MATH 996: HOPKINS SEMINAR

WENLIANG ZHANG

This is the seminar to accompany the mini-course taught by Professor Michael Hopkins in November 2014.

In preparation for the lectures, I have used the following resources:

- *Complex Analysis* by Lars Ahlfors
- *Introduction to Modular Forms* by Serge Lang
- *A first course in modular forms* by Fred Diamond and Jerry Shurman
- *A course in arithmetic* by Jean-Pierre Serre
- *Arithmetic of elliptic curves* by Joseph Silverman

Given the nature of the lectures, I've decided not to specify where any of the results discussed comes from. Needless to say, all errors are mine.

1. MODULAR CURVES

One of the underlying theme of this semester is group actions; we will see group action on sets, topological spaces, fields, functions, etc. We will start with group actions on sets.

Definition 1.1. A group action of a group G on a set $\mathcal{S}$ is an assignment $G \times \mathcal{S} \to \mathcal{S}$ such that

(1) $1_G \cdot s = s$ for all $s \in \mathcal{S}$, and
(2) $g_2 \cdot (g_1 \cdot s) = (g_2 g_1) \cdot s$.

Consider the fractional linear maps: for each $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in GL_2(\mathbb{C})$ and $z \in \mathbb{C}$, set $gz = \frac{az+b}{cz+d}$. It is straightforward to check that 1.1(1)&(2) were satisfied. However, we can't say it is a $GL_2(\mathbb{C})$-action on $\mathbb{C}$, because $\begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \cdot 1 = \frac{1}{0} \notin \mathbb{C}$. So, we need to throw in ∞.

For each $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, we define

$$\begin{cases} g \cdot \infty = \begin{cases} \frac{a}{c} & c \neq 0 \\ \infty & c = 0 \end{cases} \\ g \cdot (-\frac{d}{c}) = \infty \\ g \cdot z = \frac{az+b}{cz+d} \qquad\qquad \text{elsewhere} \end{cases}$$

Or, one can simple set $g \cdot z = \frac{az+b}{cz+d}$ and handle $\infty, -\frac{d}{c}$ by taking appropriate limits.

Exercise 1.2. Check that this defines a $GL_2(\mathbb{C})$-action on $\mathbb{C} \cup \{\infty\}$.

Proposition 1.3. *For $g, g_1 \in GL_2(\mathbb{C})$, we have $gz = g_1 z$ for all $z \in \mathbb{C} \cup \{\infty\}$ if and only if $g_1 = \lambda g$ for some $\lambda \in \mathbb{C}$.*

Proof. Assume that $g_1 = \lambda g$ for some $\lambda \in \mathbb{C}$. Then

$$g_1 z = \lambda g z = \frac{\lambda a z + \lambda b}{\lambda c z + \lambda d} = g z.$$

Conversely, assume that $gz = g_1 z$ for all $z \in \mathbb{C} \cup \{\infty\}$, it i.e. $\frac{a_1 z + b_1}{c_1 z + d_1} = \frac{az+b}{cz+d}$.

We claim that there are $\alpha, \beta \in \mathbb{C}$ such that $b_1 = \alpha b$, $d_1 = \alpha b$, $a_1 = \beta a$, and $c_1 = \beta c$ and we reason as follows. Clearly, $a_1 = 0 \Leftrightarrow g_1\infty = 0 \Leftrightarrow g\infty = 0 \Leftrightarrow a = 0$. And, likewise, $b_1 = 0 \Leftrightarrow b = 0$, $c_1 = 0 \Leftrightarrow c = 0$, and $d_1 = 0 \Leftrightarrow d = 0$. Hence we may assume that none of them is 0.

$$\begin{aligned}
g_1 0 = g0 \Rightarrow \frac{b_1}{d_1} = \frac{b}{d} &\Rightarrow b_1 = \alpha b,\ d_1 = \alpha d \\
g_1\infty = g\infty \Rightarrow \frac{a_1}{c_1} = \frac{a}{c} &\Rightarrow a_1 = \beta a,\ c_1 = \beta c \\
g_1 1 = g1 \Rightarrow \frac{a_1 + b_1}{c_1 + d_1} &= \frac{a+b}{c+d} \\
\Rightarrow \frac{a\beta + b\alpha}{c\beta + d\alpha} &= \frac{a+b}{c+d} \\
\Rightarrow (ad - bc)(\alpha - beta) &= 0 \\
\Rightarrow \alpha &= \beta
\end{aligned}$$

□

From the proof, we can see that if $g_1 z = gz$ for $z \in \{0, 1, \infty\}$, then $g_1 = g$. The same conclusion holds for any 3 distinct points.

Proposition 1.4. *Assume z_1, z_2, z_3 are 3 distinct elements of $\mathbb{C} \cup \{\infty\}$. If $gz_i = z_i$ for $i = 1, 2, 3$, then $g = \lambda I_2$.*

Proof. Consider $\sigma = \begin{pmatrix} z_2 - z_3 & -z_1(z_2 - z_3) \\ z_2 - z_1 & -z_3(z_2 - z_1) \end{pmatrix}$. $\Sigma \in \mathrm{GL}_2(\mathbb{C})$ since $\det(sigma) = (z_1 - z_3)(z_2 - z_3)(z_2 - z_1) \neq 0$. And $\sigma z = \frac{(z-z_1)(z_2-z_3)}{(z-z_3)(z_2-z_1)}$. So $\sigma z_1 = 0$, $\sigma z_2 = 1$, and $\sigma z_3 = \infty$. Hence $(\sigma g \sigma^{-1})0 = 0$, $(\sigma g \sigma^{-1})1 = 1$, $(\sigma g \sigma^{-1})\infty = \infty$. From Proposition 1.3, $\sigma g \sigma^{-1} = \lambda i_2$ and hence $g = \lambda I_2$. □

Remark 1.5. One can also see from the proof that the $GL_2(\mathbb{C})$-action on $\mathbb{C} \cup \{\infty\}$ is transitive.

The $GL_2(\mathbb{C})$-action on $\mathbb{C} \cup \{\infty\}$ enjoys many nice properties. Here is one that we may need later.

Exercise 1.6. Prove that each $g \in GL_2(\mathbb{C})$ sends {circles and lines} to {circles and lines}.

Next, we want to consider a subgroup of $GL_2(\mathbb{C})$: $SL_2(\mathbb{Z})$, and a subset of $\mathbb{C} \cup \{\infty\}$: $\mathfrak{h} := \{z \in \mathbb{C} \mid \mathrm{Im}(z) > 0\}$.

Proposition 1.7. *The assignment $\begin{pmatrix} a & b \\ c & d \end{pmatrix} z = \frac{az+b}{cz+d}$ defines a $SL_2(\mathbb{Z})$-action on $\mathfrak{h}$.*

Proof. Since 1.1(1)&(2) hold for $GL_2(\mathbb{C})$, they must hold for $SL_2(\mathbb{Z})$. It remains to show that $\frac{az+b}{cz+d}$ is still in $\mathfrak{h}$. Write $z = x + iy$, we have

$$\frac{az+b}{cz+d} = \frac{a(x+iy)+b)(cx+d-icy)}{(cx+d+icy)(cx+d-icy)} = \frac{1}{|cz+d|^2}((ax+b)(cx+d)+acy^2+iy)$$

Therefore,

$$\mathrm{Im}(gz) = \frac{\mathrm{Im}(z)}{|cz+d|^2} \tag{1.7.1}$$

Note that $|cz+d| \neq 0$; otherwise $z = -\frac{d}{c} \notin \mathfrak{h}$. □

There are two special elements in $SL_2(\mathbb{Z})$:

$$S := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \; Sz = -\frac{1}{z}, \; T := \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \; Tz = z+1.$$

Our first goal is to understand the space of orbits of $\mathfrak{h}$ under the $SL_2(\mathbb{Z})$-action. This end, we will consider

Definition 1.8 (Fundamental domain)**.** Let Γ be a subgroup of $\mathfrak{h}$ and $\mathcal{D}$ be a closed subset of $\mathfrak{h}$. $\mathcal{D}$ is called a *fundamental domain* of $\mathfrak{h}$ under the Γ-action if

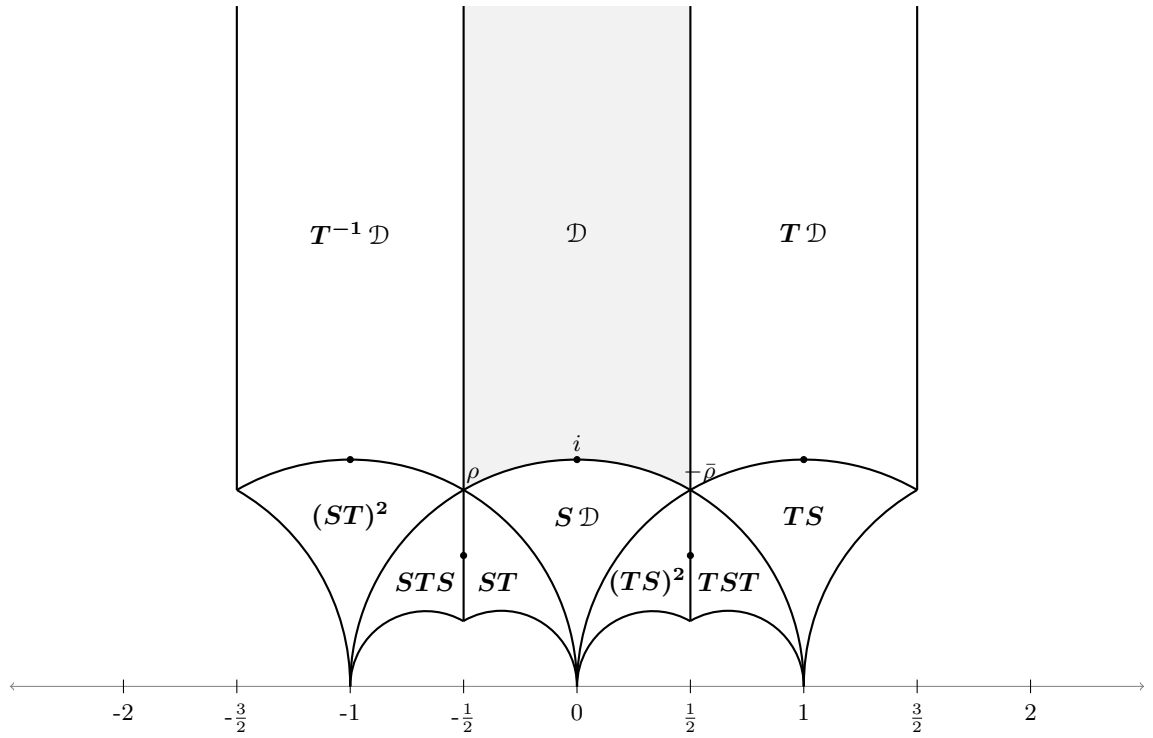
(1) $\mathfrak{h} = \bigcup_{\sigma\in\Gamma} \sigma\,\mathcal{D}$, and
(2) for any z_1, z_2 in the interior of $\mathfrak{h}$[1], $z_1 \notin \Gamma z_2$ (*i.e.* any two points in the interior of $\mathcal{D}$ belong to different orbits).

Set

$$\mathcal{D} := \{z \in \mathfrak{h} \mid |z| \geq 1 \text{ and } |\,\mathrm{Re}(z)| \leq \frac{1}{2}\}$$

Below is $\mathcal{D}$ and a few of its transformations under $SL_2(\mathbb{Z})$.

[1]under the topology on $\mathfrak{h}$ inherited from the natural topology on $\mathbb{C}$



Let G denote the subgroup of $SL_2(\mathbb{Z})$that is generated by S, T. Then

Theorem 1.9. *For each $z \in \mathfrak{h}$, there is a $g \in G$ such that $gz \in \mathcal{D}$.*

Proof. Given any $z \in \mathfrak{h}$ and a real umber $r > 0$, there are only finitely many pairs $(c, d) \in \mathbb{Z}^2$ such that $|cz+d| < r$ (one can see this as follows; $(cx+d)^2 + c^2y^2 < r^2 \Rightarrow c$ is bounded $\dots$). Therefore, from formula (1.7.1), there is a $g \in G$ such that $\mathrm{Im}(gz)$ is maximized (among all elements in G). We may choose n such that $\mathrm{Re}(T^n gz) \in [-\frac{1}{2}, \frac{1}{2}]$; note that $\mathrm{Im}(T^n gz) = \mathrm{Im}(gz)$. If $|T^n gz| < 1$, then we would have $\mathrm{Im}(ST^n gz) > \mathrm{Im}(T^n gz)$, a contradiction. Hence $|T^n gz| \geq 1$; consequently $T^n gz \in \mathcal{D}$, where $T^n g \in G$. □

To prove that $\mathcal{D}$ is a fundamental domain of $SL_2(\mathbb{Z})$, it remains to prove the following.

Theorem 1.10. *Let $z \in \mathcal{D}$. If there is $g \in SL_2(\mathbb{Z})$ such that $z \neq gz \in \mathcal{D}$, then both z and gz must be on the boundary of $\mathcal{D}$.*

Proof. Write $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Replacing (z, g) by (gz, g^1) if necessary, we may assume that $\mathrm{Im}(gz) \geq \mathrm{Im}(z)$, *i.e.* $|cz + d| \leq 1$. In particular, we have $|c| \leq 1$. There are 3 cases: $c = 0, 1, -1$.

Case 1: $c = 0$. Then $|d| \leq 1$ and hence $d = 1$ or -1. Then $a = 1$ or -1, respectively. So, g is the translation by b. Since $z, gz \in \mathcal{D}$, from the definition of $\mathcal{D}$, we can see that either $b = 0$ or $b = 1, -1$. If $b = 0$, then $g = I_2$ or $-I_2$; but then $gz = z$. If $b = 1, -1$, then by inspecting $\mathcal{D}$ one can see that z, gz must be on the lines $x = \frac{1}{2}$ and $x = -\frac{1}{2}$.

Case 2: $c = 1$. Then $|z + d| \leq 1$. Since d is an integer, by inspecting $\mathcal{D}$, we can see that, if $d \neq 0$, then $d = 1, -1$ and $z, gz \in \{\rho, -\bar{\rho}\}$. If $d = 0$, then $|z| \leq 1$. But $|z| \geq 1$ since $z \in \mathcal{D}$, we have $|z| = 1$, on the boundary of cD.

Case 3: $c = -1$. Similar to Case 2. □

Definition 1.11. Let Γ be a subgroup of $SL_2(\mathbb{Z})$. For each $z \in \mathfrak{h}$ and , all elements $\gamma \in \Gamma$ such that $\gamma z = z$ form a subgroup of Γ, called the isotropy subgroup and denoted by Γ_z.

Theorem 1.12. *For each $z \in \mathcal{D}$, we have $SL_2(\mathbb{Z})_z = \{\pm I_2\}$ except for 3 cases:*
(1) if $z = i$, then $SL_2(\mathbb{Z})_z = \langle S \rangle$;
(2) if $z = \rho$, then $SL_2(\mathbb{Z})_z = \langle ST \rangle$;
(3) if $z = -\bar{\rho}$, then $SL_2(\mathbb{Z})_z = \langle TS \rangle$.

Proof. Assume that $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z})_z$, *i.e.* $gz = z$. The formula for the imaginary part implies that $|cz + d| = 1$. Since $zin\,\mathcal{D}$, we have $\mathrm{Im}(z) \geq \frac{\sqrt{3}}{2}$ and hence $|c| \leq 1$. Also, we can see that $az + b = cz^2 + dz$. Then $c \neq 0$, otherwise $z \notin \mathfrak{h}$. The quadratic formula tells us that

$$z = \frac{(a-d) \pm \sqrt{(d-a)^2 + 4bc}}{2c} = \frac{(a-d) \pm \sqrt{(d+a)^2 - 4}}{2c}.$$

Since $\mathrm{Im}(z) > 0$, we have $|a + d| \leq 1$. Since $|\,\mathrm{Re}(z)| \leq \frac{1}{2}$, we have $|a - d| \leq 1$. Hence $a^2 + d^2 \leq 1$; either $a = 0 \& d = \pm 1$ or $d = 0 \& a = \pm 1$. Now one can simply check case by case and the rest of the proof is left as an exercise. □

Exercise 1.13. Finish the proof of Theorem 1.12.

Remark 1.14. There is a slightly more conceptual approach to proving Theorem 1.12. After one realizes $|a + d| \leq 1$, one can look at the characteristic polynomial of g which is (after simplification) $\lambda^2 - (a + d)\lambda + 1$. Hence there are 3 possibilities:

$$f_1 = \lambda^2 + 1,\ f_2 = \lambda^2 - \lambda + 1,\ f_3 = \lambda^2 + \lambda + 1.$$

Since $f_i(g) = 0$ for some $i = 1, 2, 3$, one can deduce that $g \in \langle S \rangle$, or $g \in \langle ST \rangle$, or $g \in \langle TS \rangle$.

Corollary 1.15. $SL_2(\mathbb{Z}) = G = \langle S, T \rangle$.

Proof. Given any $\sigma \in Sl_2(\mathbb{Z})$, we wish to show that $\sigma \in G$. Pick any point z_0 in the interior of $\mathcal{D}$ (*e.g.* $3i$). Set $z = \sigma z_0$. Theorem 1.9 implies that there is a $g \in G$ such that $g(\sigma z_0) \in \mathcal{D}$. Hence both z_0 and $(g\sigma)z_0$ are in $\mathcal{D}$ and z_0 is in the interior of $\mathcal{D}$, Theorem 1.10 implies that $z_0 = (g\sigma)z_0$. Again since z_0 is in the interior of $\mathcal{D}$, Theorem 1.12 implies that $g\sigma = \pm I_2$. Therefore $\sigma \in G$. □

Remark 1.16. As a matter of fact, one can show that $SL_2(\mathbb{Z})$ is the free product of $\langle S \rangle$ and $\langle ST \rangle$.

From we have seen so far, one can find a representative of each orbit of $SL_2(\mathbb{Z})$ in $\mathcal{D}$. Next, we want to treat the set of orbits as a topological space and $\mathcal{D}$ will help us visualize what this space is.

Definition 1.17. An action of a group G on a topological space X is an action of G on the underlying set X such that $x \mapsto gx$ is continuous for each $g \in G$.

Assume that a group G acts on a topological space X, the set of orbits $X/G := \{Gx \mid x \in X\}$ admits a quotient topology from X; namely a subset $U \subseteq X/G$ is open if $\pi^{-1}(U)$ is open in X where $\pi : X \to X/G$ is defined by $x \mapsto Gx$.

Exercise 1.18. Prove that under this quotient topology π is a continuous surjective open mapping.

Using $\mathcal{D}$, one can visualize $\mathfrak{h}/SL_2(\mathbb{Z})$ as follows: any pair of points in the interior of $\mathcal{D}$ represent two distinct points in $\mathfrak{h}/SL_2(\mathbb{Z})$; the line $x = -\frac{1}{2}$ should be glued to the line $x = \frac{1}{2}$; the arc between ρ and i should be glued to the arc between $-\bar{\rho}$ and i. One may view the resulted space as an envelop with infinite length. Therefore, topologically $\mathfrak{h}/SL_2(\mathbb{Z})$ is the same as $\mathbb{C}$.

Before we go any further, let's introduce some notation.

$\Gamma(N) := \{g \in Sl_2(\mathbb{Z}) \mid g \equiv I_2 \bmod N\}$, called the *principal congruence subgroup of* $Sl_2(\mathbb{Z})$ *of level* N. Clearly $X(1) = SL_2(\mathbb{Z})$. It should be clear that $\Gamma(N)$ is the kernel of the natural homomorphism $SL_2(\mathbb{Z}) \xrightarrow{\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \begin{pmatrix} \bar{a} & \bar{b} \\ \bar{c} & \bar{d} \end{pmatrix}} SL_2(\mathbb{Z}/N)$, and hence a normal subgroup of $SL_2(\mathbb{Z}) = \Gamma(1)$.

The quotient space $\mathfrak{h}/\Gamma(N)$ is denoted by $Y(N)$. Let $X(N)$ denote the compactification of $Y(N)$.

Exercise 1.19. (1) Prove that the natural map $SL_2(\mathbb{Z}) \xrightarrow{\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \begin{pmatrix} \bar{a} & \bar{b} \\ \bar{c} & \bar{d} \end{pmatrix}} SL_2(\mathbb{Z}/N)$ is surjective.

(2) Find the order of $SL_2(\mathbb{Z}/N)$.

We have seen that $Y(1)$ is homeomorphic to $\mathbb{C}$. The compactification of $\mathbb{C}$ is $\mathbb{C} \cup \{\infty\}$ (which $GL_2(\mathbb{C})$ acts on as we discussed in our first lecture), where an open neighborhood of ∞ is given by the complement of a compact subset of $\mathbb{C}$. One can directly that $\mathbb{C} \cup \{\infty\}$ is compact under this topology, or one can use the stereographic projection to give a homeomorphism between $\mathbb{C} \cup \{\infty\}$ and the unit 2-sphere S^2 as follows.

Write S^2 as $\{(x_1, x_2, x_3) \mid \sum_i x_i^2 = 1\}$ (in real coordinates). Define a function $p : S^2 \to \mathbb{C} \cup \{\infty\}$ by

$$p(x_1, x_2, x_3) = \begin{cases} \frac{x_1}{1-x_3} + i\frac{x_2}{1-x_3} & (x_1, x_2, x_3) \neq (0, 0, 1) \\ \infty & (x_1, x_2, x_3) = (0, 0, 1) \end{cases}$$

One can check that p is continuous and bijective.

Hence $X(1) = S^2$.

The fundamental domain $\mathcal{D}$ of $\Gamma(1)$ has been very helpful in determining $Y(1)$ and $X(1)$. One can actually find a fundamental domain for each $\Gamma(N)$ from $\mathcal{D}$.

Remark 1.20. Write $\Gamma(1) = \bigcup_i \Gamma(N) g_i$. Then we have

$$\mathfrak{h} = \bigcup_{g \in \Gamma(1)} g\,\mathcal{D} = \bigcup_{g \in \cup_i \Gamma(N) g_i} = \bigcup_{g \in \Gamma(N)} (\bigcup_i g g_i\,\mathcal{D}) = \bigcup_{g \in \Gamma(N)} g(\bigcup_i g_i\,\mathcal{D}).$$

One may use $\bigcup_i g_i\,\mathcal{D}$ as a fundamental domain of $\Gamma(N)$. However, this doesn't really help us visualize $Y(N)$, since the number of posets are usually big.

We will follow a different strategy to study $Y(N)(N \geq 2)$: we will prove that $X(N)$ is a compact Riemann surface and use the classification of compact Riemann surfaces. First, we need to know if $Y(N)$ and $X(N)$ are Hausdorff.

Definition 1.21. A group G acts *properly discontinuously* on a topological space X if, for any two points $x_1, x_2 \in X$, there are open neighborhoods U_1, U_2 of x_1, x_2 respectively such that $\mid \{g \in G \mid gU_1 \cap U_2 \neq \emptyset\} \mid < \infty$.

Proposition 1.22. *Each subgroup $\Gamma \subseteq SL_2(\mathbb{Z})$ acts properly discontinuously on $\mathfrak{h}$.*

Proof. The idea is to find an open neighborhood that is bounded and bounded away from 0 in y-coordinate. For each $z_0 = x_0 + iy_0 \in \mathfrak{h}$, set $U := \{z \in \mathfrak{h} \mid |z - z_0| < \frac{y_0}{2}\}$. Then we have $\mathrm{Im}(z) > \frac{y_0}{2}$ for each $z \in U$. Assume $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is an element of Γ such that $gU \cap U \neq \emptyset$. Then there is a $z \in U$ such that $gz \in U$. Then we have $\mathrm{Im}(gz) = \frac{\mathrm{Im}(z)}{|cz+d|^2} > \frac{y_0}{2}$. It follows immediately that there are only finitely choices for the pair (c, d). Considering the real part of gz results in finitely choices for the pair (a, b). This finishes the proof. □

Recall that we proved the following last time.

Proposition 1.23. *Each subgroup $\Gamma \subseteq SL_2(\mathbb{Z})$ acts properly discontinuously on $\mathfrak{h}$.*

Our goal today is to show that $Y(N) := \mathfrak{h}/\Gamma(N)$ is Hausdorff. We will start with a lemma.

Lemma 1.24. *Let G be a group acting on a topological space X. Assume that for two points $x_1, x_2 \in X$ there are open neighborhoods U_1, U_2 of x_1, x_2 respectively such that $gU_1 \cap U_2 \neq \emptyset$ if and only if $gx_1 = x_2$. Then X/G is Hausdorff.*

Proof. Let $\pi : X \to X/G$ be the natural surjection. Pick two distinct points $\pi(x_1) \neq \pi(x_2) \in X/G$. Let U_1, U_2 be open neighborhoods of x_1, x_2, respectively, as in the assumptions. Then we have

$$\begin{aligned} \pi(U_1) \cap \pi(U_2) \neq \emptyset &\Leftrightarrow \exists g \in G \text{ such that } gU_1 \cap U_2 \neq \emptyset \\ &\Leftrightarrow \exists g \in G \text{ such that } gx_2 = x_2 \text{ (by assumptions)} \\ &\Leftrightarrow \pi(x_1) = \pi(x_2), \text{ a contradiction} \end{aligned}$$

Hence $\pi(U_1), \pi(x_2)$ are disjoint open neighborhoods of $\pi(x_1), \pi(x_2)$ respectively. This finishes the proof. □

Theorem 1.25. *Let X be a Hausdorff topological space on which a group G acts properly discontinuously. Then assumptions in Lemma 1.24 are satisfied. In particular, X/G is Hausdorff.*

Proof. Let $x_1 \neq x_2$ be two distinct points in X. Since the action is properly discontinuous, there are open neighborhoods V_1, V_2 of x_1, x_2 respectively such that

$$| \{g \in G \mid gV_1 \cap V_2 \neq \emptyset\} | < \infty.$$

Write $\{g \in G \mid gV_1 \cap V_2 \neq \emptyset\} = \{g_1, \ldots, g_n\}$ and we may assume that there is an ℓ such that

$$g_1x_1 = x_2, \ldots, g_\ell x_1 = x_2, g_{\ell+1}x_1 \neq x_2, \ldots, g_n x_1 \neq x_2.$$

Since X is Hausdorff, for each $i = \ell + 1, \ldots, n$, there are open neighborhoods W_i of $g_i x_1$ of W_i' of x_2 such that $W - i \cap W_i' = \emptyset$. Set

$$U_1 = V_1 \bigcap (\bigcap_{i=\ell+1}^{n} g_i^{-1} W_i) \text{ and } U_2 = V_2 \bigcap (\bigcap_{i=\ell+1}^{n} W_i').$$

Then we have $gU_1 \cap U_2 \neq \emptyset \Leftrightarrow g \in \{g_1, \ldots, g_\ell\} \Leftrightarrow gx_1 = x_2$. This shows that the assumptions in Lemma 1.24 are satisfied and finishes the proof of our theorem. □

Combining Proposition 1.23 and Theorem 1.25, we have

Theorem 1.26. *$Y(N)$ is Hausdorff.*

Our ultimate goal is to apply the classification of compact Riemann surfaces to understand $Y(N)$. Unfortunately, $Y(N)$ is never compact. So, we need to compactify it. Recall that, to do so for $Y(1)$, we simply added ∞. Naturally, we want to consider $\mathfrak{h} \cup \{\infty\}$. But, $\Gamma(N)$ does *not* act on $\mathfrak{h} \cup \{\infty\}$:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \infty = \frac{a}{c} \in \mathbb{Q} \not\subset \mathfrak{h} \cup \{\infty\} \text{ (when } c \neq 0).$$

Hence it is necessary to add in $\mathbb{Q}$. As we will see it is also sufficient to add in $\mathbb{Q}$.

Set

$$\mathfrak{h}^* := \mathfrak{h} \cup \mathbb{Q} \cup \{\infty\}.$$

On the face of it, we are adding a lot more than we did for $Y(1)$. To reconcile,

Proposition 1.27. *$\Gamma(1) = SL_2(\mathbb{Z})$ acts transitively on $\mathbb{Q} \cup \{\infty\}$. Hence $\mathbb{Q} \cup \{\infty\}$ still contributes only one point, ∞, to $Y(1)$.*

Proof. For any $\frac{p}{q} \in \mathbb{Q}$, we may assume that $(p, q) = 1$. Then there are $r, s \in \mathbb{Z}$ such that $ps - qs = 1$. Then $\begin{pmatrix} p & r \\ q & s \end{pmatrix} \in Sl_2(\mathbb{Z})$ and $\begin{pmatrix} p & r \\ q & s \end{pmatrix} \infty = \frac{p}{q}$. □

Corollary 1.28. *$\mathbb{Q} \cup \{\infty\}$ has finitely many $\Gamma(N)$-orbits.*

Proof. This follows immediately from $[\Gamma(1) : \Gamma(N)] < \infty$. □

It is clear that $\Gamma(N)$ acts on $\mathfrak{h}$ and $\mathbb{Q} \cup \{\infty\}$ separately. hence $\mathfrak{h}^*/\Gamma(N)$ is the (disjoint) union of $Y(N) = \mathfrak{h}/\Gamma(N)$ and finitely many points from $(\mathbb{Q} \cup \{\infty\})/\Gamma(N)$.

Definition 1.29. Each point in $(\mathbb{Q} \cup \{\infty\})/\Gamma(N)$ is called a *cusp*.

So far, $\mathfrak{h}^*$ is just a set; we need to give it a topology. To start off, the topology on $\mathfrak{h}$ is the same as before. A basis of open neighborhoods of ∞ is given by

$$\mathcal{N}_\lambda := \{z \in \mathfrak{h} \mid \mathrm{Im}(z) > \lambda\} \cup \{\infty\},$$

where $\lambda \in \mathbb{R}_{>0}$.

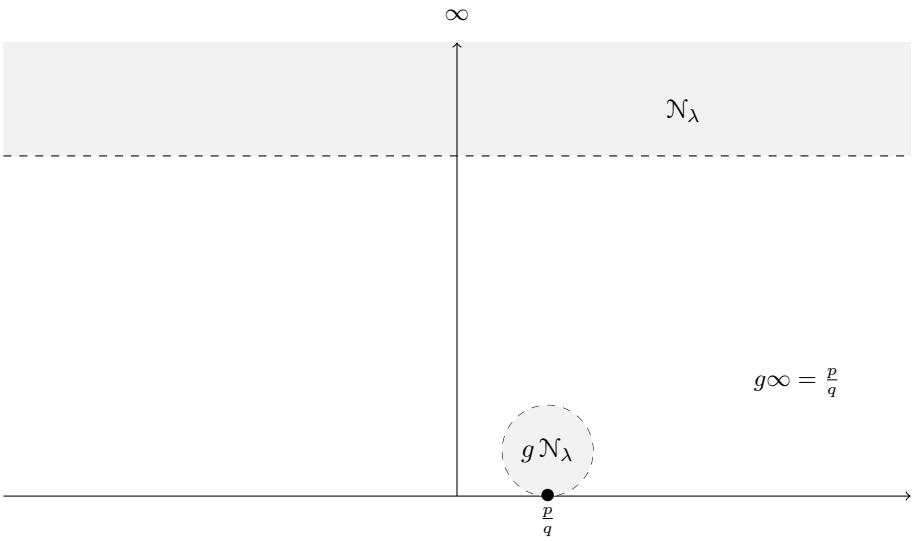
Lastly, to define the topology around each $\frac{p}{q} \in \mathbb{Q}$, we will use Proposition 1.27: there is a $g = \begin{pmatrix} p & r \\ q & s \end{pmatrix} \in SL_2(\mathbb{Z})$ such that $g\infty = \frac{p}{q}$. A basis of open neighborhoods of $\frac{p}{q}$ is given by $\{g\,\mathcal{N}_\lambda\}_\lambda$. If you have done Exercise 1.6, then it should be easy to visualize $g\,\mathcal{N}_\lambda$. In any

case, consider the following.

$$\begin{aligned} z \in g\,\mathcal{N}_\lambda &\Leftrightarrow g^{-1}z \in \mathcal{N}_\lambda \\ &\Leftrightarrow \mathrm{Im}(g^{-1}z) > \lambda \\ &\Leftrightarrow \frac{\mathrm{Im}(z)}{|-qz+p|^2} > \lambda, \text{ since } g^{-1} = \begin{pmatrix} s & -r \\ -q & p \end{pmatrix} \\ &\Leftrightarrow |-qz+p|^2 < \frac{\mathrm{Im}(z)}{\lambda} \\ &\Leftrightarrow (p-qx)^2 + q^2y^2, \frac{y}{\lambda}, \ z = x+iy \\ &\Leftrightarrow (x-\frac{p}{q})^2 + (y - \frac{1}{2\lambda q^2}) < (\frac{1}{2\lambda q^2})^2 \end{aligned}$$

Hence, the line $y = \lambda$ and the region $\{z \in \mathfrak{h} \mid \mathrm{Im}(z) > \lambda\}$ are turned into the circle $\{x+iy \in \mathfrak{h} \mid (x-\frac{p}{q})^2 + (y - \frac{1}{2\lambda q^2}) < (\frac{1}{2\lambda q^2})^2\}$ and the open disc, $\mathbb{D}((\frac{p}{q}, \frac{1}{2\lambda q^2}), \frac{1}{2\lambda q^2})$, centered at $(\frac{p}{q}, \frac{1}{2\lambda q^2})$ with radius $\frac{1}{2\lambda q^2}$, respectively.

Therefore, a basis of open neighborhoods of $\frac{p}{q}$ is given by $\left\{\mathbb{D}((\frac{p}{q}, \frac{1}{2\lambda q^2}), \frac{1}{2\lambda q^2}) \cup \{\frac{p}{q}\}\right\}_\lambda$.



From this calculation, we can also get the following.

Proposition 1.30. *Assume that $\lambda \geq 1$. Then $g\,\mathcal{N}_\lambda \cap \mathcal{N}_\lambda \neq \emptyset \Leftrightarrow g\infty = \infty$.*

Exercise 1.31. The isotropy group $\Gamma(N)_\infty$ of ∞ is $\{\begin{pmatrix} 1 & nN \\ 0 & 1 \end{pmatrix} \mid n \in \mathbb{Z}\}$.

Exercise 1.32. Prove that $\mathcal{D} \cup \{\infty\}$ is a compact subset of $\mathfrak{h}^*$ (under the topology on $\mathfrak{h}^*$).

We want to show that $\mathfrak{h}^*/\Gamma(N)$ is

(1) Hausdorff,
(2) connected,
(3) compact, and
(4) a complex manifold.

We begin by proving that $\mathfrak{h}^*/\Gamma(N)$ is Hausdorff.

Remark 1.33. $\Gamma(N)$ doesn't act properly discontinuously on $\mathfrak{h}^*$: $\Gamma(N)_\infty$ is infinite. Hence Theorem 1.25 is not applicable.

We have to directly that $\mathfrak{h}^*/\Gamma(N)$ is Hausdorff. Let $\pi : \mathfrak{h}^* \to \mathfrak{h}^*/\Gamma(N)$ be the natural surjection and let $\pi(x_1), \pi(x_2)$ be two distinct points in $\mathfrak{h}^*/\Gamma(N)$. We wish to find disjoint neighborhoods U_1, U_2 of $\pi(x_1), \pi(x_2))$ respectively. We will consider 3 cases:

(1) $x_1, x_2 \in \mathfrak{h}$,
(2) $x_1 \in \mathfrak{h}$ and $x_2 \in \mathbb{Q} \cup \{\infty\}$, and
(3) $x_1, x_2 \in \mathbb{Q} \cup \{\infty\}$.

Since $\mathfrak{h}/\Gamma(N)$ is Hausdorff, there are disjoint open neighborhoods U_1, U_2 of $\pi(x_1), \pi(x_2))$. This finishes Case (1).

Case (3): there $g_1, g_2 \in SL_2(\mathbb{Z})$ such that $g_i\infty = x_i$. Fix $\lambda \geq 1$. Set $U_i = g_i(\mathcal{N}_\lambda)$. We claim that $\pi(U_1)$ and $\pi(U_2)$ are disjoint and we reason as follows. Assume otherwise, then $\exists g \in \Gamma(N)$ and $y_1, y_2 \in \mathcal{N}_\lambda$ such that $gg_1(y_1) = g_2(y_2)$. Then $g_2^{-1}gg_1(y_1) = y_2 \in \mathcal{N}_\lambda$, hence $g_2^{-1}gg_1\,\mathcal{N}_\lambda \cap \mathcal{N}_\lambda \neq \emptyset$. By Proposition 1.30, we would have $g_2^{-1}gg_1\infty = \infty$. Since $g_i\infty = x_i$, we would have $gx_1 = x_2$ *i.e.* $\pi(x_1) = \pi(x_2)$, a contradiction.

Exercise 1.34. Assume that $x_1 \in \mathfrak{h}$ and $x_2 \in \mathbb{Q} \cup \{\infty\}$. Find disjoint open neighborhoods U_1, U_2 of $\pi(x_1), \pi(x_2))$.

Next, connectedness. It is straightforward to see that $\mathfrak{h}^*$ itself is connected. Since π is a surjective continuous map, $\mathfrak{h}^*/\Gamma(N)$ must be connected as well.

Compactness: write $\Gamma(1) = SL_2(\mathbb{Z}) = \cup_i \Gamma(N)g_i$. Since $\mathcal{D}$ is a fundamental domain of $\mathfrak{h}$ under $\Gamma(1) = SL_2(\mathbb{Z})$ and $\Gamma(1)$ acts transitively on $\mathbb{Q} \cup \{\infty\}$, we have

$$\mathfrak{h}^* = \bigcup_{g \in \Gamma(1)} g(\mathcal{D} \cup \{\infty\}) = \bigcup_{g \in \Gamma(N)} (g(\bigcup_i g_i(\mathcal{D} \cup \{\infty\}))).$$

Hence

$$\mathfrak{h}^*/\Gamma(N) = \bigcup_i \pi(g_i(\mathcal{D} \cup \{\infty\})),$$

a finite union of compact subsets (the compactness of $g_i(\mathcal{D} \cup \{\infty\})$ is from Exercise 1.32), must be compact.

Definition 1.35. Let X be a Hausdorff topological space. X is called a 1-*dimensional complex manifold* (or X *has a complex structure*) if X admits an open covering $\{U_\alpha\}_\alpha$ and homeomorphisms $f_\alpha : U_\alpha \to W_\alpha$ where W_α is a connected open subset of $\mathbb{C}$ such that, whenever $U_\alpha \cap U_\beta \neq \emptyset$,

$$f_\beta \circ f_\alpha^{-1} : f_\alpha(U_\alpha \cap U_\beta) \to f_\beta(U_\alpha \cap U_\beta)$$

is holomorphic.

Example 1.36 (S^2)**.** We have seen that there is a stereographic projection from the north pole $(0, 0, 1)$ which induces

$$f_1 : U_1 = S^2 \backslash (0, 0, 1) \to \mathbb{C},\ (x_1, x_2, x_3) \mapsto \frac{x_1}{1 - x_3} + i\frac{x_2}{1 - x_3}$$

Similarly, we have a projection from the south poly $(0, 0, -1)$ which induces

$$f_2 : U_2 = S^2 \backslash (0, 0, -1) \to \mathbb{C},\ (x_1, x_2, x_3) \mapsto \frac{x_1}{1 + x_3} - i\frac{x_2}{1 + x_3}$$

Then we have

$$f_2 \circ f_1^{-1} : \mathbb{C}\backslash\{0\} \to \mathbb{C}\backslash\{0\}$$

$$z = x + iy \mapsto (\frac{2x}{1+x^2+y^2}, \frac{2y}{1+x^2+y^2}, \frac{x^2+y^2-1}{1+x^2+y^2}) \mapsto \frac{x}{x^2+y^2} - i\frac{y}{x^2+y^2} = \frac{1}{z}$$

which is holomorphic.

Example 1.37 ($\mathbb{P}^1(\mathbb{C})$)**.** Define an equivalence relation on $\mathbb{C}^2\backslash\{(0,0)\}$ by $(z_1, z_2) \sim (z_1', z_2')$ if $(z_1, z_2) = \lambda(z_1', z_2')$ for some $\lambda \neq 0 \in \mathbb{C}$. Then $\mathbb{P}^1(\mathbb{C})$ is defined to be $\mathbb{C}^2/\sim$. Each element in $\mathbb{P}^1(\mathbb{C})$ is denoted by $[z_1, z_2]$ which is the set $\{\lambda(z_1, z_2) \mid \lambda \in \mathbb{C}\backslash\{0\}\}$.

$\mathbb{P}^1(\mathbb{C})$ is called the complex projective line.

$\mathbb{P}^1(\mathbb{C})$ admits a complex structure as follows.

$$U_1 = \{[z_1, z_2] \mid z_2 \neq 0\} \xrightarrow{f_1} \mathbb{C},\ [z_1, z_2] \mapsto \frac{z_1}{z_2}$$

$$U_2 = \{[z_1, z_2] \mid z_1 \neq 0\} \xrightarrow{f_1} \mathbb{C},\ [z_1, z_2] \mapsto \frac{z_2}{z_1}$$

Clearly $f_1^{-1}(z) = [z, 1]$, $f_2^{-1}(z) = [1, z]$ and hence $f_2 \circ f_1^{-1} : \mathbb{C}\backslash\{0\} \to \mathbb{C}\backslash\{0\}$ sends z to $\frac{1}{z}$ and is holomorphic.

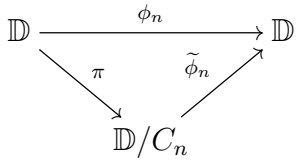
Remark 1.38. We can see that S^2 and $\mathbb{P}^1(\mathbb{C})$ share the same transition functions. As a matter of fact, they are the same; both can be viewed as $\mathbb{C} \cup \{\infty\}$.

Let's look at one more example that involves a group action.

Example 1.39. Set $\mathbb{D} := \{z \in \mathbb{C} \mid |z| < 1\}$ and $\omega_n = e^{\frac{2\pi i}{n}}$. Let $C_n = \langle\sigma\rangle$ denote the cyclic group of order n. Then C_n acts on $\mathbb{D}$ by $\sigma^j \cdot z = \omega_n^j z$. We want to understand $\mathbb{D}/C_n$ (and to give a complex structure to it).

The key is to find a function on $\mathbb{D}$ that is invariant under C_n; the exact same idea will appear in the construction of charts of $X(N)$.

Consider $phi_n : \mathbb{D} \to \mathbb{D}$ defined by $z \mapsto z^n$. Clearly $\phi_n(\sigma^j \cdot z) = \phi_n(z)$. Therefore ϕ_n induces $\widetilde{\phi}_n : \mathbb{D}/C_n \to \mathbb{D}$ which sends the orbit $C_n \cdot z$ to $\phi_n(z)$. One can check that $\widetilde{\phi}_n$ is well-defined. We have a commutative diagram:



We claim that $\widetilde{\phi}_n$ is a homeomorphism. Since ϕ_n is surjective, so is $\widetilde{\phi}_n$. Assume that $\widetilde{\phi}_n(C_n \cdot z_1) = \widetilde{\phi}_n(C_n \cdot Z_2)$. Then $\phi_n(z_1) = \phi_n(z_2)$, *i.e.* $z_1^n = z_2^n$, *i.e.* $z_1 = \sigma^j z_2$ for some j, *i.e.* $C_n \cdot z_1 = C_n \cdot z_2$. This shows that $\widetilde{\phi}_n$ is also injective.

Recall that $X(N) = \mathfrak{h}^*/\Gamma(N) = \mathfrak{h}/\Gamma(N) \cup (\mathbb{Q} \cup \{\infty\})/\Gamma(N)$. On the face of it, we have two types of points. In reality, there are 3 types of points. Recall that, in $Y(1) = \mathfrak{h}/\Gamma(1)$, the points i and ρ are fixed by elements other than $\pm I_2$. (Intuitively, but not rigorously, i and ρ give us two corner points when we glue the lines $x = 1/2$ and $x = -1/2$ and the two arcs to form $Y(1)$; they are seemingly different from other points in $Y(1)$.)

Definition 1.40. A point $z \in \mathfrak{h}$ is called an *elliptic point* of $\Gamma(N)$ (or an elliptic point of $\mathfrak{h}$ under the $\Gamma(N)$-action) if there exists $g \in \Gamma(N)$ such that $gz = z$ and $g \neq \pm I_2$.

A point $z \in \mathfrak{h}$ is called an *ordinary point* if it is not an elliptic point.

Now the benefit of restricting ourselves to $\Gamma(N)$, instead of a subgroup of $SL_2(\mathbb{Z})$ with finite index, is clear.

Proposition 1.41. *If $N \geq 2$, $\Gamma(N)$ has no elliptic points.*

Proof. Exercise. □

Recall that, when we discussed the $\Gamma(N)$-action on $\mathfrak{h}$, we proved that

Proposition 1.42. *For each $z \in \mathfrak{h}$ there is an open neighborhood U of z such that $gU \cap U \neq \emptyset$ if and only if $gz = z$.*

We start with ordinary points: let $z \in \mathfrak{h}$ be ordinary and U be the open neighborhood as in Proposition 1.42, then $gU \cap U \neq \emptyset \Leftrightarrow g = \pm I_2$. Consequently, $U \xrightarrow{\pi} \pi(U) \subseteq X(N)$ is an homeomorphism. Since U is open in $\mathfrak{h}$ and hence open in $\mathbb{C}$, we can simply take $(\pi(U), \pi^{-1})$ as our chart at z, an ordinary point.

Next, elliptic points. From Proposition 1.41, we know that we only to construct charts for i and ρ for $X(1)$.

(i) Let V be the open neighborhood in Proposition 1.42. Hence $gV \cap V \neq \emptyset$ if and only if $g \in \langle S \rangle$ ($S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$). Set $U = V \cap SV$. Then U is invariant under S and $\pi(U) \xrightarrow{\sim} U/\langle S \rangle$, where π is the natural surjection $\mathfrak{h} \to Y(N) = \mathfrak{h}/\Gamma(N)$.

Let $\varphi : \mathfrak{h} \to \mathbb{D}$ be defined by $z \mapsto \frac{z-i}{z+i}$ (one of the exercises in our first lecture asserts that φ is holomorphic and bijective). We have the following commutative diagram:

$$\begin{array}{ccc} U & \xrightarrow{\varphi} & \varphi(U) \\ \Big\downarrow{\scriptstyle S} & & \Big\downarrow{\scriptstyle z \mapsto -z} \\ U & \xrightarrow{\varphi} & \varphi(U) \end{array}$$

whose commutativity can be checked as follows

$$\varphi(S(z)) = \varphi(-\frac{1}{z}) = \frac{-\frac{1}{z} - i}{-\frac{1}{z} + i} = \frac{-z + i}{z + i} = -\frac{z - i}{z + i} = -\varphi(z).$$

The map $z \mapsto -z$ on $\varphi(U) \subseteq \mathbb{D}$ is the same as the C_2-action. Hence we will consider the function $z \mapsto z^2$ as follows.

$$\begin{array}{ccc} U/\langle S \rangle & \xrightarrow{\sim} & \varphi(U)/C_2 \\ \Big\downarrow{\scriptstyle \sim} & & \Big\downarrow{\scriptstyle z \mapsto z^2} \\ \pi(U) & \xrightarrow{z \mapsto (\frac{z-i}{z+i})^2} & \mathbb{D} \end{array}$$

Therefore, we can take $(\pi(U), (\frac{z-i}{z+i})^2)$ as a local chart at i.

(ρ) The construction is entirely similar to the one of i, so I will be brief. Let V be the open neighborhood as in Proposition 1.42 and set $U = V \cap (ST)V \cap (ST)^2V$. (Then U is invariant under $\langle ST \rangle$.) Let $\psi : U \to \mathbb{D}$ be defined by $z \mapsto \frac{z-\rho}{z-\bar{\rho}}$. The following commutative diagram

$$\begin{array}{ccc} U/\langle ST\rangle & \xrightarrow{\sim} & \psi(U)/C_3 \\ \downarrow \sim & & \downarrow z\mapsto z^3 \\ \pi(U) & \xrightarrow{(\frac{z-\rho}{z-\bar{\rho}})^3} & \mathbb{D} \end{array}$$

tells us that we can take $(\pi(U), (\frac{z-\rho}{z-\bar{\rho}})^3)$ as a local chart of ρ.

Finally, cusps. Since each $\frac{p}{q} \in \mathbb{Q}$ can be hit by ∞ by element in $\Gamma(1) = SL_2(\mathbb{Z})$, it suffices to construct a local chart of ∞ (then we can use the same trick as we defined topology of $\mathfrak{h}^*$ around $\frac{p}{q}$).

Recall that an open neighborhood of ∞ is given by $\mathcal{N}_\lambda := \{z \in \mathfrak{h} \mid Im(z) > \lambda\} \cup \{\infty\}$. Let's assume $\lambda \geq 1$. The isotropy subgroup of ∞ in $\Gamma(N)$ is $\langle \begin{pmatrix} 1 & N \\ 0 & 1 \end{pmatrix} \rangle$ (clearly $\begin{pmatrix} 1 & N \\ 0 & 1 \end{pmatrix} z = z + N$). Hence

$$\pi(\mathcal{N}_\lambda) = \mathcal{N}_\lambda / \langle \begin{pmatrix} 1 & N \\ 0 & 1 \end{pmatrix} \rangle.$$

As before, we need to find a function invariant under $\langle \begin{pmatrix} 1 & N \\ 0 & 1 \end{pmatrix} \rangle$. One such function is $e^{\frac{2\pi i z}{N}}$. Hence we can take $(\pi(\mathcal{N}_\lambda), e^{\frac{2\pi i z}{N}} : \pi(\mathcal{N}_\lambda) \to \mathbb{D}(0, e^{-\frac{2\pi\lambda}{N}}))$ as a local chart at ∞, where $\mathbb{D}(0, e^{-\frac{2\pi\lambda}{N}})) = \{z \mid |z| < e^{-\frac{2\pi\lambda}{N}})\}$.

Checking that the transition functions are holomorphic is left to you.

Remark 1.43. So far, we have shown that each $X(N) = \mathfrak{h}^*/\Gamma(N)$ is a compact Reimann surface. More generally and quite similarly to we have seen, given any subgroup Γ of $SL_2(\mathbb{Z})$ with finite index, one can show that $X(\Gamma) = \mathfrak{h}^*/\Gamma$ is a compact Riemann surface.

For the rest of this lecture, we turn to cusps of $X(N)$. For instance, we want to the number of cusps of $X(N)$. There are different approaches. We will start with a computational approach, which will actually tell us what the cusps are; next time, we will follow a more conceptual approach.

We will use $\binom{p}{q}$ to denote an element of $\mathbb{Q} \cup \{\infty\}$. As usual, we assume that $(p, q) = 1$. You may think of $\binom{p}{q}$ as $\frac{p}{q}$. We use $\binom{p}{q}$ since it is more convenient to treat the $\Gamma(N)$-action on $\mathbb{Q} \cup \{\infty\}$ as matrix multiplication. Clearly $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} ap + bq \\ cp + dq \end{pmatrix}$. Keep in mind that $\pm\binom{p}{q}$ represent the same number $\frac{p}{q}$.

Proposition 1.44. *$r = \binom{p}{q}$ and $r' = \binom{p'}{q'}$ are in the same orbit, i.e. $\Gamma(N)r = \Gamma(N)r'$, if and only if $\binom{p}{q} \equiv \pm\binom{p'}{q'}$ mod N.*

Proof. Assume that $\Gamma(N)r = \Gamma(N)r'$, *i.e.* $\binom{ap+bq}{cp+dq} = \binom{p'}{q'}$. Since $a, d \equiv 1 \bmod N$ and $b, c \equiv 0 \bmod N$, we have that $\binom{p}{q} \equiv \pm\binom{p'}{q'} \bmod N$.

Conversely, assume that $\binom{p}{q} \equiv \binom{p'}{q'} \bmod N$. We will handle the special case when $\binom{p}{q} = \binom{1}{0} = \infty$ first and then use the transitivity of the $SL_2(\mathbb{Z})$-action on $\mathbb{Q} \cup \{\infty\}$ to get the general case.

Assume that $\binom{p'}{q'} \equiv \binom{1}{0} \bmod N$. Consequently $p' = cN + 1$ for some $c \in \mathbb{Z}$ and $N \mid q'$. Since $(p', q') = 1$, there are $a, b \in \mathbb{Z}$ such that $ap' + bq' = 1$. Multiplying the both sides of it by $p' - 1 = cN$, we get

$$(1 - acN)p' - bcNq' = 1.$$

Now we have $\begin{pmatrix} p' & bcN \\ q' & 1 - acN \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} p' \\ q' \end{pmatrix}$ and $\begin{pmatrix} p' & bcN \\ q' & 1 - acN \end{pmatrix} \in \Gamma(N)$.

For each $\binom{p}{q}$, there is a $\sigma \in SL_2(\mathbb{Z})$ such that $\sigma\binom{1}{0} = \binom{p}{q}$. Then we have

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \equiv \sigma^{-1} \begin{pmatrix} p \\ q \end{pmatrix} \equiv \sigma^{-1} \begin{pmatrix} p' \\ q' \end{pmatrix} \bmod N.$$

By the special case, we know that there is a $g \in \Gamma(N)$ such that $g\sigma^{-1}\binom{p}{q} = \sigma^{-1}\binom{p'}{q'}$. Then we have $\sigma g \sigma^{-1}\binom{p}{q} = \binom{p'}{q'}$. Since $\Gamma(N)$ is a normal subgroup of $SL_2(\mathbb{Z})$, we see that $\sigma g \sigma^{-1} \in \Gamma(N)$. This finishes the proof. □

Example 1.45. When $N = 2$, it is easy to see that there are exactly 3 such pairs: $\binom{0}{1} = 0, \binom{1}{1} = 1, \binom{1}{0} = \infty$. Hence $X(2)$ has 3 cusps: $0, 1, \infty$.

Recall that we have proved that

Proposition 1.46. *Let $\binom{p}{q}$ denote an element of $\mathbb{Q} \cup \{\infty\}$ with $(p, q) = 1$ in case $q \neq 0$. Then*

$$\Gamma(N)\begin{pmatrix} p \\ q \end{pmatrix} = \Gamma(N)\begin{pmatrix} p' \\ q' \end{pmatrix} \Leftrightarrow \begin{pmatrix} p \\ q \end{pmatrix} \equiv \begin{pmatrix} p' \\ q' \end{pmatrix} \bmod N$$

It follows immediately that, the cusps of $X(N)$ are given by the pairs[2] $\pm\binom{\bar{p}}{\bar{q}}$ in $(\mathbb{Z}/N)^2$ whose order is N. The bijection between those pairs and cusps (*i.e.* orbits under $\Gamma(N)$-action on $\mathbb{Q} \cup \{\infty\}$) is given as follows:

$$\pm\begin{pmatrix} \bar{p} \\ \bar{q} \end{pmatrix} \mapsto \Gamma(N)\begin{pmatrix} p \\ q \end{pmatrix},$$

where $\binom{p}{q}$ is a lift of $\pm\binom{\bar{p}}{\bar{q}}$ in $\mathbb{Z}^2$ with $(p, q) = 1$.

Hence counting those pairs in $(\mathbb{Z}/N)^2$ gives us a formula of the number of cusps for $X(N)$; this approach also gives us explicitly the cusps of $X(N)$.

Example 1.47. (1) $X(2)$: there are clearly 3 cusps given by $0, 1, \infty$.
(2) $X(3)$: there are 4 cusps given by $0, 1, -1, \infty$; note that a tetrahedron has 4 vertices.
(3) $X(4)$: there are 6 cusps; note that an octahedron has 6 vertices.
(4) $X(5)$: there are 12 cusps; note that an icosahedron has 12 vertices.

We will explain later the connection between the number of cusps and the number of vertices in the case of $X(5)$.

[2]When $N \geq 3$, they are indeed pairs; but when $N = 2$, it is clearly $\pm\binom{\bar{p}}{\bar{q}}$ coincide.

Next, I want to take a slightly more conceptual approach to derive a formula of the number of cusps of $X(N)$ since the ingredients are needed later on.

Writing $X(1) = \mathfrak{h}^*/\Gamma(1)$ as $(\mathfrak{h}^*/\Gamma(N))/(\Gamma(1)/\Gamma(N))$, we have a surjection

$$X(N) = \mathfrak{h}^*/\Gamma(N) \twoheadrightarrow^{\pi_N} X(1)$$

given by $(\Gamma(1)/\Gamma(N))x \mapsto x$.

Let y be a point of $X(1)$. We want to understand the cardinality of $\pi_N^{-1}(y)$, *i.e.* the cardinality of $(\Gamma(1)/\Gamma(N))x$. We will start with an ordinary point.

Proposition 1.48. *If y is an ordinary point of $X(1)$, then*

$$|\pi_N^{-1}(y)| = \begin{cases} \frac{1}{2}|SL_2(\mathbb{Z}/N)| & N = 3 \\ |SL_2(\mathbb{Z}/2)| & N = 2 \end{cases}$$

Proof. From our 1st homework, we have an exact sequence

$$1 \to \Gamma(N) \to \Gamma(1) \to SL_2(\mathbb{Z}/N) \to 1.$$

Consequently, $|\Gamma(a)/\Gamma(N)| = |SL_2(\mathbb{Z}/N)|$.

When $N \geq 3$, we see that $-I_2 \notin \Gamma(N)$. This implies that $\Gamma(N)g \neq \Gamma(N)(-g)$ for each coset $\Gamma(N)g \in \Gamma(1)/\Gamma(N)$. But $\Gamma(N)gy = \Gamma(N)(-g)y$ since $gy = (-g)y$ (and no other element of $\Gamma(1)$ can fix y). Hence we have that

$$|\pi_N^{-1}(y)| = \frac{1}{2}|SL_2(\mathbb{Z}/N)|.$$

When $N = 2$, we have $-I_2 \in \Gamma(2)$. Hence we have $|\pi_2^{-1}(x)| = |SL_2(\mathbb{Z}/2)|$. □

From Proposition 1.48, we can see that for all but 3 points (2 elliptic points and one cusp) the cardinality of $\pi_N^{-1}(y)$ are the same; this cardinality is called the degree of π_N and will be denoted by d_N. The same phenomenon happens for more general maps as we will see later on.

Proposition 1.49.

$$|\pi_N^{-1}(i)| = \frac{d_N}{2},\ |\pi_N^{-1}(\rho)| = \frac{d_N}{3},\ |\pi_N^{-1}(\infty)| = \frac{d_N}{N}.$$

Proof. The idea already appears in the proof of Proposition 1.48, I will be brief.

Since $\Gamma(1)_i = \langle S \rangle$ with $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ and $\langle S \rangle / \{\pm I_2\}$ has order 2, we need to divide d_n further by 2.

Since $\Gamma(1)_\rho = \langle ST \rangle$ with $T = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ and $\langle ST \rangle / \{\pm I_2\}$ has order 3, we need to divide d_n further by 3.

For ∞, note that $\Gamma(1)_\infty = \langle \begin{pmatrix} 1 & n \\ 1 & 0 \end{pmatrix} | n \in \mathbb{Z} \rangle$ and $\Gamma(N)_\infty = \langle \begin{pmatrix} 1 & nN \\ 1 & 0 \end{pmatrix} | n \in \mathbb{Z} \rangle$. Hence $\begin{pmatrix} 1 & \ell \\ 1 & 0 \end{pmatrix}$ with $0 \leq \ell \leq n-1$ give different cosets in $\Gamma(1)/\Gamma(N)$, but the same point. Therefore, we need to divide d_n further by N. □

Note that Proposition 1.49 also gives us a formula of the number of cusps.

One may think of Proposition 1.49 as counting points with multiplicities: each point in the preimage of i has multiplicity 2 (or is counted twice), etc. Such a multiplicity can be conceptualized as follows.

Definition 1.50. Let $f : X \to Y$ be a nonconstant map between Riemann surfaces. f is called *holomorphic* is, given any charts $\{(U_\alpha, t_\alpha)\}$ on X and $\{(V_\beta, s_\beta)\}$ on Y, the map $s_\beta \circ f \circ t_\alpha^{-1}$ is holomorphic.

Remark 1.51. Let $f : X \to Y$ be a nonconstant holomorphic map between *compact* Riemann surfaces. Then $|f^{-1}(y)|$ is the same for all but finitely many points $y \in Y$.

Let $x \in f^{-1}(y)$. One may choose local charts (U, t) at x and (V, s) at y such that $t : V \to \mathbb{D}$ and $s : V \to \mathbb{D}$ are homeomorphisms, where $\mathbb{D}$ is the open unit disk in $\mathbb{C}$, and $t(x) = 0 = s(y)$. Since $g = s \circ f \circ t^{-1} : \mathbb{D} \to \mathbb{D}$ is holomorphic, we may write

$$g(z) = z^{e_{x/y}} + \text{higher order terms}$$

where $e_{x/y}$ is a positive integer and it is independent of the charts $(U, t), (V, s)$.

One may think of $e_{x/y}$ as follows. Let's ignore the higher order part and assume that $g(z) = z^{e_{x/y}}$. Then $g|_{\mathbb{D}\setminus\{0\}}$ is a $e_{x/y}$-to-1 cover, *i.e.* for each $y \in \mathbb{D}\setminus\{0\}$, $|g^{-1}(y)| = e_{x/y}$ and each $x \in g^{-1}(y)$ has an open neighborhood U such that $g|_U$ is a homeomorphism.

One may also think of $e_{x/y}$ as the multiplicity of 0 over 0 (the pre image of 0 consists of only one point 0, being counted $e_{x/y}$ times).

Definition 1.52. Let $f : X \to Y$ be a nonconstant holomorphic map between compact Riemann surfaces. The *degree* of f is the integer d such that $|f^{-1}(y)| = d$ for all but finitely many $y \in Y$.

The integer $e_{x/y}$ in Remark 1.51 is called the *ramification index* of x over y. If $e_{x/y} \geq 2$, then y is called a *ramification point*.

Example 1.53. Consider the natural surjection $\pi_N : X(N) \to X(1)$. Using the charts that we constructed during the proof of $X(N)$ being a compact Riemann surface (recall that, we used the function $z \mapsto z^2$ in the construction of a chart at i, $z \mapsto z^3$ in the construction of a chart at ρ, etc), we can see that $e_{x/i} = 2$ for each $x \in \pi_N^{-1}(i)$, $e_{x/\rho} = 3$ for each $x \in \pi_N^{-1}(\rho)$, $e_{x/\infty} = N$ for each $x \in \pi_N^{-1}(\infty)$, which are exactly the denominators in the formulas of $|\pi_n^{-1}(i)|$, $|\pi_n^{-1}(\rho)|$, $|\pi_n^{-1}(\infty)|$, respectively.

Let me remind you that, our goal is to understand $X(N)$ and our main tool is the classification of compact Riemann surfaces via genus. To this end, let's briefly review how genus is defined and its topological/geometric meaning.

Definition 1.54. Let X be a compact Hausdorff topological space. A *curved triangle* on X is a closed subspace A of X with a homeomorphism $t : \tau \to A$ where τ is a closed triangular region in the plane.

A *triangulation* of X is a finite collection $\{A_1, \ldots, A_n\}$ of curved triangles on X such that

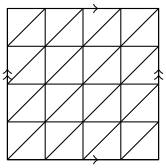
(1) $\bigcup_i A_i = X$, and

(2) for $i \neq j$, the intersection $A_i \cap A_j$ is empty, or a vertex of both A_i and A_j, or an edge of A_i and A_j.

Example 1.55. (1) The Riemann sphere S^2. One may enclose a tetrahedron with S^2 and project the edges and faces of the tetrahedron to S^2 to triangulate S^2 with 4 vertices, 6 edges, and 4 faces.

draw a picture here

(2) The torus $\mathbb{T}^1$. We know that we can get a torus by gluing the opposite sides of a rectangle (or even a parallelogram) without any twisting. A triangulation is given as follows.



We need a few facts from topology.

Fact 1.56. Each compact Riemann surface can be triangulated.

Let v_X, e_X, f_X denote the numbers of vertices, edges, and faces of a triangulation of a compact Riemann surface X.

Fact 1.57. The integer $v_X - e_X + f_X$ is independent of the triangulation on X, called the *Euler characteristic* of X, denoted by $\chi(X)$.

Example 1.58. From the triangulations of S^2 and $\mathbb{T}^1$ we constructed, we have

$$\chi(S^2) = 2,\ \chi(\mathbb{T}^1) = 0.$$

Definition 1.59. Let X be compact Riemann surface. The *genus* of X, denoted by g_X, is defined by

$$g_X = \frac{2 - \chi(X)}{2}.$$

Clearly, knowing $\chi(X)$ is equivalent to knowing g_X. Why bother introducing the notion of genus? A short answer is that the topological/geometric meaning of genus is easier for us to visualize than the one of Euler characteristic. Given any surface X, one perform the following procedure (adding a handle to X): removing two small closed disks from X, opening up a torus, and gluing the two ends to X ,without any twisting, where those 2 disks have been removed. For instance, adding a handle to the Riemann sphere creates a torus. We may say that a sphere has no handle and a torus has one handle. More generally, the genus of a compact Riemann surface is the number handles that it has.

Fact 1.60. Each compact Riemann surface of genus g is homeomorphic to a Riemann sphere with g handles added.

Therefore, to understand (or to visualize) a compact Riemann surface, it suffices to calculate its genus.

Recall that our goal is to understand $X(N)$; we will calculate $g_{X(N)}$. Since we already know that $X(1)$ is the Riemann sphere and have a surjection $X(N) \to X(1)$, naturally the tool we will use is the Riemann-Hurwitz formula.

Theorem 1.61 (Riemann-Hurwitz)**.** *Let $h : X \to Y$ be a non constant holomorphic map between compact Riemann surfaces with degree d. Then*

$$2 - 2g_X = d(2 - 2g_Y) - \sum_{y \in Y} \sum_{x \in h^{-1}(y)} (e_{x/y} - 1).$$

Proof. Triangulate Y and make sure that each ramification point is a vertex. Applying h^{-1}, we get a triangulation of X. From Remark 1.51, we can see that

$$v_X = dv_Y - \sum_{y \in Y} \sum_{x \in h^{-1}(y)} (e_{x/y} - 1),\ e_X = de_Y,\ f_X = df_Y.$$

The formula follows. □

Recall that we proved the Riemann-Hurwitz formula last time.

Theorem 1.62 (Riemann-Hurwitz)**.** *Let $h : X \to Y$ be a non constant holomorphic map between compact Riemann surfaces with degree d. Then*

$$2 - 2g_X = d(2 - 2g_Y) - \sum_{y \in Y} \sum_{x \in h^{-1}(y)} (e_{x/y} - 1).$$

We will apply this formula to calculate the genus of $X(N)$.

Theorem 1.63. *Let d_N be the degree of the natural surjection of $\pi_N : X(N) \to X(1)$. Then*

$$g_{X(N)} = 1 + \frac{d_N}{12} - \frac{d_N}{2N}. \tag{1.63.1}$$

Proof. Plugging in $|\pi_N^{-1}(i)| = \frac{d_N}{2}$, $e_{x/i} = 2$ for each $x \in \pi_N^{-1}(i)$, $|\pi_N^{-1}(\rho)| = \frac{d_N}{3}$, $e_{x/\rho} = 3$ for each $x \in \pi_N^{-1}(i)$, and $|\pi_N^{-1}(\infty)| = \frac{d_N}{N}$, $e_{x/\infty} = N$ for each $x \in \pi_N^{-1}(\infty)$ into the Riemann-Hurwitz and keeping in mind that $g_{X(1)} = 0$, we get

$$2 - 2g_{X(N)} = 2 - \frac{d_N}{2}(2 - 1) - \frac{d_N}{3}(3 - 1) - \frac{d_N}{N}(N - 1).$$

Simplifying this, we get

$$g_{X(N)} = 1 + \frac{d_N}{12} - \frac{d_N}{2N}.$$

□

Remark 1.64. As you may notice that our formula doesn't hold when $N = 1$. The reason is that there are elliptic points (i and ρ) when $N = 1$. Here is the formula that works more generally.

Let Γ be a subgroup of $SL_2(\mathbb{Z})$ that contains $\Gamma(N)$ for some positive integer N. Let $X(\Gamma)$ denote $\mathfrak{h}^*/\Gamma$ and d_Γ denote the degree of the natural projection of $X(\Gamma) \to X(1)$. Let $\varepsilon_2, \varepsilon_3, \varepsilon_\infty$ denote the number of elliptic points of order 2, the number of elliptic points of order 3, and the number of cusps, respectively. Then

$$g(X(\Gamma)) = 1 + \frac{d_\Gamma}{12} - \frac{\varepsilon_2}{4} - \frac{\varepsilon_3}{3} - \frac{\varepsilon_\infty}{2}.$$

Example 1.65. Using formula (1.63.1), we can see that

$$g_{X(2)} = g_{X(3)} = g_{X(4)} = g_{X(5)} = 0.$$

In particular, $X(5)$ is the Riemann sphere.

This explains the following diagram (that appeared in Mike's lecture):

$$\begin{array}{ccccc} S^2 & \hookleftarrow & \mathfrak{h}/\Gamma(5) & \leftarrow & \mathfrak{h} \\ \downarrow & & \downarrow & \swarrow & \\ S^2/A_5 & \hookleftarrow & \mathfrak{h}/\Gamma(1) & & \end{array}$$

In this diagram, the inclusion on the top is $\mathfrak{h}/\Gamma(5) =: Y(5) \hookrightarrow X(5) = S^2$. Since $X(5)$ has 12 cusps, this inclusion has 12 points missing on the target. Similarly, the inclusion on the bottom is $\mathfrak{h}/\Gamma(1) =: Y(1) \hookrightarrow X(1) = S^2$, taking into account that $PSL_2(\mathbb{Z}/5) = A_5$; this inclusion has one point missing.

Remark 1.66. One can construct an group iso between $PSL_2(\mathbb{Z}/5)$ and A_5 explicitly. Or, more conceptually, each $PSL_2(\mathbb{Z}/p)$ with $p > 3$ is a simple group. Since $PSL_2(\mathbb{Z}/5)$ and A_5 are both simple groups of order 60, there will be an isomorphism.

The A_5-action on S^2 can be visualized as follows. Inside the icosahedron, there are 5 tetrahedrons; each symmetry on the icosahedron will permute these 5 tetrahedrons and it is a even permutation. This shows that the symmetry group on the tetrahedron is A_5. Now enclose the icosahedron with S^2 and extend the A_5-action to S^2.

It turns out that A_5 also acts on the meromorphic functions on S^2.

Definition 1.67. Let $f : U \to \mathbb{C}$ be function from an open subset U of $\mathbb{C}$ to $\mathbb{C}$. A point $z_0 \in U$ is called a *pole* of f if $f(z_0)$ is undefined and there is a positive integer m such that $(z-z_0)^m f(z)$ is holomorphic around z_0. (Recall that we have defined holomorphic functions.) f is called *meromorphic* if f is holomorphic on U except a discrete subset of U each of which is a pole of f.

Definition 1.68. Let $f : X \to Y$ be a function between Riemann surfaces. f is called *meormorphic* if, given any charts $\{(U_\alpha, t_\alpha)\}$ on X and $\{(V_\alpha, s_\alpha)\}$ on Y, the function $s_\beta \circ f \circ t_\alpha^{-1}$ is meromorphic.

Theorem 1.69. *The meromorphic functions $S^2 \to \mathbb{C}$ are just rational functions on S^2 in one variable z. In particular, they form a field $\mathbb{C}(z)$ of transcendental degree 1 over $\mathbb{C}$.*

Proof. Since all the meromorphic functions on S^2 form a field, the second conclusion follows from the first.

We will think of S^2 as $\mathbb{C} \cup \{\infty\}$. Clearly each rational function (*i.e.* $\frac{p(z)}{q(z)}$ where both p and q are polynomials) is a meromorphic function on $\mathbb{C} \cup \{\infty\}$. it remains to show the reverse inclusion.

Let $f : \mathbb{C} \cup \{\infty\} \to \mathbb{C}$ be a meromorphic functions. Since $\mathbb{C} \cup \{\infty\}$ is compact, f has finitely many zeros and poles. Let $a_1, \ldots, a_n$ be the zeros of f with $\text{ord}_{a_i} = \alpha_i$ and $b_1, \ldots, b_m$ be the poles of f with $\text{ord}_{b_i}(f) = -\beta_i$. Consider the rational function $r(z)$ defined by

$$r(z) := \frac{\prod_{i=1}^n (z - a_i)^{\alpha_i}}{\prod_{i=1}^m (z - b_i)^{\beta_i}}$$

Clearly $1/r$ is also a meromorphic functions on $\mathbb{C} \cup \{\infty\}$. Set $h = \frac{f}{r}$. Then h is holomorphic on $\mathbb{C}$ with ∞ being the only possible pole. The order of h at ∞ must be finite. So one write $h = \sum_{j=0}^M c_i z^i$ where M is an integer. If $M \geq 1$, then by the fundamental theorem of

algebra, h will have at least M zeros; but h has no zeros by construction. So h is a constant. This shows that f is the product of a rational function and a constant; hence finishes the proof of the reverse inclusion. □

The A_5-action, *i.e.* the map $S^2 \to S^2/A_5$, induces an field extension $\mathbb{C}(z) \hookrightarrow \mathbb{C}(z')$ with Galois group A_5, where $\mathbb{C}(z)$ is the function field of S^2/A_5 (which we have seen still is S^2) and $\mathbb{C}(z')$ is the function field of S^2. This is a degree 60 extension. The icosahedron equation is the degree-60 equation over $\mathbb{C}(z)$ whose roots generate $\mathbb{C}(z')$, which plays a crucial role in Klein's solution to quintics.

Definition 1.70. Assume E/F is a finite, separable extension of fields and let $\sigma_1, \ldots, \sigma_d$ be the distinct embeddings of E into any chosen algebraic closure $\overline{F}$ of F. ($d = [E : F]$.) For $\alpha \in E$, the *norm* of α over F is

$$N_F^E(\alpha) := \prod_{i=1}^{d} \sigma_i(\alpha).$$

Proposition 1.71. *For any finite, separable extension E/F and any $\alpha \in E$, define $f_\alpha(x) = \prod_{i=1}^{d}(x - \sigma_i(\alpha))$, where $\sigma_1, \ldots, \sigma_d : E \to \overline{F}$ are all the embeddings of E into a chosen algebraic closure $\overline{F}$ of F that fix F. Then all of the coefficients of $f_\alpha(x)$ belong to F and so, in particular, $N_F^E(\alpha) \in F$.*

Proof. First, we will prove this proposition in the case when E/F is Galois. Then $f_\alpha(x) = \prod_{\sigma \in \mathrm{Gal}(E/F)}(x - \sigma(\alpha))$. For any $\tau \in \mathrm{Gal}(E/F)$, we have $f_\alpha^\tau(x) = \prod_{\sigma \in \mathrm{Gal}(E/F)}(x - \sigma(\alpha))^\tau = \prod_{\sigma \in \mathrm{Gal}(E/F)}(x - \tau\sigma(\alpha))$. Since the set $\{\tau\sigma \,|\, \sigma \in \mathrm{Gal}(E/F)\}$ coincides with $\mathrm{Gal}(E/F)$, we get $f_\alpha^\tau(x) = f_\alpha(x)$. This proves all of the coefficients of $f_\alpha(x)$ are fixed by every element of $\mathrm{Gal}(E/F)$ and hence must all belong to F.

When E/F is not Galois, we consider the normal closure E^{nor} of E over F in an algebraic closure $\overline{F}$ of F. Each σ_j can be extended to $\widetilde{\sigma}_j; E^{nor} \hookrightarrow \overline{F}$. Since E^{nor} is normal, we have $\widetilde{\sigma}_j(E^{nor}) = E^{nor}$. Hence we may compose $\tau \in Gal(E^{nor}/F)$ with $\widetilde{\sigma}_j$. Clearly $\tau \circ \widetilde{\sigma}_j \,|_E = \tau \circ \sigma_j$. Since

$$\{\tau \circ \sigma_1, \ldots, \tau \circ \sigma_d\} = \{\sigma_1, \ldots, \sigma_d\}$$

we have $f_\alpha^\tau(x) = f_\alpha(x)$ and hence $f_\alpha(x) \in F[x]$. □

Theorem 1.72 (Hilbert's Theorem 90)**.** *Suppose E/F is a finite Galois extension and its Galois group is cyclic, say generated by $\sigma \in \mathrm{Gal}(E/F)$. For $\beta \in E$, $N_F^E(\beta) = 1$ if and only if $\beta = \frac{\alpha}{\sigma(\alpha)}$ for some $\alpha \in E$.*

Proof. Assume $\beta = \frac{\alpha}{\sigma(\alpha)}$. Then

$$N_F^E(\beta) = \prod_{i=0}^{n-1} \sigma^i(\beta) = \prod_{i=0}^{n-1} \sigma^i\left(\frac{\alpha}{\sigma(\alpha)}\right) = \prod_{i=0}^{n-1} \frac{\sigma^i(\alpha)}{\sigma^{i+1}(\alpha)} = 1,$$

since $\sigma^n = 1$.

Now assume $N_F^E(\beta) = 1$. We define a function $g : E \to E$ as the following E-linear combination of $\mathrm{id}, \sigma, \ldots, \sigma^{n-1}$:

$$g := \mathrm{id} + \beta\sigma + (\beta\sigma(\beta))\sigma^2 + \cdots + (\beta\sigma(\beta)\cdots\sigma^{n-2}(\beta))\sigma^{n-1} : E \to E.$$

There exists $u \in E$ such that $g(u) \neq 0$. Set $\alpha = g(u)$. Then

$$\begin{aligned}
\beta\sigma(\alpha) &= \beta\sigma(g(u)) \\
&= \beta\sigma(u + \beta\sigma(u) + (\beta\sigma(\beta))\sigma^2(u) + \cdots + (\beta\sigma(\beta)\cdots\sigma^{n-2}(\beta))\sigma^{n-1}(u)) \\
&= \beta\sigma(u) + \beta\sigma(\beta)\sigma^2(u) + \beta\sigma(\beta)\sigma^2(\beta)\sigma^3(u) + \cdots + \underbrace{(\beta\sigma(\beta)\cdots\sigma^{n-1}(\beta))}_{=N(\beta)=1}\underbrace{\sigma^n(u)}_{=u} \\
&= g(u) = \alpha.
\end{aligned}$$

Thus $\beta = \frac{\alpha}{\sigma(\alpha)}$. □

Theorem 1.73 (Kummer)**.** *Let F be a field containing a primitive nth root of unity. Let E/F be a Galois extension with cyclic Galois group $\mathbb{Z}/n\mathbb{Z}$ (with a generator σ). Then there is an element $\theta \in E$ such that $E = F(\theta)$ and $\theta^n \in F$.*

Proof. Let ω be a primitive nth root of unity in F. Then clearly $N_F^E(\omega) = \omega^n = 1$. By Hilbert Theorem 90, there is $\theta \in E$ such that $\omega = \frac{\sigma(\theta)}{\theta}$. Then $\sigma(\theta) = \omega\theta$ and consequently $\sigma^j(\theta) = \omega^j\theta$. Then $\{g(\theta)|g \in \mathrm{Gal}(E/F)\}$ has n distinct elements; hence $\deg_F(\theta) \geq n = [E : F]$. Therefore $E = F(\theta)$. □

Kummer's theorem tells us a strategy to solve equations. If $f \in k[x]$ has degree n and if the Galois group of f is cyclic of order n, then one can solve $f(x) = 0$ by solving $x^n - a = 0$ which can be solved by finding all numbers β such that $e^\beta = a$ then setting $x = e^{\beta/n}$.

Example 1.74 (Cubic equations)**.** Consider $f(x) = x^3+px+q$ over $K = \mathbb{Q}(\omega)$ with $\omega = e^{\frac{2\pi i}{3}}$. Assume that f has 3 distinct roots x_1, x_2, x_3. Consider the $C_3 = \mathbb{Z}/3 = \langle\sigma\rangle$-action:

$$\sigma : x_1 \mapsto x_2 \mapsto x_3 \mapsto x_1$$

The expression $(x_1 - x_2)(x_2 - x_3)(x_1 - x_3)$ is *not* invariant under C_3, but its square is. Its square is

$$-4p^3 - 27q^2.$$

By adding the square root of it to K, we may assume that the Galois group of f is C_3. The proof of Hilbert Theorem 90 tells us that we should look at

$$\begin{aligned}
(id + \omega\sigma + \omega\sigma(\omega)\sigma^2)(x_1) &= x_1 + \omega x_2 + \omega^2 x_3, \text{ or} \\
(id + \omega\sigma^2 + \omega\sigma^2(\omega)\sigma^4)(x_1) &= x_1 + \omega x_3 + \omega^2 x_2, \ \sigma^2 \text{ is also a generator}
\end{aligned}$$

After normalizing, set

$$\begin{aligned}
\alpha_1 &= (x_1 + \omega x_2 + \omega^2 x_3)/3 \\
\alpha_2 &= (x_1 + \omega x_3 + \omega^2 x_2)/3
\end{aligned}$$

It can be checked that $\alpha_1\alpha_2 = -\frac{p}{3}$. Writing $t = \alpha_1^3$ leads to the substitution

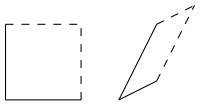
$$x = t^{1/3} - (p/3)t^{-1/3},$$

the Tartaglia substitution. This substitution turns f into a quadratic equation.

2. Elliptic curves

Definition 2.1. An *elliptic curve* E over $\mathbb{C}$ is a compact Riemann surface of genus 1.

By the classification of compact Riemann surfaces, we know that an elliptic curve E must be a complex torus.



To compare tori, we consider lattices.

Definition 2.2. A *lattice* Λ in $\mathbb{C}$ is a free abelian group $\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$ of rank 2 such that $\mathbb{R}\omega_1 + \mathbb{R}\omega_2 = \mathbb{C}$.

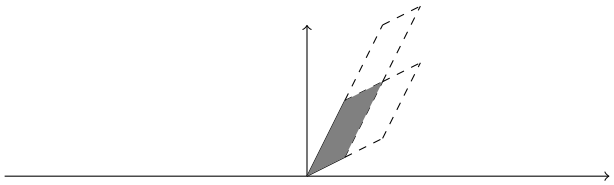
Remark 2.3. Since ω_1, ω_2 generate $\mathbb{C}$ over $\mathbb{R}$, the ratio ω_1/ω_2 can not be a real number. Since -1 is a unit in $\mathbb{Z}$, we may change one of ω_1, ω_2 to its negative without changing the lattice. Hence we may assume that the imaginary part of ω_1/ω_2 is positive, *i.e.* $\omega_1/\omega_2 \in \mathfrak{h}$.

Any complex torus has the form $\mathbb{C}/\Lambda$ for a lattice Λ in $\mathbb{C}$.

Definition 2.4. Let $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$ be a lattice. The *fundamental parallelogram* of Λ is

$$\{a_1\omega_1 + a_2\omega_2 \mid 0 \leq a_1, a_2 < 1\},$$

denoted by $\mathcal{P}$.



Definition 2.5. An *elliptic function* (relative to Λ) is a meromorphic function $f : \mathbb{C} \to \mathbb{C}$ such that

$$f(z + \omega) = f(z)$$

for all $\omega \in \Lambda$ and $z \in \mathbb{C}$.

It is clear that each elliptic function induces a function on $\mathbb{C}/\Lambda$ and all elliptic functions (relative to a fixed lattice Λ) form a field.

Recall the following result from complex analysis.

Theorem 2.6 (Liouville)**.** *A bounded holomorphic function $f : \mathbb{C} \to \mathbb{C}$ must be a constant.*

Proposition 2.7. *An elliptic function f (relative to Λ) without poles (or zeros) is constant.*

Proof. No poles implies that f is holomorphic. The periodicity of f implies that

$$\sup_{z \in \mathbb{C}} \mid f(z) \mid = \sup_{z \in \bar{\mathcal{P}}} \mid f(z) \mid .$$

since f is continuous and $\bar{\mathcal{P}}$ is compact, f is bounded. Hence it must be a constant by Liouville's Theorem. □

Theorem 2.8. *Given any two lattices Λ_1, Λ_2. The function*

$$\{\alpha \in \mathbb{C} \mid \alpha\Lambda \subseteq \lambda_2\} \to \{\text{holomorphic } \phi : \mathbb{C}/\Lambda_1 \to \mathbb{C}/\Lambda_2 \mid \phi(0) = 0\}$$

defined by $\alpha \mapsto \phi_\alpha$ is a bijection, where $\phi_\alpha(z) = \alpha z \pmod{\Lambda_2}$.

Proof. Injectivity: assume that $\phi_\alpha = \phi_\beta$. Then $\alpha z \equiv \beta z \pmod{\Lambda_2}$. Hence the multiplication by $\alpha - \beta$ sends $\mathbb{C}$ into Λ_2. Since $\mathbb{C}$ is connected and Λ_2 is discrete, this multiplication map must a constant; the only possibility is $\alpha - \beta = 0$. This proves injectivity.

Surjectivity: assume that $\phi : \mathbb{C}/\Lambda_1 \to \mathbb{C}/\Lambda_2$ is holomorphic and $\phi(0) = 0$. Since $\mathbb{C}$ is the universal covering space of any torus, we can lift ϕ to a holomorphic $\widetilde{\phi} : \mathbb{C} \to \mathbb{C}$ such that $\widetilde{\phi}(0) = 0$.

$$\begin{array}{ccc} \mathbb{C} & \xrightarrow{\widetilde{\phi}} & \mathbb{C} \\ \downarrow & & \downarrow \\ \mathbb{C}/\Lambda_1 & \xrightarrow{\phi} & \mathbb{C}/\Lambda_2 \end{array}$$

Then we have $\widetilde{\phi}(z+\omega) \equiv \widetilde{\phi}(z) \pmod{\Lambda_2}$ for all $\omega \in \Lambda_1$. Since Λ_2 is discrete, we have $\widetilde{\phi}(z+\omega) - \widetilde{\phi}(z)$ is a constant independent of z. Hence

$$\widetilde{\phi}'(z+\omega) = \widetilde{\phi}'(z),\ \forall \omega \in \Lambda_1,\ z \in \mathbb{C}.$$

So, $\widetilde{\phi}'$ is a holomorphic elliptic function; it must be a constant by Proposition 2.7, say α. Then $\phi(z) = \alpha z + \beta$. But $\beta = 0$ since $\phi(0) = 0$. □

Corollary 2.9. *A holomorphic function $\phi_\alpha : \mathbb{C}/\Lambda_1 \to \mathbb{C}/\Lambda_2$ is bijective if and only if $\alpha\Lambda_1 = \Lambda_2$.*

Proof. Exercise. □

Definition 2.10. Two lattices are called *homothetic* if there is $\alpha \in \mathbb{C}$ such that $\alpha\Lambda_1 = \Lambda_2$.

So, isomorphic tori (meaning there is a bijective holomorphic maps between) correspond to homothetic lattices.

Remark 2.11. A lattice generated by ω_1, ω_2 is homothetic to the lattice generated by $(\omega_1/\omega_2, 1)$ (keep in mind that $\omega_1/\omega_2 \in \mathfrak{h}$). Therefore, each lattice has the form $\mathbb{Z}\tau + \mathbb{Z}$ with $\tau \in \mathfrak{h}$.

We will use E_τ t denote the complex torus $\mathbb{C}/\Lambda$ where Λ is generated by τ and 1.

Proposition 2.12. *Given $\tau, \tau' \in \mathfrak{h}$. The complex tori $E_\tau, E_{\tau'}$ are isomorphic if and only if there is a $\sigma \in SL_2(\mathbb{Z})$ such that $\tau = \sigma(\tau')$.*

Proof. Assume that $E_\tau \cong E_{\tau'}$, then there is an $\alpha \in \mathbb{C}$ such that $\alpha\Lambda = \Lambda'$. Therefore, we may write

$$\alpha\tau = a\tau' + b \text{ and } \alpha 1 = c\tau' + d$$

where $a, b, c, d \in \mathbb{Z}$. Since $a\tau' + b$ and $c\tau' + d$ also generate λ', we must have that the determinant of $\sigma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is either 1 or -1. Since $\tau, \tau' \in \mathfrak{h}$ and $\sigma(\tau') = \tau$, we must have $\sigma \in SL_2(\mathbb{Z})$.

Conversely, assume that $\tau, \tau' \in \mathfrak{h}$ and $\sigma(\tau') = \tau$ for some $\sigma \in SL_2(\mathbb{Z})$. Then we have $\tau = \frac{a\tau'+b}{c\tau'+d}$. Set $\alpha = c\tau' + d$, then $\alpha\tau = a\tau' + b$. Since $\det(\sigma) = 1$, the scalar α defines an isomorphism from Λ_τ to $\Lambda_{\tau'}$. Hence $E_\Lambda \cong E_{\Lambda'}$. □

Therefore, the set of isomorphism classes of elliptic curves over $\mathbb{C}$ is in natural bijection with $\mathfrak{h}/SL_2(Z) = Y(1)$. And the bijection is given by

$$\tau \in \mathfrak{h} \leftrightarrow E_\tau$$

Hence $Y(1)$ parametrizes elliptic curves.

Clearly $E_\tau = \mathbb{C}/\Lambda_\tau$ is an abelian group under $+$. Given each integer $N \geq 2$, we may consider the N-torsion points, *e.g.* $\frac{1}{N}, \frac{\tau}{N}$.

Exercise 2.13. Prove that the N-torsion points in E_τ form an abelian group isomorphic to $(\mathbb{Z}/N)^2$ with a basis given by $\frac{1}{N}, \frac{\tau}{N}$.

Therefore, the set of isomorphism classes of the triples (E, t_1, t_2) where E is an elliptic curve over $\mathbb{C}$ and t_1, t_2 is a basis for N-torsion points in E is in natural bijection with $\mathfrak{h}/\Gamma(N) = Y(N)$. And the bijection is given by

$$\tau \in \mathfrak{h} \leftrightarrow (E_\tau, \frac{1}{N}, \frac{\tau}{N})$$

Next, we will discuss Weierstrass $\wp$-function (associated with a lattice Λ) which turns out to be the key to understand elliptic functions on Λ:

$$\wp(z) := \frac{1}{z^2} + \sum_{\omega \in \Lambda - \{0\}} (\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2})$$

To show that $\wp$ is a meromorphic function, we need the following theorem.

Theorem 2.14 (Weierstrass' M-test)**.** *Let $\{f_n : W \to \mathbb{C}\}_{n \geq 1}$ be a sequence of holomorphic function on an open subset W of $\mathbb{C}$. Assume that there is a sequence of positive real numbers $\{M_n\}_{n \geq 1}$ such that $\sum_n M_n$ converges and $\mid f_n(z) \mid \leq M_n$ for all $z \in W$ for each n. Then $\sum_n f_n(z)$ converges uniformly absolutely to a holomorphic function $f(z)$ and*

$$f'(z) = \sum_n f_n'(z).$$

Theorem 2.15. *$\wp$ is a meromorphic function with poles precisely at lattices each of which is of order 2, and*

$$\wp'(z) = -2 \sum_{\omega \in \Lambda} \frac{1}{(z-\omega)^3}.$$

We need an easy lemma first.

Lemma 2.16. *Write $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$. Then there is a positive real number ϵ such that*

$$\mid a_1\omega_1 + a_2\omega_2 \mid \geq \epsilon\sqrt{a_1^2 + a_2^2}$$

for all real numbers a_1, a_2.

Proof. Dividing both sides of the desired inequality by $\sqrt{a_1^2 + a_2^2}$, we see that it is equivalent to proving there is an $\epsilon > 0$ such that

$$\mid (\cos\theta)\omega_1 + (\sin\theta)\omega_2 \mid \geq \epsilon$$

for all $\theta \in [0, 2\pi]$. Since the function $\mid (\cos\theta)\omega_1 + (\sin\theta)\omega_2 \mid$ is a continuous function on a compact set $[0, 2\pi]$, it is bounded. Since ω_1, ω_2 are $\mathbb{R}$-linearly independent, the lower bound of $\mid (\cos\theta)\omega_1 + (\sin\theta)\omega_2 \mid$ must be positive. Set ϵ to be this lower bound. $\square$

Proof of Theorem 2.15. It suffices to show that given any $M > 0$ after removing finitely many (lattice) points the series in $\wp$ converges uniformly absolutely in open disc $\mathbb{D}(0, M) = \{z \mid |z| < M\}$. To this end, set $\Lambda_M = \{\omega \in \Lambda \mid |\omega| \leq 2M\}$. Then from Lemma 2.16, we have

$$\Lambda_M \subseteq \{n\omega_1 + m\omega_2 \mid n^2 + m^2 \leq 4M\epsilon^{-2}\}.$$

For each $z \in \mathbb{D}(o, M)$ and $\omega \in \Lambda - \Lambda_M$, we have $|z| < \frac{1}{2}|\omega|$. Then

$$|\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2}| = |\frac{z(2\omega - z)}{(z-\omega)^2\omega^2}| \leq \frac{5M\frac{|\omega|}{2}}{|\omega|^4/4} = \frac{10M}{|\omega|^3} \leq \frac{10M\epsilon^{-3}}{(n^+m^2)^{\frac{3}{2}}}.$$

By Weierstrass' M-test, $\wp(z)$ is holomorphic on the open set $\mathbb{D}(0, M) - \Lambda_M$ for each $M > 0$. Hence $\wp(z)$ is meromorphic. It also follows from Weierstrass' M-test that taking derivative of $\wp(z)$ is the same as taking derivative of each term in the series. □

Proposition 2.17. *$\wp(z)$ is an even elliptic function.*

Proof. First, $\wp(z)$ is an even function since replacing ω with $-\omega$ doesn't affect the sum (as ω runs over Λ, $-\omega$ also runs over Λ).

For each $\lambda \in \Lambda$, we have $\wp'(z+\lambda) = -2\sum_{\omega\in\Lambda}(z+\lambda-\omega)^{-3}$. When ω runs over Λ, so does $\omega - \lambda$, therefore $\wp'(z+\lambda) = \wp'(z)$ for all z. Consequently, $\wp(z+\lambda) - \wp(z) = c(\lambda)$ where $c(\lambda)$ is a constant (independent of z). Setting $z = -\frac{\lambda}{2}$ gives us

$$c(\lambda) = \wp(\frac{\lambda}{2}) - \wp(-\frac{\lambda}{2}) = 0$$

where the last equality follows from the fact that $\wp(z)$ is an even function. □

Theorem 2.18. *Let f be an elliptic function relative to a lattice $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$. Then*

$$\sum_{z\in\mathbb{C}/\Lambda} \mathrm{ord}_z(f) = 0.$$

Proof. By the Residue Theorem in complex analysis, we have

$$\sum_{z\in\mathbb{C}/\Lambda} \mathrm{ord}_z(f) = \frac{1}{2\pi i}\int_{\partial\mathcal{P}} \frac{f'(z)}{f(z)}dz,$$

where $\mathcal{P} = \{a_1 + \omega_1 + a_2\omega_2 \mid 0 \leq a_1, a_2 < 1\}$. The boundary $\partial\mathcal{P}$ of $\mathcal{P}$ is a parallelogram oriented counterclockwise; opposite sides have opposite direct. Since both f' and f are periodic with respect to Λ, this integral is 0. □

Remark 2.19. As a consequence of Theorem 2.18, the number of poles of an elliptic function f in the fundamental parallelogram $\mathcal{P}$ is the same as the number of zeros in $\mathcal{P}$, called the order of $f(z)$.

The Weierstrass $\wp$-function has order 2.

Definition 2.20. Let Λ be a lattice and $k > 2$ be an integer. The *Eisenstein series of weight k* for Λ is

$$G_k(\Lambda) := \sum_{\omega\in\Lambda-\{0\}} \frac{1}{\omega^k}.$$

Before we show that this series converges absolutely (and hence is well-defined), we want to make the following observation.

Remark 2.21. As we have seen, we can write any lattice Λ as $\mathbb{Z}\tau + \mathbb{Z}$ for some $\tau \in \mathfrak{h}$. Hence we can write

$$G_k(\Lambda) = G_k(\tau) = \sum_{m,n\in\mathbb{Z}} \frac{1}{(m+n\tau)^k}.$$

It is straightforward to check that

$$G_k(\tau+1) = G_k(\tau),\ G_k(-\frac{1}{\tau}) = \tau^k G_k(\tau).$$

It is the simplest example of a *modular form* which we will discuss in the sequel.

When k is odd, the terms $\frac{1}{\omega^k}$ and $\frac{1}{(-\omega)^k}$ will cancel; consequently $G_k(\Lambda) = 0$ for all odd k.

Proposition 2.22. *For any lattice Λ, the Eisenstein series of weight $k > 2$ converges absolutely.*

Proof. The idea is to compare this series with $\sum_{n=1}^{\infty} \frac{1}{n^{k-1}}$.

Let δ be the minimal distance between any two points in Λ. Consider the annulus A_n with inner radius n and outer radius $n+1$. One can check that there are at most $\frac{8n\pi}{\delta^2}$ lattice points inside A_n. Therefore,

$$\sum_{\omega\in\Lambda,|\omega|\geq 1} \frac{1}{|\omega|^k} \leq \sum_{n=1}^{\infty} \frac{(8\pi/\delta^2)n}{n^k} = \frac{8\pi}{\delta^2} \sum_{n=1}^{\infty} \frac{1}{n^{k-1}}$$

which converges since $k > 2$. Clearly $\sum_{\omega\in\Lambda,|\omega|<1}$ is a finite sum and hence bounded. So

$$\sum_{\omega\in\Lambda-\{0\}} \frac{1}{|\omega|^k} = \sum_{\omega\in\Lambda,|\omega|\geq 1} \frac{1}{|\omega|^k} + \sum_{\omega\in\Lambda,|\omega|<1} \frac{1}{|\omega|^k} < \infty.$$

This shows that $G_k(\Lambda)$ converges absolutely. □

Theorem 2.23. *The Laurent expansion of $\wp(z)$ (relative to a lattice Λ) at $z = 0$ is*

$$\wp(z) = \frac{1}{z^2} + \sum_{n=1}^{\infty} (2n+1) G_{2n+2}(\Lambda) z^{2n}.$$

Proof. First, we consider the expansion of $\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2}$ around 0. We may assume that $|z| < |\omega|$. Then

$$\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} = \frac{1}{\omega^2}(\frac{1}{(1-\frac{z}{\omega})^2} - 1) = \frac{1}{\omega^2} \sum_{n=1}^{\infty} (n+1)(\frac{z}{\omega})^n = \sum_{n=1}^{\infty} \frac{(n+1)z^n}{\omega^{n+2}},$$

where we use the well-known expansion

$$\frac{1}{(1-x)^2} = (1 + x + \cdots)^2 = \sum_{n=1}^{\infty} (n+1)x^n.$$

This shows that

$$\wp(z) = \frac{1}{z^2} + \sum_{\omega\in\Lambda-\{0\}} \sum_{n=1}^{\infty} \frac{(n+1)z^n}{\omega^{n+2}}.$$

Since we have shown that $\wp(z)$ converges absolutely uniformly, we can change the order of summation. Hence

$$\wp(z) = \frac{1}{z^2} + \sum_{n=1}^{\infty} \sum_{\omega \in \Lambda - \{0\}} \frac{(n+1)z^n}{\omega^{n+2}} = \frac{1}{z^2} + \sum_{n=1}^{\infty} (n+1) G_{n+2}(\Lambda) z^n = \frac{1}{z^2} + \sum_{n=1}^{\infty} (2n+1) G_{2n+2}(\Lambda) z^{2n},$$

where in the last step we use the fact that $G_k(\Lambda) = 0$ for all odd k. □

Remark 2.24. It can be shown that every elliptic function relative to a lattice λ is a rational function of $\wp(z)$ and $\wp'(z)$, *i.e.*

$$\mathbb{C}(\Lambda) = \mathbb{C}(\wp(z), \wp'(z)).$$

For dimension reason, $\mathbb{C}(\lambda)$ has transcendental degree 1 over $\mathbb{C}$. So, there must be an algebraic relation between $\wp(z)$ and $\wp'(z)$, which is given below.

Theorem 2.25. *Let Λ be a lattice and $\wp(z)$ be the Weierstrass function relative to it. Then*

$$\wp'(z)^2 = 4\wp(z)^3 - g_2(\Lambda)\wp(z) - g_3(\Lambda),$$

where $g_2(\Lambda) = 60G_4(\Lambda)$ and $g_3(\Lambda) = 140G_6(\Lambda)$.

Proof. Let's write the first few terms in the Laurent expansions

$$\begin{aligned} \wp'(z)^2 &= 4z^{-6} - 24G_4 z^{-2} - 80G_6 + \cdots \\ \wp(z)^3 &= z^{-6} + 9G_4 z^{-2} + 15G_6 + \cdots \\ \wp(z) &= z^{-2} + 3G_4 z^2 + \cdots \end{aligned}$$

Comparing these, we can see that the function

$$\ell(z) := \wp'(z) - 4\wp(z)^3 + 60G_4\wp(z) - 140G_6$$

is holomorphic near 0 and vanishes at 0. Since both $\wp(z)$ and $\wp'(z)$ are elliptic and the Eisenstein series converge absolutely, $\ell(z)$ is a holomorphic elliptic function; it must be 0. □

Remark 2.26. It is straightforward to check

$$\wp''(z) = 6\wp(z)^2 - \frac{g_2}{2}.$$

Hence the Laurent expansion of $\wp(z)$ can be written as $\wp(z) = \frac{1}{z^2} + \sum_{n=1}^{\infty} a_n z^{2n}$ where the coefficients a_n satisfy

$$a_1 = \frac{g_2}{20},\ a_2 = \frac{g_3}{28},\ a_{n+1} = \frac{6}{(2n+1)(2n+2) - 12} \sum_{j=1}^{n-1} a_j a_{n-j}\ (n \geq 2).$$

Therefore, $\wp(z)$ is uniquely determined by g_2, g_3.

Next, we want to use the relation between $\wp(z)$ and $\wp'(z)$ in Theorem 2.25 to identify the elliptic curve (or complex torus) E_Λ with a nonsingular algebraic curve.

Definition 2.27. Let $f(x, y) \in \mathbb{C}[x, y]$ be a polynomial. A point $(a, b) \in \mathbb{C}^2$ is called a *singular point* of $f(x, y)$ if $f(a, b) = 0$ and $\frac{\partial f(x,y)}{\partial x}|_{(a,b)} = \frac{\partial f(x,y)}{\partial y}|_{(a,b)} = 0$.

The curve defined by f, $\{(a, b) \in \mathbb{C} \mid f(a, b) = 0\}$, is called *nonsingular* (or *smooth*) if $f(x, y)$ has no singular point.

Example 2.28. The curve defined by $y^2 - x$ is nonsingular.

The curve defined by $y^2 - x^3$ has a singular point at $(0, 0)$.

Exercise 2.29. Let $f(x) \in \mathbb{C}[x]$ be a degree 3 polynomial. Prove that the curve defined by $y^2 - f(x)$ is nonsingular if and only if $f(x)$ has 3 distinct roots.

Lemma 2.30. *A point $z \notin \Lambda$ is a zero of $\wp'(z)$ (relative to Λ) if and only if $2z \in \Lambda$.*

Proof. Exercise. □

Theorem 2.31. *Let $\Lambda, g_2(\Lambda), g_3(\Lambda)$ be as in Theorem 2.25. Then the curve defined by $y^2 - 4x^3 - g_2(\lambda)x - g_3(\Lambda)$ is nonsingular.*

Let C denote the curve defined by $y^2 - 4x^3 - g_2(\lambda)x - g_3(\Lambda)$. Then the following map is a bijection:

$$\phi : E_\Lambda = C/\Lambda \xrightarrow{z \mapsto (\wp(z), \wp'(z))} C.$$

Proof. Let ω_1, ω_2 be a basis of Λ. Set $\omega_3 = \omega_1 + \omega_2$. By lemma 2.30, $\wp'(\frac{\omega_i}{2}) = 0$. This shows that each $\frac{\omega_i}{2}$ is a root of $4x^3 - g_2(\Lambda)x - g_3(\Lambda)$. It remains to show that they are distinct. The function $\wp(z) - \wp(\omega_i/2)$ is an even function, hence ha at least a double zero at $\omega_i/2$. But it has order 2, $\omega_i/2$ is the only zero in the fundamental parallelogram. So, $\wp(\omega_i/2) \neq \wp(\omega_j/2)$ for $i \neq j$.

Assume that (a, b) is a point on C. Then $\wp(z) - a$ is a nonconstant elliptic function, hence (again by the Residue Theorem) it must have a zero z_0. hence $\wp'(z_0)^2 = b^2$. Since $\wp'(z)$ is an odd function, we may replace z_0 by $-z_0$ if necessary and assume that $\wp'(z_0) = b$. Then $z_0 \mapsto (a, b)$. This proves surjectivity of ϕ.

Assume that $\phi(z_1) = \phi(z_2)$. First assume that $2z_1 \notin \Lambda$. Then function $\wp(z) - \wp(z_1)$ has order 2 and zeros $z_1, -z_1, z_2$. Hence $z_2 \equiv \pm z_1 \pmod{\Lambda}$. Consider

$$\wp'(z_1) = \wp'(z_2) = \wp'(\omega z_1) = \pm \wp'(z_1)$$

and $\wp'(z_1) \neq 0$ (since $2z_1 \notin \Lambda$), we must have $z_2 \equiv z_1 \pmod{\Lambda}$. Next assume that $2z_1 \in \Lambda$. Then $\wp(z) - \wp(z_1)$ has a double zero at z_1 and vanishes at z_2; so $z_2 \equiv z_1 \pmod{\Lambda}$. □

Definition 2.32. The *discriminant* of Λ (or of the curve $y^2 = 4x^3 - g_2(\Lambda)x - g_3(\Lambda)$) is defined as

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

The *j-invariant* of Λ (or of the curve $y^2 = 4x^3 - g_2(\Lambda)x - g_3(\Lambda)$) is defined as

$$j(\Lambda) := 1728 \frac{g_2(\Lambda)^3}{\Delta(\Lambda)}.$$

Exercise 2.33. Prove that $\Delta(\Lambda) \neq 0$ for any lattice Λ.

Theorem 2.34. *Two lattices Λ and Λ' are homothetic if and only if $j(\Lambda) = j(\Lambda')$.*

Proof. Assume that $\Lambda' = \alpha\Lambda$ for some $\alpha \in \mathbb{C}$. Then we have

$$g_2(\Lambda') = 60 \sum_{\omega \in \Lambda' - \{0\}} \frac{1}{\omega^4} = 60 \sum_{\omega \in \Lambda - \{0\}} \frac{1}{(\alpha\omega)^4} = \alpha^{-4} g_2(\Lambda).$$

Similarly, $g_3(\Lambda') = \alpha^{-6} g_3(\Lambda)$. Therefore,

$$j(\Lambda') = 1728 \frac{(\alpha^{-4} g_2(\Lambda))^3}{(\alpha^{-4} g_2(\Lambda))^3 - 27(\alpha^{-6} g_3(\Lambda))^2} = 1728 \frac{g_2(\Lambda)^3}{g_2(\Lambda)^3 - 27 g_3(\Lambda)^2} = j(\Lambda).$$

Conversely, assume that $j(\Lambda) = j(\Lambda')$. We need to find $\alpha \in \mathbb{C}$ such that $\Lambda' = \alpha\Lambda$. We consider 3 cases:

(1) $j = 0$, *i.e.* $g_2 = g_2' = 0$. Set $\alpha = (g_3/g_3')^{1/6}$.
(2) $j = 1728$, *i.e.* $g_3 = g_3' = 0$. Set $\alpha = (g_2/g_2')^{1/4}$.
(3) $j \neq 0, 1728$. Set $\alpha = (g_2/g_2')^{1/4} = (g_3/g_3')^{1/6}$.

Then we have $g_2(\Lambda') = \alpha^{-4}g_2(\Lambda) = g_2(\alpha\Lambda)$ and $g_3(\Lambda') = \alpha^{-6}g_3(\Lambda) = g_3(\alpha\Lambda)$. By Remark 2.26, we have $\wp(z;\Lambda') = \wp(z;\alpha\Lambda)$ and hence $\Lambda' = \alpha\Lambda$. □

Write $\Lambda = \mathbb{Z}\tau + \mathbb{Z}$, then we can view j as a function $\mathfrak{h} :\to \mathbb{C}$ by setting $j(\tau) = j(\Lambda_\tau)$.

Remark 2.35. It follows from Exercise 2.33 that $j : \mathfrak{h} \to \mathbb{C}$ is holomorphic.
Also, Theorem 2.34 implies that $j : \mathfrak{h} \to \mathbb{C}$ is invariant under $SL_2(\mathbb{Z})$.

Exercise 2.36. Prove that $j(i) = 1728$ and $j(\rho) = 0$.

Proposition 2.37. *Let* $\Delta(\tau)$ *denote* $\Delta(\Lambda_\tau)$. *Then*

$$\lim_{\mathrm{Im}(\tau)\to\infty} \Delta(\tau) = 0.$$

Proof. Write

$$g_2(\tau) = 60 \sum_{(m,n)\neq(0,0)} \frac{1}{(m+n\tau)^4} = 60\Big(2\sum_{m=1}^{\infty}\frac{1}{m^4} + \sum_{n\neq 0}\frac{1}{(m+n\tau)^4}\Big).$$

When $\mathrm{Im}(\tau) \to \infty$, we have $\frac{1}{(m+n\tau)^4} \to 0$. Hence

$$\lim_{\mathrm{Im}(\tau)\to\infty} g_2(\tau) = 120\sum_{m=1}^{\infty}\frac{1}{m^4} = 120\frac{\pi^4}{90} = \frac{4\pi^4}{3}.$$

Similarly,

$$\lim_{\mathrm{Im}(\tau)\to\infty} g_3(\tau) = 280\frac{\pi^6}{945} = \frac{8\pi^6}{27}.$$

Then

$$\lim_{\mathrm{Im}(\tau)\to\infty} \Delta(\tau) = (\frac{4\pi^4}{3})^3 - 27(\frac{8\pi^6}{27})^2 = 0.$$

□

Theorem 2.38. *j defines a bijection $Y(1) = \mathfrak{h}/\Gamma(1) \to \mathbb{C}$.*

Proof. The injectivity follows from Theorem 2.34.

Proposition 2.37 implies that j is not a constant. Hence j is a nonconstant holomorphic function on an open set $\mathfrak{h}$, hence the Open Mapping Theorem says that $j(\mathfrak{h})$ is open in $\mathbb{C}$. to conclude that $j(\mathfrak{h}) = \mathbb{C}$, we wish to show that $j(\mathfrak{h})$ is also closed. Let $j(z_1), j(z_2), \ldots$ be an arbitrary convergent sequence in $j(\mathfrak{h})$, converging to $\beta \in \mathbb{C}$. Since j is invariant under $SL_2(\mathbb{Z})$, we may assume that $z_i \in \mathcal{D}$ for each i. It follows from Proposition 2.37, that j is unbounded when $\mathrm{Im}(\tau) \to \infty$; hence we may assume that $\{\mathrm{Im}(z_i)\}_i$ is bounded. So, $z_1, z_2, \ldots$ come from compact subset Ω of $\mathcal{D}$. Thus, $z_1, z_2, \ldots$ converges to $z \in \Omega \subseteq \mathcal{D}$. Since j is continuous, $j(z) = \beta$. This shows that $j(\mathfrak{h})$ is closed and finishes the proof. □

Usually en elliptic is defined by $y^2 - f(x)$ where $f(x)$ is a degree 3 polynomial in x with 3 distinct roots. Theorem 2.31 tells us that each complex torus is an elliptic curve. To really prove that our Definition 2.1 agrees with the usual one, we also have to prove that following.

Theorem 2.39. *Assume that $f(x, y) = y^2 - 4x^3 - g_2x - g_3$ defines a nonsingular curve[3]. Then there is a unique lattice Λ such that $g_2 = g_2(\Lambda)$ and $g_2 = g_2(\Lambda)$.*

Proof. We will consider 3 cases.

(1) $g_2 = 0$. In this case, since f is nonsingular, we must have $g_3 \neq 0$. Set

$$\omega_1 = (\frac{g_3(\Lambda_\rho)}{g_3})^{1/6} \text{ and } \omega_2 = \rho\omega_1.$$

Set $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$. Then

$$g_2(\Lambda) = g_2(\omega_1\Lambda_\rho) = \frac{1}{\omega^4}g_2(\Lambda_\rho) = 0 = g_2$$
$$g_3(\Lambda) = g_3(\omega_1\Lambda_\rho) = \frac{1}{\omega^6}g_3(\Lambda_\rho) = g_3$$

(2) $g_3 = 0$. Then $g_2 \neq 0$. Choose ω_1 such that $\omega_1^4 = \frac{g_2\Lambda_i}{g_2}$ and $\omega_2 = i\omega_1$.

(3) $g_2g_3 \neq 0$. Theorem 2.38 guarantees the existence of $\tau \in \mathfrak{h}$ with $j(\tau) = \frac{g_2^3}{g_2^3 - 27g_3^2}$. Then it follows that

$$\frac{j(\tau) - 1}{j(\tau)} = \frac{27g_3^2}{g_2^3}. \tag{2.39.1}$$

Choose ω_1 such that

$$\omega_1^2 = \frac{g_2}{g_3}\frac{g_3(\Lambda_\tau)}{g_2(\Lambda_\tau)}$$

and set $\omega_2 = \tau\omega_1$, $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$. Then

$$\frac{g_2(\Lambda)}{g_3(\Lambda)} = \frac{\omega^{-4}g_2(\Lambda_\tau)}{\omega^{-6}g_3(\Lambda_\tau)} = \omega_1^2\frac{g_2(\Lambda_\tau)}{g_3(\Lambda_\tau)} = \frac{g_2}{g_3},$$

and hence

$$g_3(\Lambda) = \frac{g_3}{g_2}g_2(\Lambda) \tag{2.39.2}$$

On the other hand, we have

$$\frac{j(\tau) - 1}{j(\tau)} = \frac{27g_3^2(\Lambda)}{g_2^3(\Lambda)} = \frac{27(g_3/g_2)^2g_2^2(\Lambda)}{g_2^3(\Lambda)} = \frac{27g_3^2}{g_2^2g_2(\Lambda)}.$$

Comparing this with (2.39.1), we can see that $g_2(\Lambda) = g_2$ and by (2.39.2) we have $g_3(\Lambda) = g_3$.

□

[3]After an appropriate substitution, we may always assume that each degree 3 polynomial has the form $4x^3 - g_2x - g_3$.

3. MODULAR FORMS

Definition 3.1. A subgroup Γ of $SL_2(\mathbb{Z})$ is called *a congruence subgroup* if it contains $\Gamma(N)$ for some N.

Remark 3.2. There are many commonly used congruence subgroups of $SL_2(\mathbb{Z})$; we have seen $\Gamma(N)$ and there are two more:

$$\Gamma_1(N) := \{\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z}) \mid \begin{pmatrix} a & b \\ c & d \end{pmatrix} \equiv \begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix} \bmod N\}$$

$$\Gamma_0(N) := \{\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z}) \mid c \equiv 0 \bmod N\}$$

Clearly,

$$\Gamma(N) \subseteq \Gamma_1(N) \subseteq \Gamma_0(N) \subseteq SL_2(\mathbb{Z}).$$

Definition 3.3. Let k be an integer and Γ be a congruence subgroup of $SL_2(\mathbb{Z})$. A meromorphic function $f : \mathfrak{h} \to \mathbb{C}$ is called *weakly modular of weight k for* Γ if

$$f(\sigma(z)) = (cz + d)^k f(z) \tag{3.3.1}$$

for each $\sigma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$ and $z \in \mathfrak{h}$.

Remark 3.4. If f satisfies (3.3.1) for σ_1, σ_2, then f satisfies (3.3.1) for $\sigma_1\sigma_2$. Consequently, to check whether f is weakly modular with weight k for Γ, it suffices to check if f satisfies (3.3.1) for generators of Γ.

Example 3.5. Since $SL_2(z)$ is generated by $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, from Remark 3.4, f is weakly modular of weight k for $SL_2(\mathbb{Z})$ if and only if

$$f(z + 1) = f(z) \text{ and } f(-\frac{1}{z}) = z^k f(z).$$

Example 3.6. The congruence subgroup $\Gamma_0(4)$ can be generated by $\pm \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $\pm \begin{pmatrix} 1 & 0 \\ 4 & 1 \end{pmatrix}$. Hence f is weakly modular of weight 2 for $\Gamma_0(4)$ if

$$f(z + 1) = f(z) \text{ and } f(\frac{z}{4z + 1}) = (4z + 1)^2 f(z).$$

Remark 3.7. Assume that $-I_2 \in \Gamma$. If a nonzero function f is weakly modular of weight k for γ, then k must be an even integer since

$$f(z) = f(-I_2(z)) = (-1)^k f(z).$$

Example 3.8. Recall the Eisenstein series G_k ($k \geq 4$ is even) relative to a lattice Λ is defined as

$$G_k(\Lambda) = \sum_{\omega \in \Lambda - \{0\}} \frac{1}{\omega^k}.$$

For each $z \in \mathfrak{h}$, let Λ_z denote the lattice $\mathbb{Z}z + \mathbb{Z}$. Set

$$G_k(z) := G_k(\Lambda_z).$$

Then

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

where the equality holds since $(m+n,n)$ runs over $\mathbb{Z}^2 - \{(0,0)\}$ when (m,n) does so. Similarly,

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

where the last step uses the fact that the series converges absolutely uniformly. Hence $G_k(z)$ is weakly modular of weight k for $SL_2(\mathbb{Z})$ (when $k \geq 4$).

When $k=2$, one has to be careful; G_2 does *not* converges absolutely. One still has the following

$$G_2(z) = \sum_{m\neq 0} \frac{1}{m^2} + \sum_{n\neq 0}\sum_{m\in\mathbb{Z}} \frac{1}{(m+nz)^2}$$

but one can not change the order of the double sum. Consequently, $G_2(Z)$ does not satisfy $G_2(-\frac{1}{z}) = z^2 G_2(z)$. Indeed

$$G_2(\frac{az+b}{cz+d}) = (cz+d)^2 G_2(z) - \pi i c(cz+d)^{-1}.$$

But one can modify $G_2(z)$ by adding a correction term to make it modular, which is one of the key points to solve the sum of 4 squares problem as we will see later.

Exercise 3.9. Prove that

(1) if f is weakly modular of weight k for Γ, then f^ℓ is weakly modular of weight ℓk for Γ;
(2) if f, g are weakly modular of weights m, n for Γ respectively, then fg and f/g are weakly modular of weights $m+n$ and $m-n$ for Γ respectively.

Example 3.10. Combining Example 3.8 and Exercise 3.9, we can see that $G_4^3(z)$ and $G_6^2(z)$ are weakly modular of weight 12 for $SL_2(\mathbb{Z})$. Hence the discriminant $\Delta(z)$ is weakly modular of weight 12 for $SL_2(\mathbb{Z})$; consequently, the j-function $j(z) = 1728\frac{g_2^3(z)}{\Delta(z)}$ is weakly modular of weight 0 for $SL_2(\mathbb{Z})$.

Remark 3.11. Let Γ be a congruence subgroup and f be a weakly modular form of weight k for Γ. Let h be the least positive integer such that $T_h = \begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}$. Let $\mathbb{D}$ be the open disk $\{z \mid |z| < 1\}$ and $\mathbb{D}' = \mathbb{D} - \{0\}$. The function $q(z) = e^{2\pi i z}$ takes $\mathfrak{h}$ to $\mathbb{D}'$. Define $g : \mathbb{D}' \to \mathbb{C}$ by $g(z) = f(\frac{\log(z)}{2\pi i})$. Then

$$f(z) = g(q_h), \text{ where } q_h = e^{2\pi i z/h}.$$

If f is holomorphic on $\mathfrak{h}$, then g is also holomorphic (on the punctured disk). So g has a Laurent expansion $g(z) = \sum_{n\in\mathbb{Z}} a_n z^n$ for $z \in \mathbb{D}'$.

We will say that f is *holomorphic at* ∞ if the Laurent expansion of g sums over $n \in \mathbb{N}$.

In this case, f will have a Fourier expansion

$$f(z) = \sum_{n\in\mathbb{Z}} a_n(f) q_h^n, \ q_h = e^{2\pi i z/h}.$$

Example 3.12 (some q-expansions)**.** Consider the following identities:

$$\frac{1}{z}+\sum_{n=1}^{\infty}(\frac{1}{z-n}+\frac{1}{z+n})=\pi\cot(\pi z)=\pi i-2\pi i\sum_{m=0}^{\infty}q^m,\ q=e^{2\pi iz}.$$

By differentiating $k-1$ (with k even) times, we have

$$G_k(z)=2\sum_{n\neq 0}\frac{1}{n^k}+2\frac{(2\pi i)^k}{(k-1)!}\sum_{n=1}^{\infty}\sigma_{k-1}(n)q^n,\ q=e^{1\pi iz}$$

where $\sigma_{k-1}(n)=\sum_{d|n,d>0}d^{k-1}$.

Once we have the q-expansion of G_k, we have

$$\Delta(z)=q\prod_{n=1}^{\infty}(1-q^n)^24,\ q=e^{1\pi iz}.$$

And hence

$$j(z)=\frac{1}{q}+744+196884q+21493760q^2+\cdots$$

Definition 3.13. For each $\sigma=\begin{pmatrix}a & b\\ c & d\end{pmatrix}$ and integer k, we define the *weight-k operator* $[\sigma]_k$ on functions $f:\mathfrak{h}\to\mathbb{C}$ by

$$(f[\sigma]_k)(z)=(cz+d)^{-k}f(\sigma(z)),\ z\in\mathfrak{h}.$$

To simply notation, the factor $(cz+d)^{-k}$ is denoted by $j(\sigma,z)$, called the *factor of automorphy*.

Definition 3.14. Let Γ be a congruence subgroup of $SL_2(\mathbb{Z})$ and let k be an integer. A weakly modular form f of weight k with respect to Γ is called *modular* if

(1) f is holomorphic,
(2) $f[\sigma]_k$ is holomorphic at ∞[4] for all $\sigma\in SL_2(\mathbb{Z})$.

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

Definition 3.15. Let f be a modular form of weight k with respect to a congruence subgroup Γ. If $a_0=0$ in the Fourier expansion of $f[\sigma]_k$ for all $\sigma\in SL_2(\mathbb{Z})$, then f is called a *cusp form* of weight k with respect to Γ.

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

Example 3.16. G_k with $k\geq 4$ is a modular form of weight k for $SL_2(\mathbb{Z})$.
Δ is a cusp form of weight 12 for $SL_2(\mathbb{Z})$.
j is weakly modular, but not modular.

It should be clear that both $\mathcal{M}_k(\Gamma)$ and $\mathcal{S}_k(\Gamma)$ are $\mathbb{C}$-vector spaces. The condition (2) in Definition 3.14 is to guarantee that both $\mathcal{M}_k(\Gamma)$ and $\mathcal{S}_k(\Gamma)$ are finite dimensional. As a matter of fact, one has the following theorem for even integers[5] k.

[4]One may define holomorphy at ∞ as $\lim_{\mathrm{Im}(z)\to\infty}f(z)$ exists.
[5]For odd integers, the dimension formula is more complicated; one has to differentiate regular cusps (integer weights) from irregular cusps (half integer weights). We will not discuss it here.

Theorem 3.17. *Let k be an even integer, Γ be a congruence subgroup of $SL_2(\mathbb{Z})$, g_Γ be the genus of $X(\Gamma)$, ε_i denote the number of elliptic points of period i ($i = 2, 3$), ε_∞ denote the number of cusps. Then*

$$\dim_{\mathbb{C}}(\mathcal{M}_k(\Gamma)) = \begin{cases} (k-1)(g-\Gamma-1) + \lfloor \frac{k}{4} \rfloor \varepsilon_2 + \lfloor \frac{k}{3} \rfloor \varepsilon_3 + \frac{k}{2}\varepsilon_\infty & k \geq 2 \\ 1 & k = 0 \\ 0 & k < 0 \end{cases}$$

and

$$\dim_{\mathbb{C}}(\mathcal{S}_k(\Gamma)) = \begin{cases} (k-1)(g-\Gamma-1) + \lfloor \frac{k}{4} \rfloor \varepsilon_2 + \lfloor \frac{k}{3} \rfloor \varepsilon_3 + (\frac{k}{2}-1)\varepsilon_\infty & k \geq 4 \\ g_\Gamma & k = 2 \\ 0 & k \leq 0 \end{cases}$$

The proof of this theorem will be postponed. We will look at an application first: sums of squares and the theta function. It may get a bit technical. So, let me explain the underlying idea/strategy: assume that we already identify a modular form related to the problem at hand

(1) the dimension formula implies that, oftentimes $\dim_{\mathbb{C}}(\mathcal{M}_k(\Gamma))$ is small; for instance
$$\dim_{\mathbb{C}}(\mathcal{M}_2(\Gamma_0(4))) = 2;$$
(2) find a basis for $\dim_{\mathbb{C}}(\mathcal{M}_k(\Gamma))$ and write the modular form that we are interested in terms of this basis;
(3) comparing the coefficients in the q-expansions will give us some desirable identities.

3.1. **Theta function.** We begin the following definition.

Definition 3.18. The theta function[6] $\theta : \mathfrak{h} \to \mathbb{C}$ is defined by

$$\theta(z) = \sum_{n=-\infty}^{\infty} e^{2\pi i n^2 z} = 1 + 2q + 2q^4 + 2q^9 + \cdots,$$

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

The connection between θ and the problem of writing an integer as a sum of a certain number of squares is indicated in the following exercise.

Exercise 3.19. Set

$$r(n,k) = \mid \{(n_1, \ldots, n_k) \in \mathbb{Z}^k \mid n = n_1^2 + \cdots + n_k^2\} \mid .$$

Prove that $\theta(z)^k = \sum_{n=0}^{\infty} r(n,k) q^n,\ q = e^{2\pi i z}$.

Hence the coefficient of q^n in θ^k is precisely the number of ways to write n as a sum of k squares. We will consider θ^4 and solve the problem of writing an integer n as a sum of 4 squares. To this end, we need some general facts on $\theta(z)$.

[6]More generally, the theta function is defined by

$$\theta(z,\tau) = \sum_{n=-\infty}^{\infty} e^{2\pi i n^2 z} e^{2\pi i n \tau}.$$

It appears in many different branches of mathematics; for instance, it is the solution to the heat equation, it induces the theta divisor on an abelian variety, etc. David Mumford wrote 3 volumes to explore theta functions from many different viewpoints. Our definition here is the simplified version.

Theorem 3.20. *The theta function satisfies*

$$\theta(z+1) = \theta(z), \quad \theta(\frac{-1}{4z}) = \sqrt{\frac{2z}{i}}\theta(z).$$

Proof. The first equality is clear (since $e^{2\pi im} = 1$ for any integer m). For the second equality, we need some Fourier analysis. The idea is to show that the second equality holds on the line segment $z = i\frac{t}{2}$ $(t > 0)$ and use the fact that any two holomorphic functions agreeing on a line segment must coincide (by analytic continuation). Denote q^{n^2} by $f(n)$ (here n is the considered as the input); hence $\theta(z) = \sum_{n=-\infty}^{\infty} f(n)$. We will set $z = i\frac{t}{2}$ $(t > 0)$. Then $f(x) = e^{-\pi x^2 t}$. From Fourier analysis, we have

$$\begin{aligned}
\hat{f}(y) &= \int_{-\infty}^{\infty} e^{2\pi ixy} f(x)dx = \int_{-\infty}^{\infty} e^{2\pi ixy - \pi x^2 t} dx \\
&= \int_{-\infty}^{\infty} e^{-\pi(\sqrt{t}x - i\frac{y}{\sqrt{t}})^2 - \pi\frac{y^2}{t}} dx \\
&= \frac{e^{-\pi y^2/t}}{\sqrt{t}} \int_{-\infty}^{\infty} e^{-\pi u^2} dx, \ (u = \sqrt{t}(x - iy/t)) \\
&= \frac{e^{-\pi y^2/t}}{\sqrt{t}}
\end{aligned}$$

hence

$$\sum_{n=-\infty}^{\infty} e^{-\pi n^2 t} = \frac{1}{\sqrt{t}} \sum_{n=-\infty}^{\infty} e^{-\pi n^2/t}$$

which shows the second transformation rule for $z = i\frac{t}{2}$. □

Corollary 3.21. *We have*

$$\theta(\frac{z}{4z+1}) = \sqrt{4z+1}\theta(z).$$

Proof.

$$\begin{aligned}
\theta(\frac{z}{4z+1}) = \theta(-\frac{1}{4(-\frac{1}{4z}-1)}) &= \sqrt{2i(\frac{1}{4z}+1)}\theta(-\frac{1}{4z}-1) \\
&= \sqrt{2i(\frac{1}{4z}+1)}\theta(-\frac{1}{4z}) = \sqrt{2i(\frac{1}{4z}+1)(-2iz)}\theta(z) \\
&= \sqrt{4z+1}\theta(z)
\end{aligned}$$

□

In particular, this shows that

$$\theta(\frac{z}{4z+1})^4 = (4z+1)^2\theta(z)^4.$$

And hence $\theta(z)^4$ is a modular form of weight 2 with respect to $\Gamma_0(4)$ (since $\Gamma_0(4)$ is generated by $\pm\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}$ and $\pm\begin{pmatrix}1 & 0\\ 4 & 1\end{pmatrix}$), *i.e.*

$$\theta^4 \in \mathcal{M}_2(\Gamma_0(4)).$$

It turns out that $\dim_{\mathbb{C}}(\mathcal{M}_2(\Gamma_0(4))) = 2$ and a basis is given by

$$G_{2,2}(z) = G_2(z) - 2G_2(2z)$$
$$G_{2,4}(z) = G_2(z) - 4G_2(4z)$$

Thus, $\theta^4 = aG_{2,2} + bG_{2,4}$. To find a, b, we will consider the Fourier expansions of $G_{2,2}$ and $G - 2, 4$.

To get Fourier expansions of $G_{2,2}$ and $G_{2,4}$, consider the following identities:

$$\frac{1}{z} + \sum_{n=1}^{\infty}(\frac{1}{z-n} + \frac{1}{z+n}) = \pi\cot(\pi z) = \pi i - 2\pi i\sum_{m=0}^{\infty} q^m,\ q = e^{2\pi i z}.$$

Differentiating both sides gives us

$$\sum_{n\in\mathbb{Z}} \frac{1}{(z+n)^2} = (2\pi i)^2 \sum_{\ell=1}^{\infty} \ell q^{\ell}.$$

Then

$$\begin{aligned} G_2(z) = \sum_{m,n\in\mathbb{Z};(m,n)\neq(0,0)} \frac{1}{(mz+n)^2} &= \sum_{n\neq 0}\frac{1}{n^2} + 2\sum_{m=1}^{\infty}\Big(\sum_{n\in\mathbb{Z}}\frac{1}{(mz+n)^2}\Big) \\ &= 2\sum_{n\neq 0}\frac{1}{n^2} + 2(2\pi i)^2\sum_{m=1}^{\infty}\sum_{\ell=1}^{\infty}\ell q^{m\ell} \\ &= 2\sum_{n\neq 0}\frac{1}{n^2} + 2(2\pi i)^2\sum_{n=1}^{\infty}\sigma_1(n)q^n, \end{aligned}$$

where $\sigma_1(n) = \sum_{d|n,d>0} d$.

More generally, by differentiating $k-1$ (with k even) times, we have

$$G_k(z) = 2\sum_{n\neq 0}\frac{1}{n^k} + 2\frac{(2\pi i)^k}{(k-1)!}\sum_{n=1}^{\infty}\sigma_{k-1}(n)q^n,$$

where $\sigma_{k-1}(n) = \sum_{d|n,d>0} d^{k-1}$.

From this, we can deduce that

$$G_{2,2}(z) = -\frac{\pi^2}{3}\Big(1 + 24\sum_{n=1}^{\infty}(\sum_{0<d|n;d\ \text{odd}} d)q^n\Big)$$
$$G_{2,4}(z) = -\pi^2\Big(1 + 8\sum_{n=1}^{\infty}(\sum_{0<d|n;4\nmid d} d)q^n\Big)$$

The expansions:

$$\theta(z)^4 = 1 + 8q + \cdots$$
$$-\frac{3}{\pi^2}G_{2,2}(z) = 1 + 24q + \cdots$$
$$-\frac{1}{\pi^2}G_{2,4}(z) = 1 + 8q + \cdots$$

show that[7] $\theta(z)^4 = -\frac{1}{\pi^2}G_{2,4}(z)$. Therefore they must have the same coefficients in their Fourier expansions, *i.e.*

$$r(n,4) = 8 \sum_{0<d|n;4\nmid d} d. \tag{3.21.1}$$

In particular, we recover the following result by Lagrange.

Theorem 3.22 (Lagrange)**.** *Each positive integer can be represented as the sum of 4 integer squares.*

Proof. Since $1 \mid n$ and $4 \nmid 1$, we have $r(n,4) > 0$ in (3.21.1). □

Remark 3.23. The sum-of-4-squares problem can also be solved via other methods; for instance the norm of an element in the quaternion algebra over $\mathbb{Z}$ is a sum of 4 squares and Hurwitz (as in Riemann-Hurwitz) proved Laglange's Theorem from this point of view. However, as we have seen, the approach via theta functions works more generally; it works for the sum of 6 squares, 8 squares, etc[8].

Department of Mathematics, University of Nebraska, Lincoln, Nebraska 68588
E-mail address: wzhang15@unl.edu

[7]One also see from these expansions that $G_{2,2}$ and $G_{2,4}$ are linearly independent; one is not a multiple of the other. Since $\dim(\mathcal{M}_2(\Gamma_0(4))) = 2$, they must form a basis.

[8]For odd number of squares, we need to discuss modular forms with half integer weights which we won't touch in this course.