Fall 2024 FLT Seminar: Modular Forms

Zachary Lihn

October 28, 2024

1 Hyperbolic plane and q-expansions

Let us begin with some examples. Recall that in our study of elliptic and theta functions, we were naturally led to consider functions of a *lattice* $\Lambda \subset \mathbb{C}$. We usually require these functions to be homogeneous, which allows us to scale and assume Λ can be written as the set

$$\Lambda = \mathbb{Z} + \mathbb{Z}\tau$$

where $\tau \in \mathbb{C}$ with $\operatorname{Im}\tau > 0$. As a result, these functions of lattices become functions of τ, and they often satisfy certain nice relations. One class of examples are the Eisenstein series, holomorphic homogeneous functions of degree $2k$ defined by and satisfying:

$$G_{2k}(\tau) = \sum_{(m,n)\in\mathbb{Z}^2\setminus(0,0)} \frac{1}{(m+n\tau)^{2k}}, \quad G_{2k}(\tau+1) = G_{2k}, \quad G_{2k}(-\frac{1}{\tau}) = \tau^{2k}G_{2k}(\tau).$$

Another function was the φ function defined as follows: setting $q = e^{2\pi i\tau}$ for the rest of the talk, we have

$$\varphi(\tau) = 2^8 q \prod_{n=1}^{\infty}(1-q^n)^{24},$$

which satisfies the identities

$$\varphi(\tau+1) = \varphi(\tau), \quad \varphi(-\frac{1}{\tau}) = \tau^{12}\varphi(\tau).$$

This function is defined a little differently from the Eisenstein series (i.e. in terms of q); we will see that they are really the same thing in a few minutes.

Going back to Eisenstein series, we saw how they show up naturally when studying elliptic curves. Indeed, the elliptic functions for a lattice Λ, which we assume to be of the form $\mathbb{Z} + \mathbb{Z}\tau$, are associated with the cubic

$$Y^2 = 4X^3 - g_4(\tau)X - g_6(\tau),$$

where $g_4 = 60G_4$ and $g_6 = 140G_6$. We define the Δ and j-invariants

$$\Delta(\tau) = g_4(\tau)^3 - 27g_6(\tau)^2$$
$$j(\tau) = 1728\frac{g_4(\tau)^3}{\Delta(\tau)}.$$

The various constants are chosen to make residue computations equal nice numbers (e.g. 1) down the road. The main point is that Δ is homogeneous of degree $4 \times 3 = 6 \times 2 = 12$, while j is homogeneous of degree 0. From the identities for the Eisenstein series we obtain

$$\Delta(\tau+1) = \Delta, \quad \Delta(-\frac{1}{\tau}) = \tau^{12}\Delta(\tau),$$
$$j(\tau+1) = j(\tau), \quad j(-\frac{1}{\tau}) = j(\tau).$$

We clearly see that these functions of τ all have similar structural properties, and it makes sense to study them from a more general perspective.

The first general observation is that the set $\{\tau \in \mathbb{C} \mid \operatorname{Im}\tau\}$ may be more properly called the half-plane $\mathcal{H}$. This is a frequently studied geometric object in complex analysis and hyperbolic geometry and is also known as the *hyperbolic plane*. With this geometric picture in mind, it becomes much clearer how e.g. Eisenstein series, defined in terms of $\tau \in \mathcal{H}$, are similar to φ, defined in terms of the parameter q lying in the unit disk D in $\mathbb{C}$.

Lemma 1.1. *Let f be a holomorphic function in $\mathcal{H}$ with period 1, i.e. $f(\tau+1) = f(\tau)$. Then there exists a unique holomorphic function g defined in the punctured disk $D^* = \{z \in \mathbb{C}, 0 < |z| < 1\}$, such that*

$$g(e^{2\pi i\tau}) = f(\tau).$$

Proof. The half plane maps to D^* via the map

$$\mathcal{H} \ni \tau \mapsto e^{2\pi i\tau} = e^{2\pi i \operatorname{Re}\tau}e^{-2\pi \operatorname{Im}\tau}.$$

Under the map, as $\operatorname{Im}\tau \to \infty$, we approach the origin in the disk which is why we have the puncture (draw a picture!). The map is a biholomorphism up to period 1, so composing it with a 1-periodic function gives the unique holomorphic function g in D^*. □

Of course, we would like to remove the puncture and obtain a holomorphic function in all of D. Entirely reasonable assumptions can make this happen.

Theorem 1.2. *Let f be a holomorphic function in $\mathcal{H}$ with period 1. Suppose that*

$$\lim_{|\operatorname{Im}\tau| \to +\infty} f(\tau) \to a_0$$

uniformly. Then f admits a q-expansion of the type

$$f(\tau) = g(e^{2\pi i\tau}) = \sum_{n=0}^{\infty} a_n q^n, \quad q = e^{2\pi i}.$$

Proof. Since $f(\tau) \to a_0$ as $|\operatorname{Im} \tau| \to \infty$ uniformly, also the function $g(q) \to a_0$ as $q \to 0$ (equivalently, as $|\operatorname{Im} \tau| \to \infty$). Thus g is holomorphic on the disk. □

Thus, we see that the function φ was just the q-expansion of some 1-periodic function with suitable behavior at infinity. In fact, *Jacobi's formula* says that up to scaling, φ is the q-expansion of Δ.

Remark 1.3. A very interesting argument in page 278 of Hellegouarch fully computes the q-expansion of G_{2k} in terms of the ζ-function and the divisor functions $\sigma_k(n) = \sum_{d|n} d^k$.

2 The modular group

We have already seen how seeing the parameter τ as lying in the geometric half-plane $\mathcal{H}$ led to a better geometric understanding of q-expansions for 1-periodic functions. We now turn towards understanding the other transformation relation $\tau \mapsto -\frac{1}{\tau}$ as well.

The main idea is that the group $SL_2(\mathbb{R})$ consisting of 2×2 real matrices of determinant 1 acts on $\mathcal{H}$ via *Mobius transformations*

$$\tau \mapsto \frac{a\tau + b}{c\tau + d}, \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{R}).$$

The determinant 1 condition is needed to send the half-plane to the half-plane. The idea is that $SL_2(\mathbb{R})$ is the group of *all* analytic automorphisms of $\mathcal{H}$, but it actually turns out to be too big: obviously I and $-I$ both fix every point. This turns out to be the only obstruction, and the group

$$PSL_2(\mathbb{R}) = SL_2(\mathbb{R})/\{I, -I\}$$

is the group of all analytic isomorphisms of $\mathcal{H}$.

Since our interesting functions behave well with respect to the transformations $\tau \mapsto \tau + 1$ and $\tau \mapsto -\frac{1}{\tau}$, both of which are analytic, we look for matrices that encode these transformations. Upstairs, they turn out to correspond to the following elements of $SL_2(\mathbb{R})$:

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

It follows that our functions behave in a controlled manner with respect to S and T. By induction, they behave in a controlled manner with respect to the entire subgroup generated by S and T! Fortunately, one can describe this group:

Fact: S and T generate

$$SL_2(\mathbb{Z}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{R}) \mid a, b, c, d \in \mathbb{Z} \right\}.$$

As for $SL_2(\mathbb{R})$, we need to quotient by $\{I, -I\}$ to account for equivalent automorphisms of $\mathcal{H}$. We arrive at the *modular group*

$$G = SL_2(\mathbb{Z})/\{I, -I\},$$

which is also generated by (the equivalence classes of) S and T.

Inspired by the examples from the start of the talk, we define the following.

Definition 2.1. A *weakly modular form* of weight k is a meromorphic function f in $\mathcal{H}$ satisfying

$$f(\frac{a\tau + b}{c\tau + d}) = (c\tau + d)^k f(\tau)$$

for every automorphism $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in G$.

The Eisenstein series are therefore weakly modular forms of weight $2k$; φ and Δ have weight 12; and j has weight 0.

To better understand modular forms, we need a piece of general terminology about group actions.

Definition 2.2. A *fundamental domain* of $\mathcal{H}$ for G is an open subset D of $\mathcal{H}$ which meets every orbit of G at exactly one point, and whose closure $\overline{D}$ contains exactly one point of each orbit.

By definition, a weakly modular form is determined by its values on a fundamental domain. Fortunately, some computations give a fundamental domain for G, which we omit since they are relatively straightforward but long.

Theorem 2.3. *A fundamental domain for G of $\mathcal{H}$ is*

$$D = \{\tau \in \mathcal{H} \mid |\operatorname{Re}\tau| < \frac{1}{2},\ |\tau| > 1\}.$$

Figure 1 gives a picture for the fundamental domain D, as well as its image under S and T. Note that D is a triangle: it has 3 vertices, one of which is the point at infinity $i\infty$. D and the rest of its orbits form a tiling of the hyperbolic plane.

While some of the vertices of these triangles meet in the interior of $\mathcal{H}$, others lie on the boundary line $\mathbb{R}$ and the point $i\infty$. We call these *cusps*. Cusps occur precisely at $i\infty$ and the rationals $\mathbb{Q}$: indeed, they are precisely the orbit of $i\infty$ under the $SL_2(\mathbb{Z})$ action, which can be easily seen to consist of the rationals.

Lastly, we remark that there are three points with nontrivial stabilizer: the point $i = e^{i\pi/2}$ (with stabilizer size 2) and $\rho = e^{2\pi i/3}$ and its reflection across the y-axis (with stabilizer size 3).

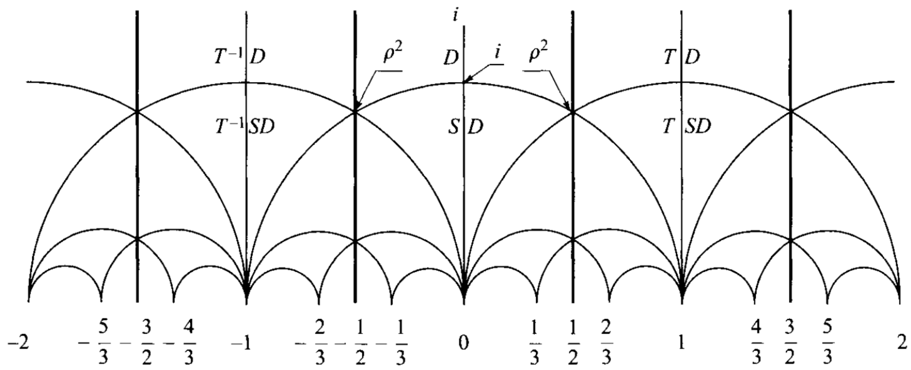

Figure 1: Tilings by fundamental domain, page 282 of Hellegouarch.

3 Modular Forms

We now come to modular forms, which combine the nice transformation properties of weakly modular forms with sufficient growth conditions at infinity, i.e.the cusps, to give q-expansions.

Definition 3.1. A meromorphic function $f : \mathcal{H} \to \mathbb{C}$ is called a *meromorphic modular form of weight* k if:

1. (Modularity) For every $\tau \in \mathcal{H}$ and $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z})$,

$$f(\frac{a\tau + b}{c\tau + d}) = (c\tau + d)^k f(\tau);$$

2. (Meromorphicity at the cusps) For every $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z})$, the function

$$(c\tau + d)^{-k} f(\frac{a\tau + b}{c\tau + d})$$

has an expansion in powers of $q^{\frac{1}{N}}$ convergent in the punctured disk $\{q \in \mathbb{C} \mid 0 < |q| < 1\}$, with only a finite number of terms with negative exponents.

While the first condition may seem complicated, in view of the generating set S and T it really just says that f is 1-periodic and has the nice transformation $f(-\frac{1}{\tau}) = \tau^k f(\tau)$.

The second condition means that f can be suitably extended as a meromorphic function along the cusps, like our assumption of uniform convergence to a constant in the theorem about q-expansions.

Note that any (weakly) modular form of odd weight equals zero by taking $a = d = -1$, $c = b = 0$.

As we have seen, the Eisenstein series G_{2k} are modular forms of weight 2k. It follows that Δ (and hence φ) and j are modular forms also.

We will now try to prove some properties of modular forms and connect them back to elliptic curves. Recall that if f is a meromorphic function around a point x_0, then its order may be defined as follows: we know there is a Laurent's series expansion

$$f(x) = \sum_{i=k}^{\infty} a_i (x - x_0)^i$$

in a neighborhood of x_0, where $a_k \neq 0$. The order $\nu_{x_0}(f)$ of f at x_0 is then equal to this integer k. Equivalently, it is the unique k such that $(x - x_0)^{-k} f$ is an invertible holomorphic function in a neighborhood of x_0 (i.e. it doesn't vanish at x_0).

When f obeys the modularity condition, the order becomes G-invariant, i.e.

$$\nu_{g \cdot x_0}(f) = \nu_{x_0}(f)$$

for any $g \in G$, the modular group. We may therefore speak of the order of f, denoted $\nu_P(f)$, at $P \in \mathcal{H}/G$, by defining it to be the order of f at any point in the preimage of P under the quotient. If also f is meromorphic at the cusps, we may define $\nu_\infty(f)$ to be the order of its q-expansion at the origin.

Theorem 3.2. *Let f be a modular form of weight k. Then,*

$$\nu_\infty(f) + \frac{1}{2}\nu_i(f) + \frac{1}{3}\nu_\rho(f) + \sum_{P \in \mathcal{H}/G, P \neq i, \rho} \nu_P(f) = \frac{k}{12}.$$

Here i and ρ denote the projections of the points $e^{i\pi/2}$ and $e^{2\pi i/3}$ with nontrivial stabilizer size 2 and 3, respectively.

Note that the sum is finite, i.e. $\nu_P(f)$ is zero for all but finitely many points, by the meromorphicity assumption at the cusps.

Proof. The idea is to apply Cauchy's theorem to the some cutoff $D_\alpha = \{\tau \in D \mid \operatorname{Im} \tau \leq \alpha\}$ of the fundamental domain D containing all the zeroes and poles.

When f has no zeroes or poles on the boundary, Cauchy's theorem says

$$\frac{1}{2\pi i} \int_{\partial D_\alpha} \frac{df}{f} = \sum_{P \in \mathcal{H}/G} \nu_P(f).$$

Since f has period 1, the integrals over AB and DE cancel each other out. If $g(q)$ is the q-expansion of f, then the integral over AE is

$$\frac{1}{2\pi i} \int_{EA} \frac{df}{f} = \frac{1}{2\pi i} \int_{\gamma} \frac{dg}{g} = -\nu_\infty(f)$$

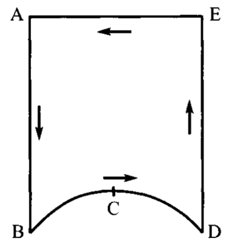

Figure 2: The domain D_α. Here $B = e^{2\pi i/3}$ and $D = e^{\pi i/3}$

where γ is some small circle, oriented clockwise, around the origin. The minus comes from it being opposite the usual orientation.

It remains to integrate along BD, which we divide into BC and CD. Using the transformation $f(-\frac{1}{\tau}) = \tau^k f(\tau)$, the integral over CD can be written

$$\int_{CD} \frac{df}{f}(\tau) = \int_{CB} \frac{df}{f}(-\frac{1}{\tau}) = \int_{CB} \frac{d(\tau^k f(\tau))}{\tau^k f(\tau)} = \int_{CB} \frac{kd\tau}{\tau} + \frac{df(\tau)}{f(\tau)}.$$

Therefore,

$$\frac{1}{2\pi i}\int_{BD} \frac{df}{f} = \frac{1}{2\pi i}\int_{BC} \frac{kd\tau}{\tau} = \frac{1}{2\pi i}\int_{\theta=\pi/3}^{\pi/2} \frac{kd(e^{i\theta})}{e^{i\theta}} = \frac{k}{12}$$

which proves the theorem when f doesn't have zeroes or poles on the boundary.

When f does have zeroes or poles on the boundary, we make a small detour in our path and take limits as the detour deforms to the original path. □

Applying our theorem to the j-invariant yields a powerful corollary.

Corollary 3.3. *Let*

$$j = 1728\frac{g_4^3}{\Delta} = 1728\frac{g_4^3}{g_4^3 - 27g_6^2}$$

be the modular invariant*, which we know is a modular function.*

1. *The function j is holomorphic in $\mathcal{H}$ and has a simple pole at $i\infty$ with residue 1.*
2. *Let $\widehat{\mathcal{H}/G}$ be $\mathcal{H}/G$ compactified with the point at infinity $i\infty$. The function j induces a bijection*

$$\mathcal{H}/G \to \mathbb{C}P^1 = \mathbb{C} \cup \{\infty\}.$$

Proof. We know that Δ doesn't vanish in $\mathcal{H}$ (e.g. by the q-expansion as φ), so j is holomorphic in $\mathcal{H}$. To see the pole at infinity, we use the first few terms of the q-expansions for the

Eisenstein series I didn't write down before. We have

$$g_4 = 60G_4 = \frac{4}{3}\pi^4(1 + 240q + \cdots)$$
$$g_6 = 140G_6 = \frac{8}{27}\pi^6(1 - 504q + \cdots)$$

and so

$$g_4^3 - 27g_6^2 = \frac{2^6}{3^3}\pi^{12}((1 + 240q + \cdots)^3 - (1 - 504q + \cdots)^2) = \frac{2^6}{3^3}\pi^{12}(1728q + \cdots).$$

The g_4^3 in the numerator cancels out the constants in front, and the 1728 gets canceled so that j has a simple pole of residue 1 at $i\infty$.

If $\lambda \in \mathbb{C}$, the equation $j(\tau) = \lambda$ is the same as solving for zeroes of the holomorphic weight 12 modular form

$$f_\lambda(\tau) = 1728g_4^3 - \lambda\Delta.$$

From the Theorem, we have

$$\nu_\infty(f_\lambda) + \frac{1}{2}\nu_i(f_\lambda) + \frac{1}{3}\nu_\rho(f_\lambda) + \sum_{P \in \mathcal{H}/G, P \neq i, \rho} \nu_P(f_\lambda) = 1$$

Therefore, f_λ must have at least one zero (otherwise the sum would be zero). This zero must be unique–if it were not, points other than i and ρ would make the LHS sum to greater than 1, while the contributions from both i and ρ will never sum to 1 since the denominators are coprime.

When $\lambda = \infty$, we can take $\tau = i\infty$. □

Corollary 3.4. *$\hat{\mathcal{H}}/G$ can be given the structure of a compact Riemann surface isomorphic to $\mathbb{C}P^1$ using j.*

Proof. Use the bijection j to port the charts over. □

Theorem 3.5. *Every smooth Weierstrass cubic over $\mathbb{C}$ can be parametrised by elliptic functions.*

Proof. Given a lattice $\Lambda \subset \mathbb{C}$, recall its Weierstrass $\wp$-function

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

Now let $C : Y^2 = 4X^3 - AX - B$ be our cubic and assume that $A^3 - 27B^2 \neq 0$. Then the modular invariant j_C of the cubic is finite and there is $\tau \in \mathcal{H}$ such that

$$j(\tau) = j_C$$

by the Theorem. If now Λ_τ is the lattice $\mathbb{Z} + \mathbb{Z}\tau$, then its Weierstrass function $\wp_\tau$ satisfies

$$W : \wp'(z)^2 = 4\wp_\tau(z)^3 - g_4(\tau)\wp_\tau(z) - g_6(\tau).$$

Since W and C have the same j-invariant, we know they are isomorphic, and so there is some $\lambda \in \mathbb{C}^*$ such that $A = \lambda^4 g_4(\tau), B = \lambda^6 g_6(\tau)$. Using homogeneity, it follows that $A = g_4(\lambda\tau), B = g_6(\lambda\tau)$ and so if $\wp$ is the function associated to the lattice $\lambda\Lambda_\tau$, we get

$$\wp'(z)^2 = 4\wp(z)^3 - A\wp(z) - B.$$

□

Remark 3.6. We know that the j-invariant classifies elliptic curves over $\mathbb{C}$. The Theorem says that the space of all (equivalence classes of) elliptic curves over $\mathbb{C}$ is $\widehat{\mathcal{H}/G} \cong \mathbb{P}^1$!.

Remark 3.7. Interestingly, we saw how j may have "higher order" multiplicities at the points i and ρ. This ultimately stemmed from the existence of a stabilizer subgroup of G at these points. This translates into "extra automorphisms" of the associated elliptic curves which the j-invariant doesn't detect. This means that $\mathbb{P}^1$ is a "course moduli space" for elliptic curves. To be able to keep track of these extra automorphisms requires developing a theory of stacks, and i and ρ are the "stacky/orbifold" points.

Remark 3.8. There are some distinguished subgroups of the $SL_2(\mathbb{Z})$, defined by reduction modulo N as follows. Let $N \geq 1$ and consider the reduction map

$$SL_2(\mathbb{Z}) \xrightarrow{\pi_N} SL_2(\mathbb{Z}/N\mathbb{Z})$$
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \begin{pmatrix} \overline{a} & \overline{b} \\ \overline{c} & \overline{d} \end{pmatrix}$$

where the overline denotes reduction modulo N. The kernel of π_N is the *principal congruence subgroup* of level N, written $\Gamma(N)$, i.e.

$$\Gamma(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z}) \mid \begin{pmatrix} a & b \\ c & d \end{pmatrix} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \mod N \right\}.$$

There are also *Hecke subgroups* of level N, defined as

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

$$\Gamma^0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z}) \mid c \equiv 0 \mod N \right\}.$$

Note that the Hecke subgroups and $\Gamma(2)$ contain $\{I, -I\}$ and so descend to the modular group. However, in general the $\Gamma(N)$ do not descend to the modular group.

In any case, we may form the quotient $\mathcal{H}/H$ for H any of these Hecke or congruence subgroups. This yields a noncompact Riemann surface, which we compactify by adding the cusps to obtain a *modular curve* $X(H)$. The modular curves have many interesting arithemetic and geometric properties. They come with a covering map to $\widehat{\mathcal{H}/G} =: X(1)$. We may compute the genus in several different ways, but it turns out to be the same as the dimension of weight 2 modular forms with respect to H.