectors stabilized by the action, we can start by fixing some $v_0 \in V$, and then consider $\sum_{x \in G} xv_0$. If v_0 is invariant under $G_0 \subset G$, then we need only consider $\sum_{x \in G/G_0} xv_0$. If G is infinite, this method still works provided that v_0 is invariant under a group G_0 of finite index in G, or if G_0 is a maximal stabilizer (so elements aren't repeated) and the sum converges. If we take f to be a rational function, and $G = \Gamma$, and G_0 the set of modular transformations which fix f, then we obtain a series of modular functions of weight n known as Poincare series.

$$\sum_{M \in G/G_0} (cz+d)^{-n} f\left(\frac{az+b}{cz+d}\right)$$

In particular, if we take v_0 to be the constant function, and set G_0 to be the set of $M \in \Gamma$ which preserve infinity, then the Poincare series is exactly the function E_{2n}.

We have not yet found a Modular form of weight 2. We can define the Eisenstein series

$$G_2(z) = 2\zeta(2) + \sum_{m \neq 0} \sum_{n=-\infty}^{\infty} \frac{1}{(mz+n)^2}$$

$$\begin{aligned} E_2(z) = \frac{G_2(z)}{2\zeta(2)} &= 1 + \frac{3}{\pi^2} \sum_{m \neq 0} \sum_{n=-\infty}^{\infty} \frac{1}{(mz+n)^2} \\ &= 1 + \frac{6}{\pi^2} \sum_{m=1}^{\infty} \sum_{n=-\infty}^{\infty} \frac{1}{(mz+n)^2} \end{aligned}$$

But these sums do not converge absolutely, though the inner sums do for any z and m. As when determining the q coefficients of the higher order G_k, we have

$$\sum_{n=-\infty}^{\infty} \frac{1}{(mz+n)^2} = -\frac{4}{B_2}\zeta(2) \sum_{n=1}^{\infty} nq^{nm} = -4\pi^2 \sum_{n=1}^{\infty} nq^{nm}$$

Hence

$$E_2(z) = 1 - 24 \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} nq^{nm}$$

Since the coefficients of q^n are determined here for small enough values of m, we may collect the coefficients to conclude

$$E_2(z) = 1 - 24\sum_{n=1}^{\infty} \sigma_1(n) q^n$$

Since

$$\sigma_1(n) = \sum_{m \mid n} m \leqslant \sum_{m \mid n} n \leqslant \sum_{m=1}^{n} n = O(n^2)$$

The series converges for all $|q| \leqslant 1$, so we see E_2 is a holomorphic function well defined at ∞. However, the transformation law fails. It is easy to be decieved here. We calculate

$$E_2(-1/z) = \frac{z^2}{2\zeta(2)} \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \frac{1}{(m-nz)^2} = \frac{z^2}{2\zeta(2)} \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \frac{1}{(mz+n)^2}$$

where we ignore the case where $m = n = 0$ when summing. We could conclude that E_2 was a modular form, provided that the summation operations can be swapped – this is where we need to apply absolutely convergence – something we don't have, and something which causes the transformation property to fail. Let us define

$$a_{mn} = \frac{1}{(mz+n-1)(mz+n)} = \frac{1}{mz+n-1} - \frac{1}{mz+n}$$

Then

$$\begin{aligned} \frac{1}{(mz+n)^2} - a_{mn} &= \frac{1}{(mz+n)} \left(\frac{1}{mz+n} - \frac{1}{mz+n-1} \right) \\ &= \frac{-1}{(mz+n)^2(mz+n-1)} \end{aligned}$$

So the modified series

$$\tilde{E}_2(z) = \frac{1}{2\zeta(2)} \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \frac{1}{(mz+n)^2} - a_{mn}$$

converges absolutely. The a_{mn} telescope to zero when summed over n, so $\tilde{E}_2(z) = E_2(z)$. This means that

$$E_2(-1/z) - z^2 E_2(z) = \frac{z^2}{2\zeta(2)} \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} a_{mn}$$

and now

$$\begin{aligned}\lim_{N\to\infty}\sum_{n=-N}^{N}\sum_{m=-\infty}^{\infty}a_{mn} &= \lim_{N\to\infty}\sum_{m=-\infty}^{\infty}\sum_{n=-N}^{N}a_{mn}\\ &= \lim_{N\to\infty}\sum_{m=-\infty}^{\infty}\left(\frac{1}{mz-N}-\frac{1}{mz+N}\right)\\ &= \lim_{N\to\infty}\frac{2}{z}\left(\pi\cot(-\pi N/z)+z/N\right)\\ &= \frac{2\pi i}{z}\lim_{N\to\infty}\frac{e^{-2\pi iN/z}+1}{e^{-2\pi iN/z}-1} = -\frac{2\pi i}{z}\end{aligned}$$

and we conclude that

$$E_2(-1/z) = z^2E_2(z) - \frac{6iz}{\pi} = z^2E_2(z) + \frac{12z}{2i\pi}$$

This procedure can be seen as introducing a 'correction coefficient' a_{nm} into the inner sums which make both sums absolutely converge, then the two sums would be equal, and we would find the sum above equal to

$$\sum_m\sum_n a_{mn} - \sum_n\sum_m a_{mn}$$

and can be applied to other problems of this type.

1.9 Classification of Modular Forms

We shall now show that all functions can be represented in terms of the Eisenstein series. Recall that the order $\mathrm{ord}(f,w)$ of a non-zero meromorphic function f at a point w is the unique integer k such that $(z-w)^k f(z)$ is a function non-zero and holomorphic at w. Note that if f is a weakly modular function, then $\mathrm{ord}(f,w) = \mathrm{ord}(f,Mw)$, because the behaviour of f around w is, up to a power of $(cz+d)^k$, the Certainly. same this is true

for $\gamma = T$, since f is periodic, and

$$\begin{aligned} w^m f(-1/z+w) &= w^m f\left(-\frac{1-wz}{z}\right) \\ &= w^m \left(\frac{z}{1-wz}\right)^k f\left(\frac{z}{1-wz}\right) \\ &= \left(\frac{w(1-wz)}{z}\right)^m \left(\frac{z}{1-wz}\right)^k \left(\frac{z}{1-wz}\right)^m f\left(\frac{z}{1-wz}\right) \end{aligned}$$

This function is holomorphic at $w = 0$, provided m is the order of $f(z)$, and it is nonzero at $-1/z$, so this shows that the order is preserved by $\gamma = S$. The general case is then true because Γ is generated by S and T. Thus we may talk about the order of a point on $\mathbf{H}/\Gamma$. We also define, for a modular function, the order at ∞ to be the smallest integer whose corresponding q-coefficent is non-zero.

The key assertion is that the total number of zeroes depends only on the weight of the modular form. This depends on the geometry of $\mathbf{H}/\Gamma$, especially the singularities of the region (the *elliptic fixed points*, whose stabilizer over Γ is not trivial) and the non-compactness of the region – the cusp at ∞. Recall that $\mathbf{H}/\Gamma$ is obtained from the fundamental region by attaching the boundary be reflection in the y-axis. The only problem is that this quotient has singularities at $\omega = e^{i\pi/6}$, i, ω^2, and ∞, which have neighbourhoods that aren't discs (ω and ω^2, and ∞ have arcs of length $2\pi/3$ and i has an arc of length π). These lengths give us the ratios which we can use to count the zeroes of the function.

Theorem 1.13. *If $f \neq 0$ is a modular function of weight k, then*

$$ord(f,\infty) + \frac{ord(f,i)}{2} + \frac{ord(f,\omega)}{3} + \sum_{\substack{p \in \mathbf{H}/\Gamma \\ p \neq i, e^{i\pi/6}}} ord(f,p) = \frac{k}{12}$$

or perhaps more lucidly,

$$ord(f,\infty) + \sum_{p \in \mathbf{H}/\Gamma} \frac{ord(f,p)}{\Gamma_p} = \frac{k}{12}$$

Proof. The idea of the proof results from counting zeroes and poles via integrating the logarithmic derivative. Consider the oriented curve α shown

in the image below. which is tall enough that the region contains one of each zero and pole in each orbit (orbits cannot go to infinity because f is meromorphic there). We swerve around points $p \in \partial D$ which are zeroes and poles, such that only one of each point in the orbit class in contained in the region in question. The residue theorem performs its magic, telling us that

$$\frac{1}{2\pi i}\int_\alpha \frac{f'(z)}{f(z)}dz = \sum_{p\in \mathbf{H}/\Gamma, p\neq i,\omega} \operatorname{ord}(f,p)$$

Now let's calculate the integral explicitly. Let $q = e^{2\pi i z}$ and $\tilde{f}(q) = f(z)$. Then

$$2\pi i q \tilde{f}'(q) = f'(z)$$

Thus

$$\frac{1}{2\pi i}\int_{HA} \frac{f'(z)}{f(z)}dz = \frac{1}{2\pi i}\int_\omega \frac{\tilde{f}'(q)}{\tilde{f}(q)}dq$$

Where ω is now a clockwise rotation around the origin. The only possible poles and zeroes are at $q = 0$, so this evaluates to $-\operatorname{ord}(f,\infty)$. T takes AB to HG, so

$$\frac{1}{2\pi i}\int_{AB} \frac{f'(z)}{f(z)}dz = -\frac{1}{2\pi i}\int_{GH} \frac{f'(z)}{f(z)}dz$$

so the integrals cancel out when they are added together. Similarily, S takes CD to FE, and because $f(Sz) = z^k f(z)$, we find

$$\begin{aligned}\frac{1}{2\pi i}\left(\int_{CD}\frac{f'(z)}{f(z)} - \int_{FE}\frac{f'(z)}{f(z)}dz\right) &= \frac{1}{2\pi i}\int_{CD}\left(\frac{f'(z)}{f(z)} - \frac{f'(Sz)}{f(Sz)}\right)dz \\ &= -\frac{k}{2\pi i}\int_{CD}\frac{1}{z}dz\end{aligned}$$

We will be taking the limit as we shrink the radius of the circles defining BC, DE, and FG to zero, in which case we trace out an angle of size $\pi/6$ in the limit, and so the value of the integral is $k/12$. Finally, we integrate BC, FG, and DE. In the limit, as we decrease the size of the circle, the circles have angles $\pi/6$, $\pi/2$, and $\pi/6$, so the path integrals integrate to $-\operatorname{ord}(f,\omega)/6$, $-\operatorname{ord}(f,i)/2$, and $-\operatorname{ord}(f,-\overline{\omega})/6$. Summing up these integrals, and then combining them with our original deduction in terms of the residue theorem, we obtain the required formula. □

The factor of $1/12$ can be explained as $1/4\pi$ times the volume of $\mathbf{H}/\Gamma$, taken with respect to the hyperbolic metric. The factor will change when we weaken the group we analyze – for instance, when we look at modular forms over congruence subgroups of Γ.

The pole formula for modular function places a vast restriction on the degrees of freedom of modular forms, because the order of the points in its domain must be positive, so there can only be a very limited number of zeroes. If $f \neq 0$ is a modular form of weight $k < 0$, then

$$0 \leqslant \operatorname{ord}(f,\infty) + \frac{\operatorname{ord}(f,i)}{2} + \frac{\operatorname{ord}(f,e^{\frac{i\pi}{6}})}{3} + \sum_{\substack{p \in \mathbf{H}/\Gamma \\ p \neq i, e^{i\pi/6}}} \operatorname{ord}(f,p) = \frac{k}{12} < 0$$

which is clearly impossible, $M_k = (0)$ for $k < 0$. If f has weight zero, then it certainly must have no zeroes on $\mathbf{H}$ (for the sum of zeroes must equal zero), but then if $f(w) = w'$, $f - w'$ is a form of weight zero, with a zero in $\mathbf{H}$, so $f - w = 0$, and so $f = w$ is a constant function. Suppose that f is a non-zero modular form of weight k, where $k = 4$, $k = 6$, $k = 8$, and $k = 10$, and $k = 14$. Then the zero order formula tells us that f has no zeroes at $i\infty$. But then $2\zeta(k)f - f(i\infty)E_k$ has a zero at $i\infty$, so $2\zeta(k)f = f(i\infty)E_k$, and so $M_k = \mathbf{C}E_k$. In general, we have a bound

Theorem 1.14. *The dimension on the space of modular forms is bounded by*

$$dim(M_k) \leqslant \begin{cases} [k/12] + 1 & k \not\equiv 2 \ (mod\ 12) \\ [k/12] & k \equiv 2 \ (mod\ 12) \end{cases}$$

where $[x]$ is the smallest integer greater than or equal to x.

Proof. If $f_1, \ldots, f_{n+1} \in M_k$, then we may choose $g = \sum a_i f_i$ to have zeroes at n non-elliptic points. Provided $n > k/12$, the pole formula tells us that $g = 0$. Thus the dimension of M_k must be less than or equal to n. If $k \equiv 2 \pmod{12}$, we can improve the bound by noticing that we must have at least a single zero at i and a double zero at ω, which already gives us $14/12$ 'zeroes' to start with, and we need only find a function with greater than or equal to $k - 2/12 = (k-2)/12$ zeroes to conclude it is constant. □

One of the beautiful facts about Modular forms is that low weight forms are very low dimensional, so that if two modular forms are the same

weight, chances are they are related in a very significant way. This gives rise to many of the beautiful equalities which emerge from the theory, such as Euler's pentagonal equality, and the formula for the discriminant we now show. For instance, the space of modular forms of weight 8 is one-dimensional, so E_4^2 and E_8 are scalar multiples of each other. Since E_4^2 and E_8 have the same value at ∞, they actually must be equal. Expanding out the q series of E_4^2 and E_8, we find that

$$\begin{aligned} 1 + 240 \sum_{m=1}^{\infty} \sigma_7(m) q^m &= \left(1 + 120 \sum_{m=1}^{\infty} \sigma_3(m) q^m \right)^2 \\ &= 1 + \sum_{m=1}^{\infty} \left(240\sigma_3(m) + 120^2 \sum_{n=1}^{m-1} \sigma_3(n)\sigma_3(m-n) \right) q^m \end{aligned}$$

Which gives us the arithmetic identity

$$\sigma_7(m) = \sigma_3(m) + 120 \sum_{n=1}^{m-1} \sigma_3(n)\sigma_3(m-n)$$

whose proof would be nigh impossible with the analytic facts we've uncovered.

1.10 The Modular Discriminant

We define the discriminant function as

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

This is a modular form of weight 12. To see this, note that

$$\begin{aligned} \frac{1}{2\pi i} \frac{d \log(\Delta)}{dz} &= 1 - 24 \sum_{n=1}^{\infty} \frac{n q^n}{1 - q^n} = 1 - 24 \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} n q^{nm} \\ &= 1 - 24 \sum_{n=1}^{\infty} \sigma_1(n) q^n = E_2(z) \end{aligned}$$

So using the transformation rule for E_2, we find

$$\frac{1}{2\pi i}\frac{\Delta((az+b)(cz+d)^{-1})}{(cz+d)^{12}\Delta(z)} = \frac{1}{(cz+d)^2}E_2\left(\frac{az+b}{cz+d}\right) - \frac{12}{2\pi i}\frac{c}{cz+d} - E_2(z) = 0$$

and this effectively completes the proof.

Δ has a zero at ∞, which must be of order one by the zero summation formula, and can therefore have no zeroes anywhere else. Thus the map $f \mapsto f\Delta$ is an isomorphism from M_k to S_{k+12}, and since $M_k = S_k \oplus \mathbf{C}E_k$, we find that all modular forms can be written as the sum and product of Eisenstein series. In particular, $\mathbf{C}[E_4, E_6] = M_*$, which can shown by induction since $\Delta \in \mathbf{C}[E_4, E_6]$, and for any even $n = 4m + 6l$, we find $E_4^m E_6^l$ is a modular form of order n, non-zero at ∞, and $M_n = S_n \oplus \mathbf{C}E_4^m E_6^l$. Thus all elements in M_n can be written

$$\sum_{4m+6l=n} a_{m,l} E_4^m E_6^l$$

and there is a one-to-one correspondence here.

The set of cusp forms of weight 12 form a space of dimension one, so any two cusp forms of weight twelve differ by a constant (see what I mean now?). Δ is a cusp form, as is the function

$$F(z) = \frac{1}{1728}[E_4^3(z) - E_6^2(z)]$$

The first q coefficent of Δ is 1. To calculate the first q coefficient of F, we write

$$\begin{aligned} F(z) &= \frac{1}{1728}\left[\left(1 - \frac{8}{B_4}q + \ldots\right)^3 - \left(1 - \frac{12}{B_6}q + \ldots\right)^2\right] \\ &= \frac{24}{1728}[1/B_6 - 1/B_4]q + \ldots \\ &= q + \ldots \end{aligned}$$

Since F is a scale multiple of Δ, and they both have the same first q coefficient, they must be equal!

The coefficients of the expansion of Δ are known as the Ramanujan tau function,

$$\Delta(z) = \sum_{n=1}^{\infty} \tau(n) q^n$$

τ has many interesting properties, conjectured by Ramanujan well before the theory of modular forms was properly developed. For instance, τ is a multiplicative function, and if p is prime, then $|\tau(p)| \leqslant 2p^5\sqrt{p}$. This is an incredibly deep inequality, and was only proved recently in the 1970s. It is easier to prove that $|\tau(p)| = O(p^6)$, proved by Hecke in the 30s in a general bound for cusp forms.

The coefficients of the Eisenstein's series $G_{2n}(z) = \sum a_k q^k$ grow on the order of n^{2n-1}, since

$$n^m \leqslant \sum_{k \mid n} k^m \leqslant \zeta(m) n^m$$

The upper bound is easily seen, if $n = p_1^{n_1} \dots p_m^{n_m}$, from the prime representation

$$\zeta(m) n^m \geqslant n^m \prod_{i=1}^{n} \left(1 + \frac{1}{p_i^m} + \frac{1}{p_i^{2m}} + \cdots + \frac{1}{p_i^{mn_i}}\right) = \sum_{k \mid n} k^m$$

We obtain much better decay for cusp forms.

Theorem 1.15. *If $f(z)$ is a cusp form of weight k, with Fourier expansion $f(z) = \sum a_k q^k$, then $|a_n| = O(n^{k/2})$, via constant dependant on f.*

Proof. The map $z \mapsto y^{k/2}|f(z)|$ is invariant under the modular group, since certainly $y^{k/2}|f(z+1)| = y^{k/2}|f(z)|$, and

$$\frac{y^{k/2}|f(-1/z)|}{(x^2+y^2)^{k/2}} = \frac{y^{k/2}|z|^k|f(z)|}{|z|^k} = y^{k/2}|f(z)|$$

So the map is invariant under Γ. As $z \to i\infty$, $y^{k/2}|f(z)| \to 0$ because $|f(z)|$ decays rapidly at infinity. Thus the map is bounded on the fundamental domain of $\mathbf{H}$, and thus everywhere, so that a global maximum exists, and we may assume $|f(z)| \leqslant Cy^{-k/2}$. The Fourier coefficients can be defined by the equations

$$a_n = e^{2\pi n y} \int_0^1 f(x+iy) e^{-2\pi i n x} dx$$

Since $|f(z)| \leqslant C$, $|a_n| \leqslant Ce^{2\pi n y} y^{-k/2}$, valid for any y. If we set $y = 1/n$, we find $|a_n| \leqslant Ce^{2\pi} y^{-k/2}$. □

In general, we find that the coefficients of modular form of weight k is $O(n^k)$, because any modular form can be written as the sum of a cusp form and an Eisenstein series.

Returning to the τ function, we note that it has interesting congruence properties. Indeed we have

$$\begin{aligned}\Delta &= \frac{E_4^3 - E_6^2}{1728} = \frac{(1+240A_3)^3 - (1-504A_5)^2}{1728} \\ &= \frac{5}{12}(A_3 - A_5) + A_5 + 100A_3^2 - 147A_5^2 + 8000A_5^3\end{aligned}$$

where $A_n = \sum_{m=1}^{\infty} \sigma_n(m)q^m$. But $\sigma_5(n) - \sigma_3(n)$ is divisible by 12 for every n, because $d^5 - d^3 = d^3(d-1)(d+1)$ is divisible by 12 for each d (the product contains at least 2 even numbers, and one number divisible by 3). Thus $(5/12)(A_3 - A_5)$ has integral coefficients. This shows that Δ itself has integral coefficients (an alternate proof, not assuming the product description we already constructed). We actually have $\sigma_5(n) \equiv \sigma_3(n)$ modulo 24, because 24 divides $d^3(d^2-1)$, so $A_3 - A_5$ has even coefficients, and therefore the coefficients of Δ modulo 2 are equal to the coefficients of $A_5 + A_5^2$, and since the coefficients of A_5^2 modulo 2 are just $\sum \sigma_5(n)q^{2n}$, we find

$$\tau(2n) \equiv \sigma_5(2n) + \sigma_5(n)$$

and for odd n,

$$\tau(n) \equiv \sigma_5(n)$$

Modulo 2, $\sigma_5(n)$ is just the number of odd divisors of n, and if we write the odd prime factors of n as $p_1^{k_1} \dots p_m^{k_m}$, we find that $\sigma_5(n) \equiv (k_1+1)\dots(k_m+1)$, so if n is odd, $\tau(n)$ is odd if and only if all the k_i are even – i.e. if n is the square of an odd number. If we consider the number $n = 2^m n_0$, where n_0 is an odd square, then $\tau(n) \equiv \sigma_5(2^{m-1}n_0) + \sigma_5(n) \equiv 1 + 1 \equiv 0$, and if n_0 is not an odd square, $\tau(n) \equiv 0 + 0 \equiv 0$. Thus $\tau(n)$ is odd if and only if n is the square of an odd number.

In a completely different calculation, we can calculate over modular forms of order 12 that

$$-\frac{B_{12}}{24} + \sum_{m=1}^{\infty} \sigma_{11}(n)q^n = \Delta(z) + \frac{691}{156}\left(\frac{E_4^3}{720} + \frac{E_6^2}{1008}\right)$$

So we have a famous congruence of Ramanujan, that $\tau(n) \equiv \sigma_{11}(n) \pmod{691}$.

1.11 The j-invariant

Let's get back to modular functions now. Define the function

$$j(z) = \frac{E_4^3}{\Delta} = \frac{1}{q} + 744 + 196884q + 21493760q^2 + \dots$$

a modular function of weight zero, known as the modular invariant. Since Δ has no zeroes on $\mathbf{H}$, j is holomorphic on $\mathbf{H}$, and has a simple pole at ∞, since Δ has a simple zero there.

Theorem 1.16. *j is a bijection from $\overline{H}/\Gamma$ to the Riemann sphere $\mathbf{PC} = \mathbf{C} \cup \{\infty\}$.*

Proof. j has a simple pole at infinity since E_4 does not vanish at infinity and Δ is a cusp form. For any $w \in \mathbf{C}$ the modular form $1728E_4^3 - w\Delta$ must vanish at exactly one point, for it is a modular form of order 1. But this implies $j(z) - w = 0$ for exactly one value in $\overline{H}/\Gamma$. □

Theorem 1.17. *The Modular functions of weight zero for Γ are precisely the rational functions of j.*

Proof. Certainly every element of $\mathbf{C}(j)$ is a modular function of weight zero, since j is a meromorphic modular function of weight zero. If f has poles $z_1, \dots, z_l$ of order $n_1, \dots, n_l$, then

$$g(z) = \prod_{n=1}^{\infty} [j(z) - j(z_1)]^{n_1} f(z)$$

is a modular function of weight zero with no poles in $\mathbf{H}$, so we may assume that f has no finite poles. If f has a pole at $i\infty$, define $g = \Delta^k f$, for a suitable value of k such that the pole at ∞ is removed, and so g is a modular form of weight $12k$. We may therefore write

$$f(z) = \sum_{4n+6m=12k} a_{n,m} \frac{E_4^n E_6^m}{\Delta^k}$$

We find $n = 3n'$, $m = 2m'$ by division laws, and E_4^3/Δ and E_6^2/Δ are each in $\mathbf{C}(j)$ (they are actually linear in j), and because each monomial in the sum is the product of such factors, we conclude that $\mathbf{C}(j)$ is the space of all weight zero modular functions. □

The compactification of $\mathbf{H}/\Gamma$ can be given the structure of a complex analytic manifold, and these propositions imply a Biholomorphism with $\mathbf{CP}^1$ by j, and hence the only holomorphic functions on the compactification are rational functions of j, as the only holomorphic functions on $\mathbf{CP}^1$ are rational functions.

Chapter 2

Complex Torii and Elliptic Curves

An **elliptic curve** is a non-singular genus one curve, which can be described as an equation

$$Y^2Z = X^3 + aXZ^2 + bZ^3$$

where $4a^3 + 27b^2 \neq 0$

2.1 Lattices

There is an interesting connection between the theory of modular forms and the theory of lattices which emerges because of the translation invariant properties of the spaces. As with modular forms, there are many different ways of looking at lattices, and they all have useful consequences. Recall that a **lattice** in a finite dimensional real vector-space V is an additive subgroup L of V such that one of the equivalent conditions holds.

1. L is discrete, and V/L is compact.
2. There is an $\mathbf{R}$ basis $e_1, \ldots, e_n$ of V such that $L = \mathbf{Z}e_1 \oplus \cdots \oplus \mathbf{Z}e_n$.
3. L is discrete, and spans V.

In our situation, we take $V = \mathbf{C}$, and we let $\mathcal{R}$ denote the set of all lattices on $\mathbf{C}$. Such lattices can be described by pairs of $\mathbf{R}$-independant complex numbers $(z, w) \in \mathbf{C}^2$. Given any particular pair, we let $L(z, w)$ denote the lattice generated by the two numbers. Note that z and w are dependant if

and only if $z/w \in \mathbf{R}$, so we may assume that $z/w \in \mathbf{H}$ (if z/w is in the other side of the plane, consider w/z instead). We let

$$M = \{(z, w) \in \mathbf{C}^2 : z/w \in \mathbf{H}\}$$

And we then view L as a map from M to $\mathcal{R}$.

A lattice can be defined by many different pairs of complex numbers. For instance, if $w' = -z$, and $z' = w$, then $L(z, w) = L(z', w')$, and we find

$$z'/w' = -w/z = S(z/w)$$

Furthermore, if we let $z' = z + w$ and $w' = z$, then $L(z', w') = L(z, w)$ and

$$z'/w' = z/w + 1 = T(z/w)$$

So we begin to see the modular group Γ giving us information about the various different bases of any particular lattice. Indeed, if $M = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$, and $(z', w') = M(z, w)$ (where we define the action of Γ on M in the obvious way), then $L(z', w') = L(z, w)$, because $(z, w) = M^{-1}(z', w')$. What's more,

$$z'/w' = \frac{az + bw}{cz + dw} = \frac{a(z/w) + b}{c(z/w) + d} = M(z/w)$$

The group Γ plays a role in the theory of lattices because the theory has discrete scale invariance, which is essentially the role of Γ as a subgroup of $PGL_2(\mathbf{C})$.

Theorem 2.1. *If $z/w, z'/w' \in \mathbf{H}$, then $L(z, w) = L(z', w')$ if and only if (z, w) is congruent to (z', w') relative to Γ.*

Proof. If $L(z, w) = L(z', w')$, we can write

$$z' = az + bw \qquad w' = cz + dw$$

It then follows that

$$z = \frac{az' - bw'}{ad - bc} \qquad w = \frac{dw' - cz'}{ad - bc}$$

Since all of the coefficients of these relations must be integer valued, we have $ad - bc \mid a, b, c, d$. We may assume the a, b, c, d are relatively prime, for

if we write $a = \lambda a_1$, $b = \lambda b_1$, $c = \lambda c_1$, and $d = \lambda d_1$, with a_1, b_1, c_1, and d_1 relatively prime, we find

$$z = \frac{1}{\lambda}\frac{a_1 z' - b_1 w'}{a_1 d_1 - b_1 c_1} \qquad w = \frac{1}{\lambda}\frac{d_1 w' - c_1 z'}{a_1 d_1 - b_1 c_1}$$

from which we conclude that $\lambda \mid a_1, b_1, c_1, d_1$ again, hence $\lambda = 1$. It then follows that $ad - bc = \pm 1$. But if $ad - bc = -1$, we find z'/w' and z/w lie on opposite sides of the complex plane, contradicting the fact that they both lie on $\mathbf{H}$. Thus we find $(z', w') = M(z, w)$ for $M \in \Gamma$, and then $z'/w' = M(z/w)$. □

We can therefore identify $\mathcal{R}$ with M/Γ. Now $\mathbf{C}^*$ acts on $\mathcal{R}$ by scaling, and this is the same map induced on the quotient M/Γ from M by scaling. Now we may identify $M/\mathbf{C}^*$ with $\mathbf{H}$, by the quotient map $(z, w) \mapsto z/w$, and the action of Γ on M induces the action of Γ on $\mathbf{H}$, so

$$\mathcal{R}/\mathbf{C}^* \cong (M/\Gamma)/\mathbf{C}^* \cong (M/\mathbf{C}^*)/\Gamma \cong \mathbf{H}/\Gamma$$

This is the main reason why Γ is called the modular group, because it characterizes the *moduli space* of lattices.

This observation can be used to construct modular functions from lattices. Let F be a complex valued function on $\mathcal{R}$. These are just functions $F(z, w)$ on M which are invariant under the action of Γ. If

$$\lambda^k F(\lambda z, \lambda w) = F(z, w)$$

Then we obtain a function $f : \mathbf{H} \to \mathbf{C}$ defined by $f(z) = F(z, 1)$, and this function satisfies

$$f\left(\frac{az+b}{cz+d}\right) = F\left(\frac{az+b}{cz+d}, 1\right) = (cz+d)^k F(az+b, cz+d) = (cz+d)^k f(z)$$

Conversely, given any f, we may define $F(z, w) = w^{-k} f(z/w)$. We shall call F a modular function if it induces a meromorphic f which is meromorphic at infinity.

2.2 Class Numbers

Let us apply what we have learned to the study of quadratic forms $Q(x, y) = Ax^2 + Bxy + Cy^2$, for $A, B, C \in \mathbf{Z}$, a classic problem in number theory.

Provided that the discriminant $D = B^2 - 4AC < 0$, $Q(x,y)$ has no non-zero roots in $\mathbf{R}^2$, and so either $Q(x,y) > 0$ for all $(x,y) \neq 0$, or $Q(x,y) < 0$. Without loss of generality, we assume Q is a positive quadratic form, so that $A, C > 0$. We also assume that Q is primitive, in the sense that $\gcd(A,B,C) = 1$. Let $\mathfrak{Q}_D$ denote the set of all primitive quadratic forms with discriminant D. Note that $Q(\lambda x, \lambda y) = \lambda^2 Q(x,y)$, and if $ad - bc = 1$, and Γ acts on $\mathfrak{Q}_D$ by composition. That is,

$$\begin{aligned}\left[\begin{pmatrix} a & b \\ c & d \end{pmatrix} Q\right](x,y) &= Q(ax+by, cx+dy) \\ &= (Aa^2 + Bac + Cc^2)x^2 \\ &\quad + (2Aab + B(ad+bc) + 2Ccd)xy \\ &\quad + (Ab^2 + Bbd + Cd^2)y^2\end{aligned}$$

and this form has discriminant D, verified by calculation. Thus we expect the theory of modular forms to apply somewhere due to the scale invariance. The number of orbits of this action is actually finite, and is called the class number of the discriminant. To prove this, we associate with any $Q \in \mathfrak{Q}_D$ the unique root of z_Q in $\mathbf{H}$ for which $Q(x, z_Q x) = 0$. One verifies that $z_Q = (-B + i\sqrt{-D})/2A$. If the roots of two positive primitive quadratic forms $Q = [A,B,C]$ and $R = [A',B',C']$ are equal, then $Q = R$, because if

$$Q(x,y) = A(x - zy)(x - \overline{z}y) = Ax^2 - 2A\mathfrak{Re}(z)xy + A|z|^2y^2$$

$$R(x,y) = A'(x - zy)(x - \overline{z}y) = A'x^2 - 2A'\mathfrak{Re}(z)xy + A'|z|^2y^2$$

then $B'/B = A'/A = C'/C$, so R is a multiple of Q, and since R is also primitive and positive, R must be a multiple of Q by 1, i.e. $Q = R$. Thus we can uniquely identify an element of $\mathfrak{Q}_D$ from its roots. One can also see that if $M \in \Gamma$, then $z_{MQ} = M^{-1} z_Q$, so by applying the classification of the orbit space of Γ over $\mathbf{H}$, we find each equivalence classes in $\mathfrak{Q}_D$ has a unique representive of $Q = [A,B,C]$ with z_Q in the fundamental domain. A short calculation shows this means exactly that

$$-A \leqslant B \leqslant A \qquad C \geqslant A$$

This set is finite, because

$$|D| = 4AC - B^2 \geqslant 3A^2$$

so A is bounded, which implies B is bounded, and then C is bounded because there is at most one C for each pair of A and B. We remark that the sequence of all class numbers form the coefficients of a certain modular form, but we do not go into detail about this.

2.3 Hecke Theory

Consider the space of formal sums of lattices, and define the operator T_n on this space, mapping a lattice L to the formal sum of sublattices L' of index n with respect to L. If $(L : L') = n$, then clearly L' is contains the lattice nL, for if $x \in L$ is arbitrary, then by Lagrange's theorem we must have $nx = 0 \in L/L'$, so $nx \in L'$. If we consider the image L'/nL of L' in L/nL, then we see from one of the isomorphism theorems that $(L/nL, L'/nL) = n$, and $(L/nL)/(L'/nL) \cong L/L'$. Conversely, if G is a subgroup of L/nL of index n, then we may identify G with its inverse image in L, and we find this group has index n in L. Since $L/nL \cong \mathbf{Z}_n^2$, so a group has index n in L/nL if and only if it is a group of size n, and it is normally much easier to count the groups of size n then the subgroups of index n of L.

Since we can identify the group L with $\mathbf{Z}^2$, and L/nL with $\mathbf{Z}_n^2$, the operator T_n is not really that interesting to the theory of lattices, since it is locally (with respect to sublattices of a particular lattice) just a function related to $\mathbf{Z}^2$. The fun begins when we add additional operators. For instance, we have the scale operators $R_\lambda(L) = \lambda L$. Together, these operators have nice composition properties.

Theorem 2.2. *The Lattice operators satisfy*

(a) $R_\lambda \circ R_\gamma = R_{\lambda\gamma}$.

(b) $R_\lambda \circ T_n = T_n \circ R_\lambda$.

(c) If n and m are relatively prime, then $T_n \circ T_m = T_{nm}$.

(d) If p is prime, $T_{p^n} \circ T_p = T_{p^{n+1}} + pT_{p^{n-1}}R_p$

Proof. (a) is obvious, and (b) follows because R_λ preserves the index of lattices. Now (c) is true if there is a unique lattice of degreee n between a lattice of degree nm and L. The group $L/nmL \cong \mathbf{Z}_{nm}^2$ decomposes into the direct product of $\mathbf{Z}_n^2$ and $\mathbf{Z}_m^2$. Given a group L' of index nm, L'/nmL

is a group of size nm so we need only need to prove that every group G of size nm in $\mathbf{Z}_n^2 \oplus \mathbf{Z}_m^2$ extends to a unique group of size nm^2. This is certainly true if $\mathbf{Z}_n^2 \oplus \mathbf{Z}_m^2/G \cong \mathbf{Z}_n \oplus \mathbf{Z}_m$, but this isomorphism always holds, because $\mathbf{Z}_n^2 \oplus \mathbf{Z}_m^2/G$ is a group of order nm containing elements of order n and m, but no element has an order which properly divides n or m. Now (d) is the really interesting property. Note that the range of all three operators consists of sums of lattices of index p^{n+1} with respect to the original lattice. Let L be the lattice we are computing the coefficients of, and L' a lattice of index p^{n+1} in L, with coefficients (a, b, c) relative to the operators $T_{p^n}T_p$, $T_{p^{n+1}}$, and $T_{p^{n-1}}R_p$. Note that b is always equal to 1, so suffices to prove that $a = 1 + pc$. We split the proof into two cases

- ($L' \not\subset pL$). Then $c = 0$, and we need to prove $a = 1$. We want to prove that there is a unique lattice L'' with $L' \subset L'' \subset L$, with L'' index p in L. L''/pL is a group of order p. But the image of L' in L''/pL is nontrivial, and contained within L''/pL, so $L'/pL = L''/pL$, and this uniquely identifies L''.

- ($L' \subset pL$) Then $c = 1$, and we need to prove $a = p + 1$. That is, L' is contained within $p + 1$ lattices of index p with respect to L. But if a lattice has index p it contains pL, so we need only count the number of lattices of index p, which is the number of size p subgroups of $L/pL \cong \mathbf{Z}_p^2$. This then follows from vector space theory.

□

There are some nice corollaries to this theorem. Indeed, as with most multiplicative functions, T_{p^n} can be written as a polynomial in R_p and T_p, and thus T_n as a polynomial in the primes which form T_n. Furthermore, the algebra generated by all prime T_p and R_p is abelian, and contains all T_n.

Recall that a function F on lattices is weakly modular if $F(\lambda L) = \lambda^{-k}F(L)$. This can be rewritten in terms of the operators we have defined, via the equation $F \circ R_\lambda = \lambda^{-k}F$. If we view R_λ as an operator on the set of lattice functions, the weakly modular functions are exactly the eigenvectors of eigenvalue λ^{-k}. Since T_n commutes with R_λ, we find

$$R_\lambda T_n F = T_n R_\lambda F = \lambda^{-k} T_n F$$

So that $T_n F$ is also weakly modular (we can extend a function on lattices to formal sums of lattices by linearity). In order to apply the operator T_n, we shall construct a more explicit formula.

Lemma 2.3. *The index of the lattice $L(az+bw, cz+dw)$ in $L(z,w)$ is the determinant of the matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$.*

Proof. We may reduce the matrix above into smith normal form by multiplying to the left and right by invertible integer matrices, which must therefore have determinant ± 1, and thus the lattice we obtain from the new matrix is the same as the original lattice, but now of the form $L(xz, yw)$. But then $L(z,w)/L(xz,yw) \cong \mathbf{Z}_x \oplus \mathbf{Z}_y$, so $L(xz,yw)$ has index xy in $L(z,w)$. □

Lemma 2.4. *For a fixed lattice $L(z,w)$, the map*

$$\begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \mapsto L(az+bw, dw)$$

is a bijection between matrices with $ad = n$ and $0 \leqslant b < d$, and index n sublattices of $L(z,w)$.

Proof. Certainly the image of any matrix with $ad = n$ is a sublattice of index n. On the other hand, let L' be an index n sublattice, and consider

$$Y = L/(L' + \mathbf{Z}w) \qquad Z = \mathbf{Z}w/(L' \cap \mathbf{Z}w)$$

Then Y is cyclic and generated by the image of z, and Z is cyclic generated by the image of w. Let Y have order a, and Z order d. We have an exact sequence

$$0 \to Z \to L/L' \to Y \to 0$$

so $ad = n$. If $w' = dw$, then $w' \in L'$. Furthermore, there is $z' \in L'$ such that $z' = az$ modulo $\mathbf{Z}w$. Write $z' = az + bw$. Then b is uniquely determined modulo d relative to z'. Thus we have shown the bijection between the two sets. □

Given a weakly modular weight k lattice function F, we may define a function $f(z) = F(z,1)$, and then one lets the operators T_n act on f, by

$$T_n f = n^{k-1} T_n F$$

The n^{k-1} is for now an arbitrary constant, but we shall see that if the q-expansion of f is integral, then the q expansion of $T_n f$ will be integral, when we include the factor of n^{k-1}. The previous lemma tells us that

$$T_n f(z) = n^{k-1} \sum_{ad=n} \sum_{b=0}^{d-1} F(az+b,d) = n^{k-1} \sum_{ad=n} \sum_{b=0}^{d-1} d^{-k} f\left(\frac{az+b}{d}\right)$$

By our previous discussion, we known f is weakly modular. The T_n also transforms the q coefficients of f in a predictable way, assuming f is a modular function.

Theorem 2.5. *If $f(z) = \sum b_m q^m$, then $T_n f = \sum c_m q^m$,*

$$c_m = \sum_{a \mid gcd(n,m)} a^{2k-1} b_{nm/a^2}$$

where we sum only over positive choices of a.

Proof. If we write $a/d = r$, then

$$\begin{aligned}
T_n f &= \sum_{ad=n} \sum_{m=-\infty}^{\infty} \left(\sum_{b=0}^{d-1} e^{2\pi i m b/d} \right) b_m n^{k-1} d^{-k} e^{2\pi i m a z/d} \\
&= \sum_{m=-\infty}^{\infty} \sum_{ad=n} b_{md} (n/d)^{k-1} q^{maz} \\
&= \sum_{r=-\infty}^{\infty} \left(\sum_{ad=n} \sum_{ma=r} b_{md} (n/d)^{k-1} \right) q^r \\
&= \sum_{r=-\infty}^{\infty} \left(\sum_{a \mid \gcd(n,r)} b_{rn/a^2} a^{k-1} \right) q^r
\end{aligned}$$

So the equation is obtained by combining terms. □

If f is a modular function, then b_m is zero for m negative enough. We can use the formula above to conclude that c_m is zero for m negative enough. In fact, if $b_m = 0$ for $m < -M$, then $c_m = 0$ for $m < -n^2 M$, for then $mn/a^2 \leqslant -Mn^3/a^2 < -M$, when $a \mid \gcd(m,n)$. Thus $T_n f$ is a modular function whenever f is a modular function, and is in fact a modular form if f is a modular form. What's more, we find by direct calculation that $c_0 = b_0 \sigma_{k-1}(n)$, $c_1 = b_n$, and if n is a prime number p, then $c_m = b_{mp}$ for all m reatively prime to p, because then $\gcd(m,p) = 1$.

2.4 Hecke Eigenfunctions

A modular form f is a Hecke eigenfunction if there is a sequence $z_0, z_1, z_2, \ldots$ such that $T_n f = z_n f$. The eigenfunctions form a subspace of all modular forms.

Theorem 2.6. *If $f = \sum a_n q^n$ is a non-zero Hecke eigenform, then $a_1 \neq 0$, and if $a_1 = 1$, then $a_n = z_n$ are the sequence of eigenvalues defining the eigenform.*

Proof. If $T_n f = \sum b_n q^n$, then we have already shown that $b_1 = a_n$, so that $a_n = z_n a_1$. If $a_1 = 0$, then $a_n = 0$ for all n, and hence $f = 0$. If $a_1 = 0$, then $a_n = z_n$. □

This also tells us that we can identify an eigenfunction by its eigenvalues, up to a scalar multiple. What's more, the q-coefficients of eigenfunctions are now seen to have very useful properties. By the composition properties of the operators T_n, we find that $a_m a_n = b_{mn}$ when m and n are relatively prime, and $a_p a_{p^n} = a_{p^{n+1}} + p^{k-1} a_{p^{n-1}}$ for p prime.

Given a Hecke eigenfunction, the q-coefficients are multiplicative, which allows us to consider the Dirichlet series of f, denoted

$$\phi_f(s) = \sum_{n=1}^{\infty} \frac{a_n}{n^s}$$

and for all s where the series is defined, we have

$$\sum_{n=1}^{\infty} \frac{a_n}{n^s} = \prod_{p \text{ prime}} (1 + a_p p^{-s} + \cdots + a_{p^n} p^{-ns} + \ldots)$$

essentially for the same reason that the product formula holds for the Riemann zeta function. In fact, we have the formula

$$\phi_f(s) = \prod_{p \text{ prime}} \frac{1}{1 - c_p p^{-s} + p^{k-1-2s}}$$

Which follows because of the recurrence relation for the coefficients of prime powers.

The first example of an eigenfunction we can construct is the Eisenstein series E_k. First we rely on the fact that we need only prove that E_k is

an eigenfunction of T_p for p prime, because all of the operators T_n are obtained from prime operators by composition. We reduce to the lattice definition of the series

$$E_k(L) = \sum_{\substack{x \in L \\ x \neq 0}} \frac{1}{|x|^k}$$

Now we have

$$T_p E_k(L) = \sum_{(L':L)=p} \sum_{\substack{x \in L' \\ x \neq 0}} \frac{1}{|x|^k}$$

Fix $x \in L$. If $x \in pL$ then $x \in L'$ for each of the $p+1$ index p sublattices of L. If x is not in pL then it belongs to exactly one of the sublattices L', since these sublattices intersect trivially. It follows that

$$T_p E_k(L) = E_k(L) + p E_k(pL) = (1 + p^{1-k}) E_k(L)$$

and thus E_k is an eigenfunction. This shows that the divisor functions σ_k are multiplicative, but this is already obvious from their definition.

The second example results from the low dimensionality of low order modular forms. The set of cusp forms of order 12 is one dimensional, because the set of all modular forms of order 12 is two dimensional, spanned by E_4^3 and E_3^4. Since $T_n\Delta$ is a cusp form of order 12 for every n, $T_n\Delta$ must be a multiple of Δ for each n, so Δ must be an eigenfunction! Since the coefficients of Δ are the values of the Ramanujan tau function, we have proved the famous result that τ is multiplicative.

Chapter 3

Congruence Modular Forms

Two subgroups H and K are commensurable if $H \cap K$ has finite index in both H and K. Given these two groups, for a fixed $x \in G$, we can consider the double coset HxK.

Theorem 3.1. *Let H be a subgroup of a group G and let $x \in G$ be such that H and $x^{-1}Hx$ are commensurable. Let $K = x^{-1}Hx \cap H$ and let $n = [H : K]$. Then if $y_1, \ldots, y_n$ is a set of coset representatives for H/K, then*

$$HxH = \bigcup_{j=1}^{n} Hxy_n$$

is a disjoint union of right cosets. Conversely, if HxH is a disjoint union of right cosets with representatives $y_1, \ldots, y_n$ then $H = \bigcup_{j=1}^{n} Ky_j$.

Proof. By coset arithmetic, we find

$$HxH = \bigcup_{j=1}^{n} HxKy_j \subset \bigcup_{j=1}^{n} Hxx^{-1}Hxy_j = \bigcup_{j=1}^{n} Hxy_j$$

That the reverse relation holds is trivial, because $K \subset H$. This is a disjoint union, for if $zxy_n = z'xy_m$ then $y_m y_n^{-1} = x^{-1}z'^{-1}zx \in x^{-1}Hx$, and trivially $y_m y_n^{-1} \in H$, so $y_m = y_n$.

Conversely, suppose that $HxH = \bigcup_{j=1}^{n} Hxy_n$. Then

$$H = Hxx^{-1} \subset HxHx^{-1} = \bigcup_{j=1}^{n} Hxy_n x^{-1} \subset \bigcup_{j=1}^{n} Ky_n$$

The last inequality follows because xy_nx^{-1} is in the same coset of K as y_n. □