21.1 Modular curves

Definition 21.1. The *principal congruence subgroup* $\Gamma(N)$ is defined by

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

A *congruence subgroup* (of level N) is any subgroup of $\mathrm{SL}_2(\mathbb{Z})$ that contains $\Gamma(N)$. A *modular curve* is a quotient of $\mathbb{H}^*$ (or just $\mathbb{H}$) by a congruence subgroup.

Two families of congruence subgroups are of particular interest:

$$\Gamma_1(N) = \left\{ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z}) : a \equiv d \equiv 1 \bmod N \text{ and } c \equiv 0 \bmod N \right\};$$
$$\Gamma_0(N) = \left\{ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z}) : c \equiv 0 \bmod N \right\};$$

Note that $\Gamma(1) = \Gamma_1(1) = \Gamma_0(1) = \mathrm{SL}_2(\mathbb{Z})$. We now define the modular curves

$$X(N) = \mathbb{H}^*/\Gamma(N), \qquad X_1(N) = \mathbb{H}^*/\Gamma_1(N), \qquad X_0(N) = \mathbb{H}^*/\Gamma_0(N).$$

and similarly define $Y(N)$, $Y_1(N)$, and $Y_0(N)$, with $\mathbb{H}^*$ replaced by $\mathbb{H}$. Following the same strategy we used for $X(1)$, one can show that these are all compact Riemann surfaces.

Having defined the modular curves $X(N)$, $X_1(N)$, and $X_0(N)$, we now want to consider the meromorphic functions on these curves. We are specifically interested in $X_0(N)$, for reasons that will become clear shortly, but we begin with the general setup.

21.2 Modular functions

Modular functions are meromorphic functions on a modular curve. To make this statement more precise, we first need to discuss q-expansions. The map $q\colon \mathbb{H} \to \mathbb{D}$ defined by

$$q(\tau) = e^{2\pi i \tau} = e^{-2\pi \operatorname{im}\tau}\left(\cos(2\pi \operatorname{re}\tau) + i\sin(2\pi \operatorname{re}\tau)\right)$$

bijectively maps each vertical strip $\{\tau : n \le \operatorname{im}\tau > n+1\}$ of the upper half plane $\mathbb{H}$ to the open unit disk $\mathbb{D}$. If $f\colon \mathbb{H} \to \mathbb{C}$ is a meromorphic function that satisfies $f(\tau+1) = f(\tau)$ for all $\tau \in \mathbb{H}$, then we can write f in the form $f(\tau) = f^*(q(\tau))$, where f^* is a meromorphic function on the punctured unit disk $\mathbb{D}\backslash\{0\}$. The *q-series* for $f(\tau)$ is the Laurent-series expansion of f^* about 0 composed with $q(\tau)$:

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

where we typically just write q for $q(\tau)$ (as we will henceforth). We say that that f is *meromorphic at* ∞ whenever f^* is meromorphic at 0 (note that $\lim_{\operatorname{im}\tau\to\infty} q(\tau) = 0$). When this condition holds, we can write

$$f(\tau) = \sum_{n=n_0}^{\infty} a_n q^n,$$

with $a_{n_0} \neq 0$. The integer n_0 is the *order of* f *at* ∞.

More generally, if f satisfies $f(\tau+N)=f(\tau)$ for all $\tau\in\mathbb{H}$, then we can write f as

$$f(\tau)=f^*(q(\tau)^{1/N})=\sum_{n=-\infty}^{\infty}a_nq^{n/N}, \tag{1}$$

and we say that f is meromorphic at ∞ if f^* is meromorphic at 0.

If Γ is a congruence subgroup of level N, then for any Γ-invariant function f we have $f(\tau+N)=f(\tau)$ (consider $\gamma=\left(\begin{smallmatrix}1&N\\0&1\end{smallmatrix}\right)$), so f can be written in the form (1), and the same is true of the function $f(\gamma\tau)$, for any fixed $\gamma\in\Gamma$.

Definition 21.2. Let Γ be a congruence subgroup and let $f:\mathbb{H}\to\mathbb{C}$ be a Γ-invariant meromorphic function. The function $f(\tau)$ is said to be *meromorphic at the cusps* if for every $\gamma\in\mathrm{SL}_2(\mathbb{Z})$ the function $f(\gamma\tau)$ is meromorphic at ∞.

In terms of the extended upper half-plane $\mathbb{H}^*$, notice that for any $\gamma\in\mathrm{SL}_2(\mathbb{Z})$,

$$\lim_{\mathrm{im}\,\tau\to\infty}\gamma\tau\in\mathbb{H}^*\backslash\mathbb{H}=\mathbb{P}^1(\mathbb{Q}).$$

Thus to say that $f(\gamma\tau)$ is meromorphic at ∞ is the same thing as saying that $f(\tau)$ is meromorphic at the cusp $\gamma\infty$. Note that since f is Γ-invariant, in order to check whether or not f is meromorphic at the cusps, it suffices to consider a set of cusp representatives $\gamma_0\infty,\gamma_1\infty,\dots,\gamma_k\infty$ for Γ, which is a finite set, since Γ has finite index in $\Gamma(1)$.

Definition 21.3. Let Γ be a congruence subgroup. A *modular function* for Γ is a meromorphic function $g\colon\mathbb{H}^*/\Gamma\to\mathbb{C}$, equivalently, a Γ-invariant meromorphic function $f\colon\mathbb{H}\to\mathbb{C}$ that is meromorphic at the cusps.

It is straight-forward to check that constant functions and all sums, products, and quotients of modular functions Γ are also modular functions for Γ, thus the set of all modular functions for Γ forms a field. Notice that if $f(\tau)$ is a modular function for a congruence subgroup Γ, then $f(\tau)$ is also a modular function for every congruence subgroup Γ' contained in Γ: clearly $f(\tau)$ is Γ'-invariant since $\Gamma'\subseteq\Gamma$, and $f(\tau)$ is meromorphic at the cusps since $f(\gamma\tau)$ is meromorphic at ∞ for every $\gamma\in\Gamma$, which includes all $\gamma\in\Gamma'$.

21.3 Modular Functions for $\Gamma(1)$

We first consider the modular functions for $\Gamma(1)=\mathrm{SL}_2(\mathbb{Z})$. In Lecture 18 we proved that the j-function is $\mathrm{SL}_2(\mathbb{Z})$-invariant and holomorphic (hence meromorphic) on $\mathbb{H}$. To show that the $j(\tau)$ is a modular function for $\Gamma(1)$ we just need to show that it is meromorphic at the cusps. In this case the cusps are all $\Gamma(1)$-equivalent, so it suffices to show that the $j(\tau)$ is meromorphic at ∞, which we do by computing its q-expansion. We first note the following lemma, part of which was used in Problem Set 8.

Lemma 21.4. *Let $\sigma_k(n)=\sum_{d|n}d^k$, and let $q=e^{2\pi i\tau}$. We have*

$$g_2(\tau)=\frac{4\pi^4}{3}\left(1+240\sum_{n=1}^{\infty}\sigma_3(n)q^n\right),$$

$$g_3(\tau)=\frac{8\pi^6}{27}\left(1-504\sum_{n=1}^{\infty}\sigma_5(n)q^n\right),$$

$$\Delta(\tau)=g_2^3(\tau)-27g_3^2(\tau)=(2\pi)^{12}q\sum_{n=1}^{\infty}(1-q^n)^{24}.$$

Proof. See Washington [4, pp. 273-274]. □

Corollary 21.5. *With $q = e^{2\pi i\tau}$ we have*

$$j(\tau) = \frac{1}{q} + 744 + \sum_{n=1}^{\infty} a_n q^n,$$

where the a_n are integers.

Proof. We have

$$g_2^3(\tau) = \frac{64}{27}\pi^{12}(1 + 240q + O(q^2))^3 = \frac{64}{27}\pi^{12}(1 + 720q + O(q^2)),$$
$$\Delta(\tau) = \frac{64}{27}\pi^{12}(3^3 \cdot 2^6)q(1 - 24q + O(q^2)),$$

where each $O(q^2)$ denotes sums of higher order terms with integer coefficients. Thus

$$j(\tau) = \frac{1728 g_2^3(\tau)}{\Delta(\tau)} = \frac{1}{q} + 744 + \sum_{n=1}^{\infty} a_n q^n,$$

for some integers a_n, as desired. Note that the factor $1728 = 3^3 \cdot 2^6$ in the definition of the j-function is the smallest integer that makes all the a_n integers. □

The corollary implies that the j-function is a modular function for $\Gamma(1)$, with a simple pole at ∞. We proved in Theorem 18.5 that the j-function defines a holomorphic bijection from $Y(1) = \mathbb{H}/\Gamma$ to $\mathbb{C}$. If we extend the domain of j to $\mathbb{H}^*$ by defining $j(\infty) = \infty$, then the j-function defines an isomorphism from $X(1)$ to the Riemann sphere $\mathcal{S} = \mathbb{P}^1(\mathbb{C})$ that is holomorphic everywhere except for a simple pole at ∞. In fact, if we fix $j(\rho) = 0$, $j(i) = 1728$, and $j(\infty) = \infty$, then the j-function is uniquely determined by this property (fixing $j(i) = 1728$ ensures that it has an integeral q-expansion, a fact we will use shortly). It is for this reason that the j-function is sometimes referred to as *the* modular function. Indeed, every modular function for $\mathrm{SL}_2(\mathbb{Z})$ can be expressed in terms of the j-function.

Theorem 21.6. *Every modular function for $\Gamma(1)$ is a rational function of $j(\tau)$. Equivalently, $\mathbb{C}(j)$ is the field of modular functions for $\Gamma(1)$.*

Proof. Let $g\colon X(1) \to \mathbb{C}$ be a modular function for $\Gamma(1)$. Then $f = g \circ j^{-1}\colon \mathcal{S} \to \mathbb{C}$ is meromorphic. By Lemma 21.7 below, this implies that f is a rational function. Therefore $g = f \circ j \in \mathbb{C}(j)$, as desired. □

Lemma 21.7. *If $f\colon \mathcal{S} \to \mathbb{C}$ is meromorphic, then $f(z)$ is a rational function.*

Proof. We assume without loss of generality that f has no zeros or poles at ∞ (the north pole): if not, rotate it by replacing $f(z)$ by $f(z - c)$ with an appropriate constant c (in terms of $\mathbb{P}^1(\mathbb{C})$ this corresponds to applying a linear fractional transformation).

Let $\{p_i\}$ be the set of poles of f, and let m_i be the order of pole p_i. Similarly, let $\{q_j\}$ be the set of zeros of f, and let n_j be the order of zero q_j. We must have $\sum_i m_i = \sum_j n_j$, since f is a meromorphic function on $\mathcal{S}$ (to prove this, triangulate $\mathcal{S}$ so that all the poles and zeros of f lie in the interior of a triangle and note that the sum of the contour integrals

of f about all of the triangles (oriented counter-clockwise) must be zero). The function $h\colon \mathcal{S} \to \mathbb{C}$ defined by

$$h(z) = f(z) \cdot \frac{\prod_i (z-p_i)^{m_i}}{\prod_j (z-q_j)^{n_j}}$$

has no zeros or poles at points $P \neq \infty$, and since $\sum_i m_i = \sum_j n_j$ it cannot have a zero or pole at infinity. By Liouville's theorem, h is a constant function. Therefore

$$f(z) = c\,\frac{\prod_j (z-q_j)^{n_j}}{\prod_i (z-p_i)^{m_i}},$$

for some constant c, which is indeed a rational function. □

Corollary 21.8. *Every modular function $f(\tau)$ for $\Gamma(1)$ that is holomorphic on $\mathbb{H}$ is a polynomial in $j(\tau)$.*

Proof. Theorem 21.6 implies that f is a rational function in j, which we may write as

$$f(\tau) = c\,\frac{\prod_i (j(\tau)-\alpha_i)}{\prod_k (j(\tau)-\beta_k)},$$

for some $c, \alpha_i, \beta_k \in \mathbb{C}$. Recall that $j\colon \mathcal{F} \to \mathbb{C}$ is a bijection, so f has a pole at each $j^{-1}(\beta_k) \in \mathcal{F}$. But f is holomorphic on $\mathcal{F}$ and therefore has no poles in $\mathcal{F}$, so the denominator must be 1 and f is a polynomial in j. □

21.3.1 Modular functions for $\Gamma_0(N)$

We now consider the modular functions for the congruence subgroup $\Gamma_0(N)$.

Theorem 21.9. *The function $j_N(\tau) = j(N\tau)$ is a modular function for $\Gamma_0(N)$.*

Proof. The function j_N is obviously meromorphic (in fact holomorphic) on $\mathbb{H}$. That j_N is meromorphic at the cusps follows from the fact that j is meromorphic at the cusps, since τ is a cusp if and only if $N\tau$ is. We just need to show that j_N is $\Gamma_0(N)$-invariant.

Let $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0(N)$. We have

$$j_N(\gamma\tau) = j(N\gamma\tau) = j\left(\frac{N(a\tau+b)}{c\tau+d}\right) = j\left(\frac{aN\tau+bN}{\frac{c}{N}N\tau+d}\right) = j(\gamma' N\tau),$$

where

$$\gamma' = \begin{pmatrix} a & bN \\ c/N & d \end{pmatrix}.$$

We now note that $\gamma' \in \mathrm{SL}_2(\mathbb{Z})$, since $\det(\gamma') = \det(\gamma) = 1$ and $c \equiv 0 \pmod N$ implies that c/N is an integer. But, j is $\mathrm{SL}_2(\mathbb{Z})$-invariant, therefore

$$j_N(\gamma\tau) = j(\gamma' N\tau) = j(N\tau) = j_N(\tau),$$

hence j_N is $\Gamma_0(N)$-invariant, as desired. □

Theorem 21.10. *$\mathbb{C}(j, j_N)$ is the field of modular functions for $\Gamma_0(N)$.*

Cox gives a very concrete proof of this result in [1, Thm. 11.9]; here we give a simpler, but somewhat more abstract proof that is adapted from Milne [3, Thm. V.2.3].

Proof. Let $\{\gamma_1, \cdots, \gamma_m\} \subset \Gamma = \mathrm{SL}_2(\mathbb{Z})$ be a set of right coset representatives for $\Gamma_0(N)$ as a subgroup of Γ; this means that the cosets $\Gamma_0(N)\gamma_1, \cdots, \Gamma_0(N)\gamma_m$ are distinct and cover Γ. Let K_N denote the field of modular functions for $\Gamma_0(N)$. The field K_N is an extension of the field $\mathbb{C}(j)$, since the j-function is a modular function for $\Gamma_0(N)$ for any N, and K_N contains j_N, by Theorem 21.9, so it is also an extension of $\mathbb{C}(j, j_N)$, we just need to show that it is an extension of degree 1.

Consider any function $f \in K_N$. For all $\gamma \in \Gamma$, the set of functions $\{f(\gamma_i\gamma\tau)\}$ is equal to the set $\{f(\gamma_i\tau)\}$, since multiplying the right cosets by γ simply permutes them. Thus any symmetric polynomial in the $f(\gamma_i\tau)$ is Γ-invariant, and therefore a rational function of $j(\tau)$, by Theorem 21.6. Now let

$$P(Y) = \prod_{i\in\{1,\cdots,m\}} (Y - f(\gamma_i\tau)).$$

Then f is a root of P, since $f(\tau) = f(\gamma_i\tau)$ for the right coset $\Gamma_0(N)\gamma_i = \Gamma_0(N)$, and the coefficients of $P(Y)$ lie in $\mathbb{C}(j)$, since they are all symmetric polynomials in the $f(\gamma_i\tau)$. Since every $f \in K_N$ is the root of a monic polynomial over $\mathbb{C}(j)$ of degree m, it follows from the primitive element theorem that $[K_N : \mathbb{C}(j)] \leq m$.

Let $F \in \mathbb{C}(j)[Y]$ be the minimal polynomial for f, so that $F(j(\tau), f(\tau)) = 0$, where F is monic and irreducible in $\mathbb{C}(j)[Y]$. If we replace τ with $\gamma_i\tau$ then we have

$$F(j(\gamma_i\tau), f(\gamma_i\tau)) = F(j(\tau), f(\gamma_i\tau)) = 0,$$

so the functions $f(\gamma_i\tau)$ all have the same minimal polynomial as $f(\tau)$.

Now consider $f = j_N$. If we can show that the functions $j_N(\gamma_i\tau)$ are distinct, then the minimal polynomial of j_N must have degree equal to m, meaning that $[\mathbb{C}(j, j_N) : \mathbb{C}(j)] = m$, and therefore $[K : \mathbb{C}(j, j_N)] = 1$.

Assume to the contrary that $j(N\gamma_i\tau) = j(N\gamma_k\tau)$ (as functions of τ), for some $i \neq k$. Then, since j is injective on $\mathcal{F}$, there is a $\gamma \in \Gamma$ such that $N\gamma_i\tau = \gamma N\gamma_i\tau$ for all $\tau \in \mathbb{H}$. Indeed, pick $\alpha, \beta \in \Gamma$ so that $\alpha N\gamma_i\tau, \beta N\gamma_j\tau \in \mathcal{F}$, and then note that $j(\alpha N\gamma_i\tau) = j(\beta N\gamma_k\tau)$ if and only if $\alpha N\gamma_i = \beta N\gamma_k$.[1] So we may take $\gamma = \alpha^{-1}\beta = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. We then have

$$\begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix} \gamma_i \;=\; \pm \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix} \gamma_k,$$

and therefore

$$\gamma_i\gamma_k^{-1} \;=\; \pm \begin{pmatrix} 1/N & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix} \;=\; \pm \begin{pmatrix} a & b/N \\ cN & d \end{pmatrix}.$$

The matrix $\gamma_i\gamma_k^{-1}$ lies in Γ, since $\gamma_i, \gamma_k \in \Gamma$, so b/N is an integer, and $cN \equiv 0 \bmod N$, so in fact $\gamma_i\gamma_k^{-1} \in \Gamma_0(N)$. But then γ_i and γ_k lie in the same right coset, a contradiction. □

[1] Here we are thinking of N as the matrix $\left(\begin{smallmatrix} N & 0 \\ 0 & 1 \end{smallmatrix}\right)$, so that $N\tau = \frac{N\tau+0}{0\tau+1}$.

21.4 The modular polynomial

Definition 21.11. The *modular polynomial* Φ_N is the minimal polynomial of j_N over $\mathbb{C}(j)$.

As in the proof of Theorem 21.10, we can write $\Phi_N \in \mathbb{C}(j)(Y)$ as

$$\Phi_N(Y) = \prod_{i=1}^{m}(Y - j_N(\gamma_i\tau)),$$

where the γ_i are right coset representatives for $\Gamma_0(N)$. The coefficients of $\Phi_N(Y)$ are symmetric polynomials in $j_N(\gamma_i\tau)$, so, as in the proof of Theorem 21.10 they are Γ-invariant, and they are holomorphic on $\mathbb{H}$, and so they are polynomials in j, by Corollary 21.8. Thus $\Phi_N \in \mathbb{C}[j, Y]$, and we may regard Φ_N as a polynomial in two variables write it as $\Phi_N(X, Y)$.

Our next task is to prove that the coefficients of Φ_N are integers. To simplify the presentation, we will only prove this for prime N, which is all we need in most practical applications and suffices to prove the main theorem of complex multiplication. The proof for composite N is essentially the same, but explicitly writing down a set of right coset representatives γ_i and computing the q-expansions of the functions $j_N(\gamma_i\tau)$ is more complicated.

Lemma 21.12. *If N is prime, then the right cosets of $\Gamma_0(N)$ in Γ are*

$$\Big\{\Gamma_0(N)\Big\} \cup \Big\{\Gamma_0(N)ST^k : 0 \le k < N\Big\},$$

where $S = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and $T = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$.

Proof. We first show that the union of the cosets is Γ. Let $\gamma = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \in \Gamma$. If $C \equiv 0 \bmod N$, then $\gamma \in \Gamma_0(N)$ lies in the first coset. Otherwise, we note that

$$ST^k = \begin{pmatrix} 0 & -1 \\ 1 & k \end{pmatrix} \quad \text{and} \quad (ST^k)^{-1} = \begin{pmatrix} k & 1 \\ -1 & 0 \end{pmatrix},$$

and for $C \not\equiv 0 \bmod N$, we may pick k such that $kC \equiv D \bmod N$, since N is prime. Then

$$\gamma_0 = \gamma(ST^k)^{-1} = \begin{pmatrix} kA - B & A \\ kC - D & C \end{pmatrix} \in \Gamma_0(N),$$

so $\gamma = \gamma_0(ST^k) \in \Gamma_0(N)ST^k$.

We now show the cosets are distinct. Suppose not. Then there must exist $\gamma_1, \gamma_2 \in \Gamma_0(N)$ such that either (a) $\gamma_1 = \gamma_2 ST^k$ for some $0 \le k < N$, or (b) $\gamma_1 ST^j = \gamma_2 ST^k$ with $0 \le j < k < N$. Let $\gamma_2 = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. In case (a) we have

$$\gamma_1 = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & k \end{pmatrix} = \begin{pmatrix} b & bk - a \\ d & dk - c \end{pmatrix} \in \Gamma_0(N),$$

and therefore $d \equiv 0 \bmod N$. But then $\det \gamma_2 = ad - bc \equiv 0 \bmod N$, which is a contradiction. In case (b), with $m = k - j$ we have

$$\gamma_1 = \gamma_2 ST^m S^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & m \end{pmatrix}\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} a - bm & -b \\ c - dm & -d \end{pmatrix} \in \Gamma_0(N).$$

Thus $c - dm \equiv 0 \bmod N$, so $d \equiv 0 \bmod N$, and we again have $\det \gamma_2 = ad - bc \equiv 0 \bmod N$, which is a contradiction. □

Theorem 21.13. $\Phi_N \in \mathbb{Z}[X,Y]$.

Proof (for N prime). Let $\gamma_k = ST^k$. By Lemma 21.12 we have

$$\Phi_N(Y) = \big(Y - j_N(\tau)\big) \prod_{k=0}^{N-1} \big(Y - j_N(\gamma_k\tau)\big).$$

Let $f(\tau)$ be a coefficient of $\Phi_N(Y)$. Then $f(\tau)$ is holomorphic function on $\mathbb{H}$, since $j(\tau)$ is, $f(\tau)$ is $\Gamma(1)$-invariant, since, as in the proof of Theorem 21.10, it is symmetric polynomial in $j_N(\tau)$ and the functions $j_N(\gamma_k\tau)$, corresponding to a set of right coset representatives for $\Gamma_0(N)$, and $f(\tau)$ is meromorphic at the cusps, since it is a polynomial in functions that are meromorphic at the cusps. Thus $f(\tau)$ is a modular function for $\Gamma(1)$ and therefore a polynomial in $j(\tau)$, by Corollary 21.8. By Lemma 21.14 below, if we can show that the q-expansion of $f(\tau)$ has integer coefficients, then it will follow that $f(\tau)$ is an integer polynomial in $j(\tau)$ and therefore $\Phi_N \in \mathbb{Z}[X,Y]$.

We first show that $f(\tau)$ has have rational coefficients. We have

$$j_N(\tau) = j(N\tau) = \frac{1}{q^N} + 744 + \sum_{n=1}^{\infty} a_n q^{nN},$$

where the a_n are integers, thus $j_N \in \mathbb{Z}((q))$.

For $j_N(\gamma_k\tau)$, we have

$$\begin{aligned} j_N(\gamma_k\tau) = j(N\gamma_k\tau) &= j\left(\left(\begin{smallmatrix} N & 0 \\ 0 & 1 \end{smallmatrix}\right) ST^k\tau\right) \\ &= j\Big(S\left(\begin{smallmatrix} 1 & 0 \\ 0 & N \end{smallmatrix}\right)\left(\begin{smallmatrix} 1 & k \\ 0 & 1 \end{smallmatrix}\right)\tau\Big) = j\Big(\left(\begin{smallmatrix} 1 & k \\ 0 & N \end{smallmatrix}\right)\tau\Big) = j\left(\frac{\tau+k}{N}\right), \end{aligned}$$

where we are able to drop the S because $j(\tau)$ is Γ-invariant. If we let $\zeta_N = e^{\frac{2\pi i}{N}}$, then

$$e^{2\pi i\left(\frac{\tau+k}{N}\right)} = e^{2\pi i\frac{k}{N}} q^{1/N} = \zeta_N^k q^{1/N},$$

and

$$j_N(\gamma_k\tau) = \frac{\zeta_N^{-k}}{q^{1/N}} + \sum_{n=0}^{\infty} a_n \zeta_N^{kn} q^{n/N},$$

thus $j_N(\gamma_k\tau) \in \mathbb{Q}(\zeta_N)((q^{1/N}))$. Note that $\mathrm{Gal}(\mathbb{Q}(\zeta_N)/\mathbb{Q})$ permutes the $j_N(\gamma_k\tau)$ and fixes $j_N(\tau)$; it follows that $f \in \mathbb{Q}((q^{1/N}))$. But the coefficients of $f(\tau)$ are also algebraic integers, since the coefficients of $j_N(\tau)$ and the $j_N(\gamma_k)$ are, so in fact $f(\tau) \in \mathbb{Z}((q^{1/N}))$, and $f(\tau)$ is a polynomial in $j(\tau)$, so its q-expansion has only integral powers of q, and therefore $f(\tau) \in \mathbb{Z}((q))$, as desired. □

Lemma 21.14 (Hasse q-expansion principal)**.** *Let $f(\tau)$ be a modular function for $\Gamma(1)$ that is holomorphic on $\mathbb{H}$ and whose q-expansion has coefficients that lie in an additive subgroup A of $\mathbb{C}$. Then $f(\tau) = P(j(\tau))$, for some polynomial $P \in A[X]$.*

Proof. By Corollary 21.8, we know that $f(\tau) = P(j(\tau))$ for some $P \in \mathbb{C}[X]$, we just need to show that $P \in A[X]$. We proceed by induction on $d = \deg P$. The lemma clearly holds for $d = 0$, so assume $d > 0$. The q-expansion of the j-function begins with q^{-1}, so the q-expansion of $f(\tau)$ must have the form $\sum_{n=-d}^{\infty} a_n q^n$, with $a_n \in A$ and $a_{-d} \neq 0$. Let $P_1(X) = P(X) - a_{-d}X^d$, and let $f_1(\tau) = P_1(j(\tau)) = f(\tau) - a_{-d}j(\tau)^d$. The q-expansion of the function $f_1(\tau)$ has coefficients in A, and by the inductive hypothesis, so does $P_1(X)$, and therefore $P(X) = P_1(X) + a_{-d}X^d$ also has coefficients in A. □

21.5 Isogenies

Recall from Lecture 19 that if L_1 is a sublattice of L_2, and $E_1 \simeq \mathbb{C}/L_1$ and $E_2 \simeq \mathbb{C}/L_2$ are the corresponding elliptic curves, then there is an isogeny $\phi\colon E_1 \to E_2$ whose kernel is isomorphic to the finite abelian group L_2/L_1. Indeed, we have the commutative diagram

$$\begin{array}{ccc} \mathbb{C}/L_1 & \longrightarrow^{\iota} & \mathbb{C}/L_2 \\ \downarrow \wr & & \downarrow \wr \\ E_1(\mathbb{C}) & \longrightarrow^{\phi} & E_2(\mathbb{C}) \end{array}$$

where the top map is induced by the inclusion $L_1 \subseteq L_2$ (lift from $\mathbb{C}/L_1$ to $\mathbb{C}$ then quotient by the finer lattice L_2). The relationship between $E_1(\mathbb{C})$ and $E_2(\mathbb{C})$ is symmetric, since if we replace L_1 by the homothetic lattice $\frac{1}{N}L_1$, where $N = [L_2 : L_1]$ is the degree of ϕ, then L_2 is a sublattice of L_1 and we obtain the dual isogeny $\hat{\phi}\colon E_2 \to E_1$.

Definition 21.15. If L_1 is a sublattice of L_2 for which the group L_2/L_1 is cyclic, then we say that L_1 is a *cyclic sublattice* of L_2, and call the corresponding isogeny $\phi\colon E_1 \to E_2$ a *cyclic isogeny*.

Cyclic isogenies are of particular interest to us they are effectively parameterized by the modular polynomial Φ_N. We will prove this in the case that N is prime, but it holds for all N. We first want to describe the cyclic sublattices of prime index in a given lattice.

Lemma 21.16. *Let $L = [1, \tau]$ be a lattice with $\tau \in \mathbb{H}$. The cyclic sublattices of L with prime index N are precisely the lattice $[1, N\tau]$ and the lattices $[N, \tau + k]$, for $0 \le k < N$.*

Proof. The lattices $[1, N\tau]$ and $[N, \tau+k]$ are clearly index N sublattices of L, and they must be cyclic sublattices, since N is prime. Conversely, any sublattice $L' \subseteq L$ can be written as $[d, a\tau + k]$, where d is the least positive integer in L' and the index of L' in L is equal to ad. If $[L : L'] = N$ is prime, then either $d = 1$ and $a = N$, in which case $L' = [1, N\tau]$, or $d = N$ and $a = 1$, in which case $L' = [N, \tau + k]$ and we may assume $0 \le k < N$. □

Theorem 21.17. *For all $j_1, j_2 \in \mathbb{C}$, we have $\Phi_N(j_1, j_2) = 0$ if and only if j_1 and j_2 are the j-invariants of elliptic curves related by a cyclic isogeny of degree N.*

Proof for N prime. We will prove the equivalent statement that $\Phi_N\big(j(L_1), j(L_2)\big) = 0$ if and only if L_2 is homothetic to a cyclic sublattice of L_1 with index N. We may assume without loss of generality that $L_1 = [1, \tau_1]$ and $L_2 = [1, \tau_2]$, where $\tau_1, \tau_2 \in \mathbb{H}$. With $\gamma_k = ST^k$ as in the proof of Theorem 21.13, we have

$$\Phi_N\big(j(\tau), Y\big) = \big(Y - j(N\tau)\big) \prod_{k=0}^{N-1} \big(Y - j(N\gamma_k\tau)\big),$$

where

$$j\Big(N\gamma_k\tau\Big) = j\Big(\left(\begin{smallmatrix} N & 0 \\ 0 & 1 \end{smallmatrix}\right) ST^k\tau\Big) = j\Big(S\left(\begin{smallmatrix} 1 & k \\ 0 & N \end{smallmatrix}\right)\tau\Big) = j\Big(\left(\begin{smallmatrix} 1 & k \\ 0 & N \end{smallmatrix}\right)\tau\Big) = j\Big(\frac{\tau + k}{N}\Big).$$

Thus

$$\Phi_N\big(j(L_1), j(L_1)\big) = \Phi_N\big(j([1, \tau_1]), j([1, \tau_2])\big) = \Phi_N\big(j(\tau_1), j(\tau_2)\big) = 0$$

if and only if τ_2 is $\Gamma(1)$-equivalent to $N\tau_1$ or $(\tau_1 + k)/N$, with $0 \le k < N$. By Lemma 21.16, this is true precisely when L_2 is homothetic to a cyclic sublattice of L_1 with index N. □

Remark 21.18. We should note that if $\phi\colon E_1 \to E_2$ is a cyclic N-isogeny, the pair $(j(E_1), j(E_2))$ does *not* uniquely determine ϕ, even up to isomorphism. As an example, suppose $\mathrm{End}(E_1) \simeq \mathcal{O}$ and $\mathfrak{p}$ is an unramified proper $\mathcal{O}$-ideal of prime norm p such that $[\mathfrak{p}]$ has order 2 in the class group $\mathrm{cl}(\mathcal{O})$. Then $\mathfrak{p}E_1 \simeq \bar{\mathfrak{p}}E_1$, and there are two distinct p-isogenies from E_1 to $E_2 = \mathfrak{p}E_1$.[2] These isogenies are not isomorphic (there is no automorphism we can compose with one to get the other). In this situation the polynomial $\Phi_p(j(E_1), Y)$ will have $j(E_2)$ as a double root (this corresponds to a singularity on the curve $\Phi_N(X, Y) = 0$).

Corollary 21.19. $\Phi_N(X, Y) = \Phi_N(Y, X)$

Proof. This follows immediately from the existence of the dual isogeny. □

In the same way that the j-function defines a bijection from $Y(1) = \mathbb{H}/\Gamma(1)$ to $\mathbb{C}$ (which we may regard as an affine curve), the functions $j(\tau)$ and $j_N(\tau)$ define a bijection from $Y_0(N) = \mathbb{H}/\Gamma_0(N)$ to the affine curve C_N defined by $\Phi_N(X, Y) = 0$. Each $\tau \in \mathbb{H}$ is mapped to the point $\big(j(\tau), j_N(\tau)\big)$. If we take as a fundamental region for $\Gamma_0(N)$ the union of the translates $\gamma_i\mathcal{F}$, where $\mathcal{F}$ is a fundamental region for $\Gamma(1)$ and $\Gamma_0(N)\gamma_1, \ldots, \Gamma_0(N)\gamma_m$ are a set of right coset representatives for $\Gamma_0(N)$, then every point in $\mathbb{H}$ is $\Gamma_0(N)$-equivalent to some $\gamma_i\tau$ with $\tau \in \mathcal{F}$. The point $\gamma_i\tau$ is mapped to $\big(j(\gamma_i\tau), j_N(\gamma_i\tau)\big)$, but since $j(\gamma_i\tau) = j(\tau)$, this is the same thing as $\big(j(\tau), j_N(\gamma_i\tau)\big)$. For any $\tau \in \mathbb{H}$, the roots of $\Phi_N(j(\tau), Y)$ are the images under j_N of the m translates $\gamma_1\tau, \ldots, \gamma_m\tau$, each of which corresponds to an isomorphism class of elliptic curves that is related to $j(\tau)$ by a cyclic isogeny of degree N.

The map from $Y_0(N)$ to C_N extends uniquely to a map from $X_0(N)$ to the projective closure $\overline{C}_N$ of C_N, but the curve $\overline{C}_N$ is not smooth, so this is not an isomorphism of Riemann surfaces (it is everywhere except for a finite set of singular points). This defect can be remedied by "desingularizing" $\overline{C}_N$ (this involves embedding $\overline{C}_N$ in a higher dimensional projective space), and this yields a canonical map from $X_0(N)$ to the desingularization of $\overline{C}_N$ that is an isomorphism of compact Riemann surfaces. In this sense, the curve defined by $\Phi_N(X, Y) = 0$ can be viewed as a canonical model for $X_0(N)$ defined over $\mathbb{Q}$. This is a remarkable feature of the modular curves $X_0(N)$ that distinguishes them from most other modular curves.

21.6 Modular curves as moduli spaces

We have seen that the modular curve $X_0(N)$ can be viewed as parameterizing (isomorphism classes of) cyclic N-isogenies between elliptic curves over $\mathbb{C}$. This point of view gives an alternative way to *define* $X_0(N)$: it is the *moduli space* of cyclic N-isogenies of elliptic curves.[3] This may sound rather abstract, but it can be made quite rigorous in the language of algebraic geometry. Doing so is well beyond the scope of this course, but it is useful to have this perspective in mind, since it applies to other modular curves.

We have already seen that the modular curve $X(1)$ is the moduli space of all elliptic curves, and the modular curve $X(N)$ is the moduli space of triples (E, P_1, P_2), where $\{P_1, P_2\}$ is a basis for the N-torsion subgroup of E. The modular curve $X_1(N)$ is the moduli space of pairs (E, P), where P is a point of order N in $E(\mathbb{C})$. Of course one needs to define a suitable notion of isomorphism in each case.

[2] Recall that if $E_1 \simeq \mathbb{C}/L$ then $\mathfrak{p}E_1$ denotes the elliptic curve $E_2 \simeq \mathbb{C}/\mathfrak{p}^{-1}L$, see Lecture 19.

[3] Here we have specified the isomorphism class of an N-isogeny as the pair of j-invariants of its domain and codomain. Equivalently, $X_0(N)$ can be defined as the moduli space of (isomorphism classes of) pairs (E, C), where E is an elliptic curve and C is a cyclic subgroup of $E(\mathbb{C})$ of order N (any such C is the kernel of a cyclic N-isogeny), which is often done.

From our perspective the most critical fact about moduli spaces is that they can be interpreted over any field, not just $\mathbb{C}$. In the case of $X_0(N)$, the modular polynomial $\Phi_N(X,Y)$ has integer coefficients, so it defines a curve $\Phi_N(X,Y) = 0$ over any field. In particular, in any finite field $\mathbb{F}_q$, two elements $j_1, j_2 \in \mathbb{F}_q$ satisfy $\Phi_N(j_1,j_2) = 0$ if and only if they are the j-invariants of elliptic curves $E_1/\mathbb{C}$ and $E_2/\mathbb{C}$ that are related by a cyclic N-isogeny. One note of caution: over a non-algebraically closed field one needs to choose the curves E_1 and E_2 appropriately, the fact that $\Phi_N(j(E_1),j(E_2)) = 0$ does not guarantee that E_1 and E_2 are related by a cyclic N-isogeny, it only guarantees that there is a twist $\tilde{E}_2$ of E_2 for which this holds.[4]

References

[1] David A. Cox, *Primes of the form $x^2 + ny^2$: Fermat, class field theory, and complex multiplication*, Wiley, 1989.

[2] J. S. Milne, *Elliptic curves*, BookSurge Publishers, 2006.

[3] Lawrence C. Washington, *Elliptic curves: number theory and cryptography*, second edition, Chapman & Hall/CRC, 2008.

[4] Recall that $\tilde{E}/k$ is a *twist* of E/k if they are isomorphic over $\bar{k}$, equivalently, $j(\tilde{E}) = j(E)$.