Modular Functions and Modular Forms

J. S. Milne

ABSTRACT. These are the notes for Math 678, University of Michigan, Fall 1990, exactly as they were handed out during the course except for some minor revisions and corrections.

Please send comments and corrections to me at jmilne@umich.edu.

CONTENTS

Introduction

It is easy to define modular functions and forms, but less easy to say why they are important, especially to number theorists. Thus I will begin with a rather long overview of the subject.

Riemann surfaces. Let X be a connected Hausdorff topological space. A *coordinate neighbourhood* of $P \in X$ is a pair (U, z) with U an open neighbourhood of P and z a homeomorphism of U onto an open subset of the complex plane. A *complex structure* on X is a compatible family of coordinate neighbourhoods that cover X. A *Riemann surface* is a topological space together with a complex structure.

For example, any open subset X of $\mathbb{C}$ is a Riemann surface, and the unit sphere can be given a complex structure with two coordinate neighbourhoods, namely the complements of the north and south poles mapped onto the complex plane in the standard way. With this complex structure it is called the *Riemann sphere*. We shall see that a torus can be given infinitely many different complex structures.

Let X be a Riemann surface, and let V be an open subset of X. A function $f : V \to \mathbb{C}$ is said to be *holomorphic* if, for all coordinate neighbourhoods (U, z) of X, $f \circ z^{-1}$ is a holomorphic function on $z(U)$. Similarly, one can define the notion of a *meromorphic* function on a Riemann surface.

The general problem. We can state the following grandiose problem: study all holomorphic functions on all Riemann surfaces. In order to do this, we would first have to find all Riemann surfaces. This problem is easier than it looks.

Let X be a Riemann surface. From topology, we know that there is a simply connected topological space $\tilde{X}$ (the universal covering space of X) and a map $p : \tilde{X} \to X$ which is a local homeomorphism. There is a unique complex structure on $\tilde{X}$ for which $p : \tilde{X} \to X$ is a local isomorphism of Riemann surfaces. If Γ is the group of covering transformations of $p : \tilde{X} \to X$, then $X = \Gamma\backslash\tilde{X}$.

THEOREM 0.1. *A simply connected Riemann surface is isomorphic to (exactly) one of the following three:*

(a) $\mathbb{C}$;
(b) *the open unit disk* $D \stackrel{\text{df}}{=} \{z \in \mathbb{C} \mid |z| < 1\}$;
(c) *the Riemann sphere.*

PROOF. This is the famous Riemann mapping theorem. □

The main focus of this course will be on Riemann surfaces with D as their universal covering space, but we shall also need to look at those with $\mathbb{C}$ as their universal covering space; the third type will not occur.

Riemann surfaces that are quotients of D. In fact, rather than working with D, it will be more convenient to work with the complex upper half plane:

$$\mathbb{H} = \{z \in \mathbb{C} \mid \Im(z) > 0\}.$$

The map $z \mapsto \frac{z-i}{z+i}$ is an isomorphism of $\mathbb{H}$ onto D (in the jargon the complex analysts use, $\mathbb{H}$ and D are conformally equivalent). We want to study Riemann surfaces of the form $\Gamma\backslash\mathbb{H}$, where Γ is a discrete group acting on $\mathbb{H}$. How do we find such Γ's? There is an obvious big group acting on $\mathbb{H}$, namely, $\mathrm{SL}_2(\mathbb{R})$. For $\alpha = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R})$ and $z \in \mathbb{H}$, define

$$\alpha(z) = \frac{az+b}{cz+d}.$$

Then

$$\Im(\alpha z) = \Im\left(\frac{az+b}{cz+d}\right) = \Im\left(\frac{(az+b)(c\bar{z}+d)}{|cz+d|^2}\right) = \frac{\Im(adz+bc\bar{z})}{|cz+d|^2}.$$

But

$$\Im(adz+bc\bar{z}) = (ad-bc)\cdot\Im(z) = \Im(z),$$

because $\det(\alpha)=1$. Hence $\Im(\alpha z) = \Im(z)/|cz+d|^2$ for $\alpha \in \mathrm{SL}_2(\mathbb{R})$. In particular,

$$z \in \mathbb{H} \implies \alpha(z) \in \mathbb{H}.$$

Later we shall see that there is an isomorphism

$$\mathrm{SL}_2(\mathbb{R})/\{\pm I\} \to \mathrm{Aut}(\mathbb{H})$$

(bi-holomorphic automorphisms of $\mathbb{H}$). There are some obvious discrete groups in $\mathrm{SL}_2(\mathbb{R})$, for example, $\Gamma = \mathrm{SL}_2(\mathbb{Z})$. This is called the *(full) elliptic modular group.* For any $N \geq 0$, we define

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

and call it the *principal congruence subgroup* of level N; in particular, $\Gamma(1) = \mathrm{SL}_2(\mathbb{Z})$. There are many discrete subgroups in $\mathrm{SL}_2(\mathbb{R})$, but those of most interest to number theorists are the ones containing a principal congruence subgroup as a subgroup of finite index.

Let $Y(N) = \Gamma(N)\backslash\mathbb{H}$ and endow it with the quotient topology. Let $p : \mathbb{H} \to Y(N)$ be the quotient map. There is a unique complex structure on $Y(N)$ such that a function f on an open subset U of $Y(N)$ is holomorphic if and only if $f \circ p$ is holomorphic on $p^{-1}(U)$. Thus $f \mapsto f \circ p$ defines a one-to-one correspondence between holomorphic functions on $U \subset Y(N)$ and holomorphic functions on $p^{-1}(U)$ invariant under $\Gamma(N)$, i.e., such that $g(\gamma z) = g(z)$ for all $\gamma \in \Gamma(N)$.

The Riemann surface $Y(N)$ is not compact, but there is a natural way of compactifying it by adding a finite number of points. The compact Riemann surface is denoted by $X(N)$. For example, $Y(1)$ is compactified by adding a single point.

Modular functions. A *modular function* $f(z)$ of level N is a meromorphic function on $\mathbb{H}$ invariant under $\Gamma(N)$ and "meromorphic at the cusps". Because it is invariant under $\Gamma(N)$, it can be regarded as a function on $Y(N)$, and the second condition means that it remains meromorphic when considered as a function on $X(N)$, i.e., it has at worst a pole at each point of $X(N) \setminus Y(N)$.

In the case of the full modular group, it is easy to make explicit the condition "meromorphic at the cusps" (in this case, cusp). To be invariant under the full modular group means that

$$f\left(\frac{az+b}{cz+d}\right) = f(z) \text{ for all } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}).$$

Since $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$, we have that $f(z+1) = f(z)$. Any function satisfying this condition can be written in the form $f(z) = f^*(q)$, $q = e^{2\pi iz}$. As z ranges over the upper half plane, $q(z)$ ranges over a punctured disk in $\mathbb{C}$ (open disk with centred at 0, with 0 removed). To say that $f(z)$ is meromorphic at the cusp means that $f^*(q)$ is meromorphic on the whole disk; hence that f has an expansion

$$f(z) = \sum_{n \geq -N_0} a_n q^n.$$

Modular forms. To construct a modular function, we have to construct a meromorphic function on $\mathbb{H}$ that is invariant under the action of $\Gamma(N)$. This is difficult. It is easier to construct functions that transform in a certain way under the action of $\Gamma(N)$; the quotient of two such functions of same type will then be a modular function.

This is analogous to the following situation. Let $\mathbb{P}^1(k) = (k \times k \setminus \{(0,0)\})/k^\times$. We seek rational functions $f(X,Y) = g(X,Y)/h(X,Y)$ such that $(a,b) \mapsto f(a,b)$ is a function on $\mathbb{P}^1$ (with a few points removed). Thus we need $f(X,Y)$ to be invariant under the action of $k^\times$, i.e., such that $f(aX,aY) = f(X,Y)$, all $a \in k^\times$. Recall that a homogeneous form of degree d is a polynomial $h(X,Y)$ such that $h(aX,aY) = a^d h(X,Y)$ for all $a \in k^\times$. If g and h are homogeneous forms of the same degree, then g/h will be a rational function on $\mathbb{P}^1$.

The relation of homogeneous forms to rational functions on $\mathbb{P}^1$ is exactly the same as the relation of modular forms to modular functions.

DEFINITION 0.2. A holomorphic function $f(z)$ on $\mathbb{H}$ is a *modular form of level N and weight* 2k if

(a) $f(\alpha z) = (cz+d)^{2k} \cdot f(z)$, all $\alpha = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma(N)$;
(b) $f(z)$ is "holomorphic at the cusps".

For the full modular group, note that (a) again implies that $f(z+1) = f(z)$, and so f can be written as a function of $q = e^{2\pi iz}$; condition (b) then says that this function is holomorphic at 0, so that

$$f(z) = \sum_{n \geq 0} a_n q^n,\ q = e^{2\pi iz}.$$

Plane affine algebraic curves. . Let k be a field. A *plane affine algebraic curve* C over k is defined by a nonzero polynomial $f(X,Y) \in k[X,Y]$. The *points of C with coordinates in a field $K \supset k$* are the zeros of $f(X,Y)$ in $K \times K$; we write $C(K)$ for this set. Let $k[C] = k[X,Y]/(f(X,Y))$, and call it the *ring of regular functions* on C. When $f(X,Y)$ is irreducible (this is the most interesting case so far as we are concerned), we write $k(C)$ for the field of fractions of $k[C]$, and call it the *field of rational functions* on C.

We say that C is *nonsingular* if $f(X,Y)$, $\frac{\partial f}{\partial X}$, $\frac{\partial f}{\partial Y}$ have no common zero in the algebraic closure of k. A point where all three vanish is called a *singular point* on the curve.

EXAMPLE 0.3. Let C be the curve defined by $Y^2 = 4X^3 - aX - b$, i.e., by the polynomial

$$f(X,Y) = Y^2 - 4X^3 + aX + b.$$

Assume char $k \neq 2$. The derivatives of f are $2Y$ and $-12X^2 + a$. Thus a singular point on C is a pair (x,y) such that $y = 0$ and x is a repeated root of $4X^3 - aX - b$. Therefore C is nonsingular if and only if the roots of $4X^3 - aX - b$ are all simple, which is true if and only if its discriminant $\Delta = a^3 - 27b^2 \neq 0$.

PROPOSITION 0.4. *Let C be a nonsingular affine algebraic curve over $\mathbb{C}$; then $C(\mathbb{C})$ has a natural structure as a Riemann surface.*

PROOF. Let P be a point in $C(\mathbb{C})$, and suppose $(\partial f/\partial Y)(P) \neq 0$. Then the implicit function theorem shows that the projection $(x,y) \mapsto x : C(\mathbb{C}) \to \mathbb{C}$ defines a homeomorphism of an open neighbourhood of P onto an open neighbourhood of $x(P)$ in $\mathbb{C}$. This we take to be a coordinate neighbourhood of P. □

Plane projective curves. A *plane projective curve* C over k is defined by a nonconstant homogeneous polynomial $F(X,Y,Z)$. Let $\mathbb{P}^2(k) = (k^3 \setminus \{(0,0,0)\})/k^\times$, and write $(a:b:c)$ for the equivalence class of (a,b,c) in $\mathbb{P}^2(k)$. As $F(X,Y,Z)$ is homogeneous, $F(cx,cy,cz) = c^m \cdot F(x,y,z)$ for every $c \in k^\times$, where $m = \deg(F(X,Y,Z))$. Thus it makes sense to say $F(x,y,z)$ is zero or nonzero for $(x:y:z) \in \mathbb{P}^2(k)$. The *points of C with coordinates in a field $K \supset k$* are the zeros of $F(X,Y,Z)$ in $\mathbb{P}^2(K)$. Write $C(K)$ for this set. Let

$$k[C] = k[X,Y,Z]/(F(X,Y,Z)),$$

and call it the *homogeneous coordinate ring* of C. When $F(X,Y,Z)$ is irreducible, so that $k[C]$ is an integral domain, we write $k(C)$ for the subfield of the field of fractions of $k[C]$ of elements of degree zero (i.e., quotients of elements of the same degree), and we call it the *field of rational functions* on C.

A plane projective curve C is the union of three affine curves C_X, C_Y, C_Z defined by the polynomials $F(1,Y,Z)$, $F(X,1,Z)$, $F(X,Y,1)$ respectively, and we say that C is nonsingular if all three affine curves are nonsingular. There is a natural complex structure on $C(\mathbb{C})$, and the Riemann surface $C(\mathbb{C})$ is compact.

More generally, one can define a nonsingular projective algebraic curve C in projective space $\mathbb{P}^n$ of any dimension. Its points $C(\mathbb{C})$ again form a Riemann surface.

THEOREM 0.5. *Every compact Riemann surface S is of the form $C(\mathbb{C})$ for some nonsingular projective algebraic curve C, and C is uniquely determined up to isomorphism. Moreover, $\mathbb{C}(C)$ is the field of meromorphic functions on S.*

The statement is not true for noncompact Riemann surfaces, for example, $\mathbb{H}$ is not of the form $C(\mathbb{C})$ for C an algebraic curve, and neither is the graph of the exponential function $z \mapsto e^z$.

Arithmetic of Modular Curves. The theorem shows that we can regard $X(N)$ as an algebraic curve, defined by some homogeneous polynomial(s) with coefficients in $\mathbb{C}$. The central fact underlying the arithmetic of the modular curves (and hence of modular functions and modular forms) is that this algebraic curve is defined, in a natural way, over $\mathbb{Q}[\zeta_N]$, where $\zeta_N = \exp(2\pi i/N)$, i.e., the polynomials defining $X(N)$ (as an algebraic curve) can be taken to have coefficients in $\mathbb{Q}[\zeta_N]$, and there is a natural way of doing this.

This statement has as a consequence that it makes sense to speak of the points of $X(N)$ with coordinates in any field containing $\mathbb{Q}[\zeta_N]$. In the remainder of the introduction, I will explain what the points of $Y(1)$ are in any field containing $\mathbb{Q}$.

Elliptic curves. An *elliptic curve* E over a field k (of characteristic zero) is a plane projective curve given by an equation:

$$Y^2Z = 4X^3 - aXZ^2 - bZ^3, \quad \Delta \overset{\text{df}}{=} a^3 - 27b^2 \neq 0.$$

When we replace X with X/c^2 and Y with Y/c^3, some $c \in k^\times$, and multiply through by c^6, the equation becomes

$$Y^2Z = 4X^3 - ac^4XZ^2 - bc^6Z^3,$$

and so we should not distinguish the curve defined by this equation from that defined by the first equation. Note that

$$j(E) \overset{\text{df}}{=} 1728a^3/\Delta$$

is invariant under this change. In fact one can show (with a suitable definition of isomorphism) that two elliptic curves E and E' are isomorphic over an algebraically closed field if and only if $j(E) = j(E')$.

Elliptic functions. What are the quotients of $\mathbb{C}$? A *lattice* in $\mathbb{C}$ is a subset of the form

$$\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$$

with ω_1 and ω_2 complex numbers that are linearly independent over $\mathbb{R}$. The quotient $\mathbb{C}/\Lambda$ is (topologically) a torus. Let $p : \mathbb{C} \to \mathbb{C}/\Lambda$ be the quotient map. The space $\mathbb{C}/\Lambda$ has a unique complex structure such that a function f on an open subset U of $\mathbb{C}/\Lambda$ is holomorphic if and only if $f \circ p$ is holomorphic on $p^{-1}(U)$.

To give a meromorphic function on $\mathbb{C}/\Lambda$ we have to give a meromorphic function f on $\mathbb{C}$ invariant under the action of Λ, i.e., such that $f(z + \lambda) = f(z)$ for all $\lambda \in \Lambda$. Define

$$\wp(z) = \frac{1}{z^2} + \sum_{\lambda \in \Lambda, \lambda \neq 0} \left(\frac{1}{(z-\lambda)^2} - \frac{1}{\lambda^2} \right).$$

This is a meromorphic function on $\mathbb{C}$, invariant under Λ, and the map

$$[z] \mapsto (\wp(z) : \wp'(z) : 1) : \mathbb{C}/\Lambda \to \mathbb{P}^2(\mathbb{C})$$

defines an isomorphism of the Riemann surface $\mathbb{C}/\Lambda$ onto the Riemann surface $E(\mathbb{C})$, where E is the elliptic curve,

$$Y^2Z = 4X^3 - g_2XZ - g_3Z^3$$

for certain g_2 and g_3.

Elliptic curves and modular curves. We have a map $\Lambda \mapsto E(\Lambda) \stackrel{\text{df}}{=} \mathbb{C}/\Lambda$ from lattices to elliptic curves. When is $E(\Lambda) \approx E(\Lambda')$? If $\Lambda' = c\Lambda$ for some $c \in \mathbb{C}$, then

$$[z] \mapsto [cz] : \mathbb{C}/\Lambda \to \mathbb{C}/\Lambda'$$

is an isomorphism of Riemann surfaces, and in fact one can show

$$E(\Lambda) \approx E(\Lambda') \iff \Lambda' = c\Lambda, \text{ some } c \in \mathbb{C}^\times.$$

By scaling with an element of $\mathbb{C}^\times$, we can normalize our lattices so that they are of the form

$$\Lambda(\tau) = \mathbb{Z} \cdot 1 + \mathbb{Z} \cdot \tau, \text{ some } \ \tau \in \mathbb{H}.$$

Note that $\Lambda(\tau) = \Lambda(\tau')$ if and only if there is a matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$ such that $\tau' = \frac{a\tau+b}{c\tau+d}$. Thus we have a map

$$\tau \mapsto E(\tau) : \mathbb{H} \to \{\text{elliptic curves over } \mathbb{C}\}/\approx,$$

and the above remarks show that it gives an injection

$$\Gamma(1)\backslash\mathbb{H} \hookrightarrow \{\text{elliptic curves over } \mathbb{C}\}/\approx .$$

One shows that the function $\tau \mapsto j(E(\tau)) : \mathbb{H} \to \mathbb{C}$ is holomorphic, and has only a simple pole at the cusp; in fact

$$j(\tau) = q^{-1} + 744 + 196884q + 21493760q^2 + \cdots, \quad q = e^{2\pi i\tau}.$$

It is therefore a modular function for the full modular group. One shows further that it defines an isomorphism $j : Y(N) \to \mathbb{C}$. The surjectivity of j implies that every elliptic curve over $\mathbb{C}$ is isomorphic to one of the form $E(\tau)$, some $\tau \in \mathbb{H}$. Therefore

$$\Gamma(1)\backslash\mathbb{H} \ \stackrel{1:1}{\leftrightarrow} \ \{\text{elliptic curves over } \mathbb{C}\}/\approx .$$

The algebraic curve $Y(1)_{\mathbb{Q}}$ over $\mathbb{Q}$ naturally attached to $Y(1)$ has the property that its points with coordinates in L, L a field containing $\mathbb{Q}$, are given by

$$Y(1)(L) = \{\text{elliptic curves over } L\}/\sim,$$

where $E \sim E'$ if E and E' become isomorphic over the algebraic closure of L. Moreover, this property determines $Y(1)_{\mathbb{Q}}$.

From this, one sees that arithmetic facts about elliptic curves correspond to arithmetic facts about special values of modular functions and modular forms. For example, let E be an elliptic curve over a number field L; then, when regarded as an elliptic curve over $\mathbb{C}$, E is isomorphic to $E(\tau)$ for some $\tau \in \mathbb{C}$, and we deduce that

$$j(\tau) = j(E(\tau)) = j(E) \in L,$$

i.e., the transcendental function j takes a value at τ which is algebraic! For example, if $\mathbb{Z} + \mathbb{Z}\tau$ is the ring of integers in a quadratic imaginary field K, one can prove in this fashion that, not only is $j(\tau)$ algebraic, but it in fact generates the Hilbert class field of K (largest abelian extension of K unramified over K at all primes, including the infinite primes).

Books. Ahlfors, L.V., Complex Analysis, McGraw-Hill, 1953.

Cartan, H., Elementary Theory of Analytic Functions of One or Several Complex Variables, Addison Wesley, 1963. I much prefer this to Ahlfors's book.

Fulton, W., Algebraic Curves, Benjamin, 1969.

Griffiths, P. A., Introduction to Algebraic Curves, AMS, 1989.

Gunning, R., Lectures on Modular Forms, Princeton U. P., 1962. One of the first, and perhaps still the best, treatments of the basic material.

Koblitz, N., Introduction to Elliptic Curves and Modular Forms, Springer, 1984. He studies elliptic curves, and uses modular curves to help with this; we do the opposite. Nevertheless, there will be a large overlap between this course and the book.

Lang, S., Introduction to Modular Forms, Springer, 1976. The direction of this book is quite different from the course.

Miyake, T., Modular Forms, Springer, 1976. This is a very good source for the analysis one needs to understand the arithmetic theory, but he doesn't do much arithmetic.

Ogg, A., Modular Forms and Dirichlet Series, Benjamin, 1969. A useful book, but the organization is a little strange.

Schoeneberg, B., Elliptic Modular Functions, Springer, 1974. Again, he concentrates on the analysis rather than the arithmetic.

Serre, J.-P., Cours d'Arithmétique, Presses Univ. de France, 1970. The last chapter is a beautiful, but brief, introduction to modular forms.

Shimura, G., Introduction to the Arithmetic Theory of Automorphic Functions, Princeton U.P., 1971. An important book, but quite difficult. These notes may serve as an introduction to Shimura's book, which covers much more.

Silverman, J., The Arithmetic of Elliptic Curves, Springer, 1986.

Silverman, J., Advanced Topics in the Arithmetic of Elliptic Curves, Springer, 1994.

References of the form Math xxx are to course notes available at http://www.math.lsa.umich.edu/~jmilne

Part I: The Basic Theory

In this part, we develop the theory of modular functions and modular forms, and the Riemann surfaces on which they live.

1. Preliminaries

In this section we review some definitions and results concerning continuous group actions and Riemann surfaces. Following Bourbaki, we require (locally) compact spaces to be Hausdorff. We often use $[x]$ to denote the equivalence class containing x.

Continuous group actions. Recall that a group G with a topology is a *topological group* if the maps

$$(g,g')\mapsto gg':G\times G\to G,\quad g\mapsto g^{-1}:G\to G$$

are continuous. Let G be a topological group and let X be a topological space. An action of G on X,

$$(g,x)\mapsto gx:G\times X\to X,$$

is *continuous* if this map is continuous. Then, for each $g\in G$, $x\mapsto gx:X\to X$ is a homeomorphism (with inverse $x\mapsto g^{-1}x$). An *orbit* under the action is the set Gx of translates of an $x\in X$. The *stabilizer* of $x\in X$ (or the *isotropy group* at x) is

$$\mathrm{Stab}(x)=\{g\in G\mid gx=x\}.$$

If X is Hausdorff, then $\mathrm{Stab}(x)$ is closed (it is the inverse image of x under of $g\mapsto gx:G\to X$). There is a bijection

$$G/\mathrm{Stab}(x)\to Gx,\quad g\cdot\mathrm{Stab}(x)\mapsto gx;$$

in particular, when G acts transitively on X, there is a bijection

$$G/\mathrm{Stab}(x)\to X.$$

Let $G\backslash X$ be the set of orbits for the action of G on X. It is the set of equivalence classes for the obvious equivalence relation, and so acquires the quotient topology: if p denotes the map $x\mapsto Gx:G\to G\backslash X$, then $U\subset G\backslash X$ is open if and only if $p^{-1}(U)$ is open in G. Note that $p:X\to G\backslash X$ is both continuous and open. (It is continuous by definition; in fact, we have given $G\backslash X$ the finest topology for which p is a continuous map. Let U be an open subset of X; we want to show that $p(U)$ is open. But $p^{-1}(p(U))=\bigcup_{g\in G}gU$, which is clearly open.)

Let H be a subgroup of G. Then H acts on G on the left and on the right, and $H\backslash G$ and G/H are the spaces of right and left cosets.

LEMMA 1.1. *The space G/H is Hausdorff if and only if H is closed in G.*

PROOF. Write p for the map $G\to G/H$, $g\mapsto gH$. If G/H is Hausdorff, then the point eH in G/H is closed, and so $H=p^{-1}(eH)$ is closed (e =identity element).

Conversely, suppose H is a closed subgroup, and let aH and bH be distinct elements of G/H. Since G is a topological group, the map

$$f : G \times G \to G, \quad (g, g') \mapsto g^{-1}g',$$

is continuous, and so $f^{-1}(H)$ is closed. As $aH \neq bH$, $(a,b) \notin f^{-1}(H)$, and so there is an open neighbourhood of (a,b), which we can take to be of the form $U \times V$, that is disjoint from $f^{-1}(H)$. Now the images of U and V in G/H are disjoint open neighbourhoods of aH and bH. □

As we noted above, when G acts transitively on X, there is a bijection $G/\mathrm{Stab}(x) \to X$ for any $x \in X$. Under some mild hypotheses, this will be a homeomorphism.

PROPOSITION 1.2. *Suppose G acts continuously and transitively on X. If G and X are locally compact and Hausdorff, and there is a countable basis for the topology of G, then the map*

$$[g] \mapsto gx : G/\mathrm{Stab}(x) \to X$$

is a homeomorphism.

PROOF. We know the map is a bijection, and it is obvious from the definitions that it is continuous, and so we only have to show that it is open. Let U be an open subset of G, and let $g \in U$; we have to show that gx is an interior point of Ux.

Consider the map $G \times G \to G$, $(h, h') \mapsto ghh'$. It is continuous and maps (e,e) into U, and so there is a neighbourhood V of e, which we can take to be compact[1], such that $V \times V$ is mapped into U; thus $gV^2 \subset U$. After replacing V with $V \cap V^{-1}$, we can assume $V^{-1} = V$. (Here $V^{-1} = \{h^{-1} \mid h \in V\}$; $V^2 = \{hh' \mid h, h' \in V\}$.)

As $e \in V$, $G = \bigcup gV$ (union over $g \in G$). Each set gV is a union of open sets in the countable basis, and we only need to take enough g's in order to get each basic open set contained in a gV at least once. Therefore, there is a countable set of elements $g_1, g_2, \ldots \in G$ such that $G = \bigcup g_n V$.

As $g_n V$ is compact, its image $g_n V x$ in X is compact, and as X is Hausdorff, this implies that $g_n V x$ is closed. The following lemma shows that at least one of the $g_n V x$'s has an interior point. But $y \mapsto g_n y : X \to X$ is a homeomorphism mapping Vx onto $g_n V x$, and so Vx has interior point, i.e., there is a point $hx \in Vx$ and an open subset W of X such that $hx \in W \subset Vx$. Now

$$gx = gh^{-1} \cdot hx \in gh^{-1}W \subset gV^2x \subset Ux$$

which shows that gx is an interior point of Ux. □

LEMMA 1.3. *Let X be a nonempty locally compact (hence Hausdorff) space, and suppose $X = \bigcup V_n$ where (V_n) is a countable family of closed subsets. Then at least one of the V_n has an interior point.*

PROOF. The hypotheses imply that X is regular: the points of X are closed, and for any point x not in a closed set A, there are disjoint open sets U and V containing x and A respectively.

Suppose no V_n has an interior point. Take U_1 to be any nonempty open subset of X whose closure $\bar{U}_1$ is compact. As V_1 has empty interior, U_1 is not contained in

[1] not necessarily open

V_1, and because U_1 is regular, there is a nonempty open subset U_2 of U_1 such that $\bar{U}_2 \subset U_1 - V_1$. Continuing in this fashion, we obtain nonempty open sets U_3, U_4 ... such that $\bar{U}_{n+1} \subset U_n - V_n$. The $\bar{U}_n$ form a decreasing sequence of nonempty compact sets, and so $\cap \bar{U}_n \neq \varnothing$, which contradicts $X = \bigcup V_n$. □

Riemann surfaces: classical approach. Let X be a connected Hausdorff topological space. A *coordinate neighbourhood* for X is pair (U, z) with U an open subset of X and z a homeomorphism of U onto an open subset of the complex plane $\mathbb{C}$. Two coordinate neighbourhoods (U_i, z_i) and (U_j, z_j) are *compatible* if the function

$$z_i \circ z_j^{-1} : z_j(U_i \cap U_j) \to z_i(U_i \cap U_j)$$

is holomorphic with nowhere vanishing derivative (the condition is vacuous if $U_i \cap U_j = \emptyset$). A family of coordinate neighbourhoods $(U_i, z_i)_{i \in I}$ is a *coordinate covering* if $X = \bigcup U_i$ and (U_i, z_i) is compatible with (U_j, z_j) for all pairs $(i, j) \in I \times I$. Two coordinate coverings are said to be *equivalent* if their union is also a coordinate covering. This defines an equivalence relation on the set coordinate coverings, and we call an equivalence class of coordinate coverings a *complex structure* on X. A space X together with a complex structure is a *Riemann surface.*

Let $\mathcal{U} = (U_i, z_i)_{i \in I}$ be a coordinate covering of X. A function $f : U \to \mathbb{C}$ on an open subset U of X is said to be *holomorphic* relative to $\mathcal{U}$ if

$$f \circ z^{-1} : z(U \cap U_i) \to \mathbb{C}$$

is holomorphic for all $i \in I$. When $\mathcal{U}'$ is an equivalent coordinate covering, f is holomorphic relative to $\mathcal{U}$ if and only if it is holomorphic relative to $\mathcal{U}'$, and so it makes sense to say that f is holomorphic relative to a complex structure on X: a function $f : U \to \mathbb{C}$ on an open subset U of a Riemann surface X is *holomorphic* if it is holomorphic relative to one (hence every) coordinate covering defining the complex structure on X. Recall that a *meromorphic function* on an open subset U of $\mathbb{C}$ is a holomorphic function on f on $U - \Xi$, for some discrete subset Ξ of U, which has at worst a pole at each point of Ξ (i.e., for each $a \in \Xi$, there exists an m such that $(z - a)^m f(z)$ is holomorphic in some neighbourhood of a). A *meromorphic function* on an open subset U of a Riemann surface is defined exactly the same way.

EXAMPLE 1.4. Any open subset U of $\mathbb{C}$ is a Riemann surface with a single coordinate neighbourhood (U itself, with the identity function z). The holomorphic and meromorphic functions on U with this structure of a Riemann surface are just the usual holomorphic and meromorphic functions.

EXAMPLE 1.5. Let X be the unit sphere

$$S : x^2 + y^2 + z^2 = 1$$

in $\mathbb{R}^3$. Let P be the north pole $(0, 0, 1)$. Stereographic projection from P gives a map

$$(x, y, z) \mapsto \frac{x + iy}{1 - z} : X - P \to \mathbb{C}.$$

Take this to be a coordinate neighbourhood for X. Similarly, stereographic projection from the south pole S gives a second coordinate neighbourhood. These two coordinate neighbourhoods define a complex structure on X, and X together with this complex structure is called the *Riemann sphere.*

EXAMPLE 1.6. Let X be the torus $\mathbb{R}^2/\mathbb{Z}^2$. We shall see that there are infinitely many different complex structures on X.

A map $f : X \to X'$ from one Riemann surface to a second is *holomorphic* if for each point P of X, there are coordinate neighbourhoods (U, z) of P and (U', z') of $f(P)$ such that $z' \circ f \circ z^{-1} : z(U) \to z(U')$ is holomorphic. An *isomorphism* of Riemann surfaces is a bijective holomorphic map whose inverse is also holomorphic.

Riemann surfaces as ringed spaces. Fix a field k. Let X be a topological space, and suppose that for each open subset U of X, we are given a set $\mathcal{O}(U)$ of functions $U \to k$. Then $\mathcal{O}$ is called a *sheaf of k-algebras* on X if

(a) $f, g \in \mathcal{O}(U) \implies f \pm g,\ fg \in \mathcal{O}(U)$; the function $x \mapsto 1$ is in $\mathcal{O}(U)$;
(b) $f \in \mathcal{O}(U),\ V \subset U \implies f|V \in \mathcal{O}(V)$;
(c) let $U = \bigcup U_i$ be an open covering of an open subset U of X, and for each i, let $f_i \in \mathcal{O}(U_i)$; if $f_i|U_i \cap U_j = f_j|U_i \cap U_j$ for all i, j, then there exists an $f \in \mathcal{O}(U)$ such that $f|U_i = f_i$ for all i.

When Y is an open subset of X, we obtain a sheaf of k-algebras $\mathcal{O}|Y$ on Y by restricting the map $U \mapsto \mathcal{O}(U)$ to the open subsets of Y, i.e., for all open $U \subset Y$, define $(\mathcal{O}|Y)(U) = \mathcal{O}(U)$.

From now on, by a *ringed space* we shall mean a pair $(X, \mathcal{O}_X)$ with X a topological space and $\mathcal{O}_X$ a sheaf of $\mathbb{C}$-algebras—we often omit the subscript on $\mathcal{O}$. A *morphism* $\varphi : (X, \mathcal{O}) \to (X', \mathcal{O}')$ *of ringed spaces* is a continuous map $\varphi : X \to X'$ such that, for all open subsets U' of X',

$$f \in \mathcal{O}'(U') \implies f \circ \varphi \in \mathcal{O}(\varphi^{-1}(U')).$$

An *isomorphism* $\varphi : (X, \mathcal{O}) \to (X', \mathcal{O}')$ *of ringed spaces* is a homeomorphism such that φ and φ^{-1} are morphisms. Thus a homeomorphism $\varphi : X \to X'$ is an isomorphism of ringed spaces if, for every U open in X with image U' in X', the map

$$f \mapsto f \circ \varphi : \mathcal{O}'(U') \to \mathcal{O}(U)$$

is bijective.

For example, on any open subset V of the complex plane $\mathbb{C}$, there is a sheaf $\mathcal{O}_V$ with

$$\mathcal{O}_V(U) = \{\text{holomorphic functions } f : U \to \mathbb{C}\},$$

all open $U \subset V$. We call such a pair $(V, \mathcal{O}_V)$ a *standard ringed space*.

The following statements (concerning a Hausdorff topological space X) are all easy to prove.

1.7. Let $\mathcal{U} = (U_i, z_i)$ be a coordinate covering of X, and, for any open subset U of $\mathbb{C}$, let $\mathcal{O}(U)$ be the set of functions $f : U \to C$ that are holomorphic relative to $\mathcal{U}$. Then $U \mapsto \mathcal{O}(U)$ is a sheaf of $\mathbb{C}$-algebras on X.

1.8. Let $\mathcal{U}$ and $\mathcal{U}'$ be coordinate coverings of X; then $\mathcal{U}$ and $\mathcal{U}'$ are equivalent if and only they define the same sheaves of holomorphic functions.

Thus, a complex structure on X defines a sheaf of $\mathbb{C}$-algebras on X, and the sheaf uniquely determines the complex structure.

1.9. A sheaf $\mathcal{O}_X$ of $\mathbb{C}$-algebras on X arises from a complex structure if and only if it satisfies the following condition:

(∗) there is an open covering $X = \bigcup U_i$ of X such that each $(U_i, \mathcal{O}_X|U_i)$ is isomorphic to a standard ringed space.

Thus to give a complex structure on X is the same as to give a sheaf of $\mathbb{C}$-algebras satisfying (∗).

EXAMPLE 1.10. Let $n \in \mathbb{Z}$ act on $\mathbb{C}$ as $z \mapsto z+n$. Topologically, $\mathbb{C}/\mathbb{Z}$ is cylinder. We can give it a complex structure as follows: let $p : \mathbb{C} \to \mathbb{C}/\mathbb{Z}$ be the quotient map; for any point $P \in \mathbb{C}/\mathbb{Z}$, choose a $Q \in f^{-1}(P)$; there exist neighbourhoods U of P and V of Q such that p is a homeomorphism $V \to U$; take any such pair $(U, p^{-1} : U \to V)$ to be a coordinate neighbourhood. The corresponding sheaf of holomorphic functions has the following description: for any open subset U of $\mathbb{C}/\mathbb{Z}$, a function $f : U \to \mathbb{C}$ is holomorphic if and only if $f \circ p$ is holomorphic (check!). Thus the holomorphic functions f on $U \subset \mathbb{C}/\mathbb{Z}$ can be identified with the holomorphic functions on $p^{-1}(U)$ invariant under the action of $\mathbb{Z}$, i.e., such that $f(z+n) = f(z)$ for all $n \in \mathbb{Z}$ (it suffices to check that $f(z+1) = f(z)$, as 1 generates $\mathbb{Z}$ as an abelian group).

For example, $q(z) = e^{2\pi i z}$ defines a holomorphic function on $\mathbb{C}/\mathbb{Z}$. It gives an isomorphism $\mathbb{C}/\mathbb{Z} \to \mathbb{C}^\times$ (complex plane with the origin removed)—in fact, this is an isomorphism of both of Riemann surfaces and of topological groups. The inverse function $\mathbb{C}^\times \to \mathbb{C}/\mathbb{Z}$ is (by definition of log) $(2\pi i)^{-1} \cdot \log$.

Before Riemann (and, unfortunately, also after), mathematicians considered functions only on open subsets of the complex plane $\mathbb{C}$. Thus they were forced to talk about "multi-valued functions" and functions "holomorphic at points at infinity". This works reasonably well for functions of one variable, but collapses into total confusion in several variables. Riemann recognized that the functions were defined in a natural way on spaces that were only locally isomorphic to open subsets of $\mathbb{C}$, that is, on Riemann surfaces, and emphasized the importance of studying these spaces. In this course we follow Riemann—it may have been more natural to call the course "Elliptic Modular Curves" rather than "Modular Functions and Modular Forms".

Differential forms. We adopt a naive approach to differential forms on Riemann surfaces.

A *differential form* on an open subset U of $\mathbb{C}$ is an expression of the form $f(z)dz$ where f is a meromorphic function on U. With any meromorphic function $f(z)$ on U, we associate the differential form $df \stackrel{\text{df}}{=} \frac{df}{dz}dz$. Let $w : U \to U'$ be a mapping from U to another open subset U' of $\mathbb{C}$; we can write it $z' = w(z)$. Let $\omega = f(z')dz'$ be a differential form on U'. Then $w^*(\omega)$ is the differential form $f(w(z))\frac{dw(z)}{dz}dz$ on U.

Let X be a Riemann surface, and let (U_i, z_i) be a coordinate covering of X. To give a *differential form* on X is to give differential forms $\omega_i = f(z_i)dz_i$ on $z_i(U_i)$ for each i that agree on overlaps in the following sense: let $z_i = w_{ij}(z_j)$, so that w_{ij} is the conformal mapping $z_i \circ z_j^{-1} : z_j(U_i \cap U_j) \to z_i(U_i \cap U_j)$; then $w_{ij}^*(\omega_i) = \omega_j$, i.e.,

$$f_j(z_j)dz_j = f_i(w_{ij}(z_j)) \cdot w_{ij}'(z_j)dz_j.$$

Contrast this with functions: to give a meromorphic function f on X is to give meromorphic functions $f_i(z_i)$ on $z_i(U_i)$ for each i that agree on overlaps in the sense that

$$f_j(z_j) = f_i(w_{ij}(z_j)) \text{ on } z_j(U_i \cap U_j).$$

A differential form is said to be of the *first kind* (or holomorphic) if it has no poles on X, of the *second kind* if it has residue 0 at each point of X where it has a pole, and of the *third kind* if it is not of the second kind.

EXAMPLE 1.11. The Riemann sphere S can be thought of as the set of lines through the origin in $\mathbb{C}^2$. Thus a point on S is determined by a point (other than the origin) on the line. In this way, the Riemann sphere is identified with

$$\mathbb{P}^1(\mathbb{C}) = (\mathbb{C} \times \mathbb{C} \setminus \{(0,0)\})/\mathbb{C}^\times.$$

We write $(x_0 : x_1)$ for the equivalence class of (x_0, x_1); thus $(x_0 : x_1) = (cx_0 : cx_1)$ for $c \neq 0$.

Let U_0 be the subset where $x_0 \neq 0$; then $z_0 : (x_0 : x_1) \mapsto x_1/x_0$ is a homeomorphism $U_0 \to \mathbb{C}$. Similarly, if U_1 is the set where $x_1 \neq 0$, then $z_1 : (x_0 : x_1) \mapsto x_0/x_1$ is a homeomorphism $U_1 \to \mathbb{C}$. The pair (U_0, z_0), (U_1, z_1) is a coordinate covering of S. Note that on $U_0 \cap U_1$, z_0 and z_1 are both defined, and $z_1 = z_0^{-1}$; in fact, $z_0(U_0 \cap U_1) = \mathbb{C} \setminus \{0\} = z_1(U_0 \cap U_1)$ and the map $w_{01} : z_1(U_0 \cap U_1) \to z_0(U_0 \cap U_1)$ is $z \mapsto z^{-1}$.

A meromorphic function on S is defined by a meromorphic function $f_0(z_0)$ of $z_0 \in \mathbb{C}$ and a meromorphic function $f_1(z_1)$ of $z_1 \in \mathbb{C}$ such that for $z_0 z_1 \neq 0$, $f_1(z_1) = f_0(z_1^{-1})$. In other words, it is defined by a meromorphic function $f(z)(= f_1(z_1))$ such that $f(z^{-1})$ is also meromorphic on $\mathbb{C}$. (It is automatically meromorphic on $\mathbb{C} \setminus \{0\}$.) In all good complex analysis courses it is shown that the meromorphic functions on S are exactly the rational functions of z, namely, the functions $P(z)/Q(z)$, $P, Q \in \mathbb{C}[X]$, $Q \neq 0$ (see 1.14 below).

A meromorphic differential form on S is defined by a differential form $f_0(z_0)dz_0$ on $\mathbb{C}$ and a differential form $f_1(z_1)dz_1$ on $\mathbb{C}$, such that

$$f_1(z_1) = f_0(z_1^{-1}) \cdot \frac{-1}{z_1^2} \text{ for } z_1 \neq 0.$$

Analysis on compact Riemann surfaces. We merely sketch what we need. For details, see for example R. Gunning, Lectures on Riemann Surfaces, Princeton, 1966, or P. Griffiths, Introduction to Algebraic Curves, AMS, 1989. Note that a Riemann surface X (considered as a topological space) is orientable: each open subset of the complex plane has a natural orientation; hence each coordinate neighbourhood of X has a natural orientation, and these agree on overlaps because conformal mappings preserve orientation. Also note that a holomorphic mapping $f : X \to S$ (the Riemann sphere) can be regarded as a meromorphic function on X, and that all meromorphic functions are of this form. The only functions holomorphic on the whole of a compact Riemann surface are the constant functions.

PROPOSITION 1.12. (a) *A meromorphic function f on a compact Riemann surface has the same number of poles as it has zeros (counting multiplicities).*

(b) *Let ω be a differential form on a compact Riemann surface; then the sum of the residues of ω at its poles is zero.*

PROOF. (Sketch) We first prove (b). Recall that if $\omega = fdz$ is a differential form on an open subset of $\mathbb{C}$ and C is any closed path in $\mathbb{C}$ not passing through any poles of f, then

$$\int_C \omega = 2\pi i(\sum_{\text{poles}} \mathrm{Res}_p \omega)$$

(sum over the poles p enclosed by C). Fix a finite coordinate covering (U_i, z_i) of the Riemann surface, and choose a triangulation of the Riemann surface such that each triangle is completely enclosed in some U_i; then $2\pi i(\sum \mathrm{Res}_p \omega)$ is the sum of the integrals of ω over the various paths, but these cancel out.

Statement (a) is just the special case of (b) in which $\omega = df/f$. □

When we apply (a) to $f - c$, c some fixed number, we obtain the following result.

COROLLARY 1.13. *Let f be a nonconstant meromorphic function on a compact Riemann surface X. Then there is an integer $n > 0$ such that f takes each value exactly n times (counting multiplicities).*

PROOF. The number n is equal to the number of poles of f. □

The integer n is called the *valence* of f. A constant function is said to have valence 0. If f has valence n, then it defines a function $X \to S$ (Riemann sphere) which is n to 1 (counting multiplicities). In fact, there will be only finitely many *ramification points*, i.e., points where it is not exactly n to 1 (when one doesn't count multiplicities).

PROPOSITION 1.14. *Let S be the Riemann sphere. The meromorphic functions are precisely the rational functions of z, i.e., the field of meromorphic functions on S is $\mathbb{C}(z)$.*

PROOF. Let $g(z)$ be a meromorphic function on S. After possibly replacing $g(z)$ with $g(z-c)$, we may suppose that $g(z)$ has neither a zero nor a pole at ∞ (= north pole). Suppose that $g(z)$ has a pole of order m_i at p_i, $i = 1, \ldots, r$, a zero of order n_i at q_i, $i = 1, \ldots, s$, and no other poles or zero. The function

$$g(z)\frac{\prod(z-p_i)^{m_i}}{\prod(z-q_i)^{n_i}}$$

has no zeros or poles at a point $P \neq \infty$, and it has no zero or pole at ∞ because (see 1.12) $\sum m_i = \sum n_i$. It is therefore constant, and so

$$g(z) = \text{constant} \times \frac{\prod(z-q_i)^{n_i}}{\prod(z-p_i)^{m_i}}.$$

□

REMARK 1.15. The proposition shows that the meromorphic functions on S are all algebraic: they are just quotients of polynomials. Thus the field $\mathcal{M}(S)$ of meromorphic functions on S is equal to the field of rational functions on $\mathbb{P}^1$ as defined by algebraic geometry. This is dramatically different from what is true for meromorphic functions on the complex plane. In fact, there exists a vast array of holomorphic functions on $\mathbb{C}$—see Ahlfors for a classification of them (IV.3.3 of the first edition).

PROPOSITION 1.16. *Let f be a nonconstant meromorphic function with valence n on a compact Riemann surface X. Then any meromorphic function g on X is a root of a polynomial of degree $\leq n$ with coefficients in $\mathbb{C}(f)$.*

PROOF. (Sketch) Regard f as a mapping $X \to S$ (Riemann sphere) and let c be a point of S such that $f^{-1}(c)$ has exactly n elements $\{P_1(c), ..., P_n(c)\}$. Let $z \in X$ be such that $f(z) = c$; then

$$0 = \prod_i (g(z) - g(P_i(c))) = g^n(z) + r_1(c)g^{n-1}(z) + \cdots + r_n(c)$$

where the $r_i(c)$ are symmetric functions in the $g(P_i(c))$. When we let c vary (avoiding the c where $f(z) - c$ has multiple zeros), each $r_i(c)$ becomes a meromorphic function on S, and hence is a rational function of $c = f(z)$. □

THEOREM 1.17. *Let X be a compact Riemann surface. There exists a nonconstant meromorphic function f on X, and the set of such functions forms a finitely generated field $\mathcal{M}(X)$ of transcendence degree 1 over $\mathbb{C}$.*

The first statement is the fundamental existence theorem (Gunning, ibid., p107). Its proof is not easy (it is implied by the Riemann-Roch Theorem), but for all the Riemann surfaces in this course, we will be able to write down a nonconstant meromorphic function.

It is obvious that the meromorphic functions on X form a field $\mathcal{M}(X)$. Let f be a nonconstant such function, and let n be its valence. Then (1.16) shows that every other function is algebraic over $\mathbb{C}(f)$, and in fact satisfies a polynomial of degree $\leq n$. Therefore $\mathcal{M}(f)$ has degree $\leq n$ over $\mathbb{C}(f)$, because if it had degree $> n$ then it would contain a subfield L of finite degree $n' > n$ over $\mathbb{C}(f)$, and the Primitive Element Theorem (Math 594f, 5.1) tells us that then $L = \mathbb{C}(f)(g)$ for some g whose minimum polynomial has degree n'.

EXAMPLE 1.18. Let S be the Riemann sphere. For any meromorphic function f on S with valence 1, $\mathcal{M}(S) = \mathbb{C}(f)$.

REMARK 1.19. The meromorphic functions on a compact complex manifold X of dimension $m > 1$ again form a field that is finitely generated over $\mathbb{C}$, but its transcendence degree may be $< m$. For example, there are compact complex manifolds of dimension 2 with no nonconstant meromorphic functions.

Riemann-Roch Theorem. The Riemann-Roch theorem describes how many functions there are on a compact Riemann surface with given poles and zeros.

Let X be a compact Riemann surface. The *group of divisors* $Div(X)$ on X is the free (additive) abelian group generated by the points on X; thus an element of $Div(X)$ is a finite sum $\sum n_i P_i$, $n_i \in \mathbb{Z}$. A divisor $D = \sum n_i P_i$ is *positive* (or *effective*) if every $n_i \geq 0$; we then write $D \geq 0$.

Let f be a nonzero meromorphic function on X. For any point $P \in X$, let $\mathrm{ord}_P(f) = m$, $-m$, or 0 according as f has a zero of order m at P, a pole of order m at P, or neither a pole nor a zero at P. The *divisor* of f is

$$div(f) = \sum \mathrm{ord}_p(f) \cdot P.$$

This is a finite sum because the zeros and poles of f form discrete sets, and we are assuming X to be compact.

The map $f \mapsto div(f) : \mathbb{C}(X)^{\times} \to Div(X)$ is a homomorphism, and its image is called the group of *principal divisors*. Two divisors are said to be *linearly equivalent* if their difference is principal. The *degree* of a divisor $\sum n_i P_i$ is $\sum n_i$. The map $D \mapsto \deg(D)$ is a homomorphism $Div(X) \to \mathbb{Z}$ whose kernel contains the principal divisors. Thus it makes sense to speak of the *degree* of a linear equivalence class of divisors.

It is possible to attach a divisor to a differential form ω: let $P \in X$, and let (U_i, z_i) be a coordinate neighbourhood containing P; the differential form ω is described by a differential $f_i dz_i$ on U_i, and we set $ord_p(\omega) = ord_p(f_i)$. Then $ord_p(\omega)$ is independent of the choice of the coordinate neighbourhood U_i (because w_{ij} and its derivative have no zeros or poles), and we define

$$div(\omega) = \sum ord_p(\omega) \cdot P.$$

Again, this is a finite sum. Note that, for any meromorphic function f,

$$div(f\omega) = div(f) + div(\omega).$$

If ω is one nonzero differential form, then any other is of the form $f\omega$ for some $f \in \mathcal{M}(X)$, and so the linear equivalence class of $div(\omega)$ is independent of ω; we write K for $div(\omega)$, and $\mathfrak{k}$ for its linear equivalence class.

For a divisor D, define

$$L(D) = \{f \in \mathcal{M}(X) \mid div(f) + D \geq 0\} \cup \{0\}.$$

This is a vector space over $\mathbb{C}$, and if $D' = D + (g)$, then $f \mapsto fg^{-1}$ is an isomorphism $L(D) \to L(D')$. Thus the dimension $\ell(D)$ of $L(D)$ depends only on the linear equivalence class of D.

THEOREM 1.20 (Riemann-Roch). *Let X be a compact Riemann surface. Then there is an integer $g \geq 0$ such that for any divisor D,*

$$\ell(D) = \deg(D) + 1 - g + \ell(K - D).$$

PROOF. See Gunning 1962, §7, or Griffiths 1989, for a proof in the context of Riemann surfaces, and Fulton 1969, Chapter 8, for a proof in the context of algebraic curves. One approach to proving it is to verify it first for the Riemann sphere S (see below), and then to regard X as a finite covering of S. □

Note that in the statement of the Riemann-Roch Theorem, we could replace the divisors with equivalence classes of divisors.

COROLLARY 1.21. *A canonical divisor K has degree $2g - 2$, and $\ell(K) = g$.*

PROOF. Put $D = 0$ in (1.20). The only functions with $div(f) \geq 0$ are the constant functions, and so the equation becomes $1 = 0 + 1 - g + \ell(K)$. Hence $\ell(K) = g$. Put $D = K$; then the equation becomes $g = \deg(K) + 1 - g + 1$, which gives $\deg(K) = 2g - 2$. □

Let $K = div(\omega)$. Then $f \mapsto f\omega$ is an isomorphism from $L(D)$ to the space of holomorphic differential forms on X, which therefore has dimension g.

The term in the Riemann-Roch formula that is difficult to evaluate is $\ell(K-D)$. Thus it is useful to note that if $\deg(D) > 2g-2$, then $L(K-D) = 0$ (because, for $f \in \mathcal{M}(X)^{\times}$, $\deg(D) > 2g-2 \implies \deg(div(f)+K-D) < 0$, and so $div(f)+K-D$ can't be a positive divisor). Hence:

COROLLARY 1.22. *If* $\deg(D) > 2g-2$, *then* $\ell(D) = \deg(D)+1-g$.

EXAMPLE 1.23. Let X be the Riemann sphere, and let $D = mP_\infty$, where P_∞ is the "point at infinity". Then $L(D)$ is the space of meromorphic functions on $\mathbb{C}$ with at worst a pole of order m at infinity and no poles elsewhere. These functions are the polynomials of degree $\leq m$, and they form a vector space of $m+1$, in other words,

$$\ell(D) = \deg(D) + 1,$$

and so the Riemann-Roch theorem shows that $g = 0$. Consider the differential dz on $\mathbb{C}$, and let $z' = 1/z$. The $dz = -1/z'^2 dz'$, and so dz extends to a meromorphic differential on X with a pole of order 2 at ∞. Thus $\deg(div(\omega)) = -2$, in agreement with the above formulas.

EXERCISE 1.24. Prove (1.20) for the Riemann sphere. (Hint: use partial fractions.)

The genus of X. Let X be a compact Riemann surface. It can be regarded as a topological space, and so we can define homology groups $H_0(X,\mathbb{Q})$, $H_1(X,\mathbb{Q})$, $H_2(X,\mathbb{Q})$. It is known that H_0 and H_2 each have dimension 1, and H_1 has dimension $2g$. It is a theorem that this g is the same as that occurring in the Riemann-Roch theorem (see below). Hence g depends only on X as a topological space, and not on its complex structure. The Euler-Poincaré characteristic of X is

$$\chi(X) \stackrel{\text{df}}{=} \dim H_0 - \dim H_1 + \dim H_2 = 2 - 2g.$$

Since X is oriented, it can be triangulated. When one chooses a triangulation, then one finds (easily) that

$$2 - 2g = V - E + F,$$

where V is the number of vertices, E is the number of edges, and F is the number of faces.

EXAMPLE 1.25. Triangulate the sphere by projecting out from a regular tetrahedron whose vertices are on the sphere. Then $V = 4$, $E = 6$, $F = 4$, and so $g = 0$.

EXAMPLE 1.26. Consider the map $z \mapsto z^e : D \to D$, where D is the unit open disk. This map is exactly $e:1$ except at the origin, which is a ramification point of order e. Consider the differential dz' on D. The map is $z' = w(z) = z^e$, and so the inverse image of the differential dz' is $dz' = dw(z) = ez^{e-1}dz$. Thus $w^*(dz')$ has a zero of order $e-1$ at 0.

THEOREM 1.27 (Riemann-Hurwitz Formula). *Let* $f : Y \to X$ *be a holomorphic mapping of compact Riemann surfaces that is* $m:1$ *(except over finitely many points). For each point* P *of* X, *let* e_P *be the multiplicity of* P *in the fibre of* f; *then*

$$2g(Y) - 2 = (2g(X)-2)m + \sum (e_P - 1).$$

PROOF. Choose a differential ω on X such that ω has no pole or zero at a ramification point of X. Then $f^*\omega$ has a pole and a zero above each pole and zero of ω (of the same order as that of ω); in addition it has a zero of order $e-1$ at each ramification point in Y (cf. the above example). Thus

$$\deg(f^*\omega) = m\deg(\omega) + \sum(e_P - 1),$$

and we can apply (1.21). □

REMARK 1.28. One can also prove this formula topologically. Triangulate X in such a way that each ramification point is a vertex for the triangulation, and pull the triangulation back to Y. There are the following formulas for the numbers of faces, edges, and vertices for the triangulations of Y and X :

$$F(Y) = m \cdot F(X), \quad E(Y) = m \cdot E(X), \quad V(Y) = m \cdot V(X) - \sum(e_P - 1).$$

Thus

$$2 - 2g(Y) = (2 - 2g(X)) - \sum(e_P - 1),$$

in agreement with (1.27).

We have verified that the two notions of genus agree for the Riemann sphere S (they both give 0). But for any Riemann surface X, there is a nonconstant function $f : X \to S$ (by 1.17) and we have just observed that the formulas relating the genus of X to that of S is the same for the two notions of genus, and so we have shown that the two notions give the same value for X.

Riemann surfaces as algebraic curves. Let X be a compact Riemann surface. Then (see 1.17) $\mathcal{M}(X)$ is a finitely generated field of transcendence degree 1 over $\mathbb{C}$, and so there exist meromorphic functions f and g on X such that $\mathcal{M}(X) = \mathbb{C}(f, g)$. There is a nonzero irreducible polynomial $\Phi(X, Y)$ such that

$$\Phi(f, g) = 0.$$

Then $z \mapsto (f(z), g(z)) : X \to \mathbb{C}^2$ maps an open subset of X onto an open subset of the algebraic curve defined by the equation:

$$\Phi(X, Y) = 0.$$

Unfortunately, this algebraic curve will in general have singularities. A better approach is the following. Suppose that the genus of X is ≥ 2 (and X is not hyperelliptic), and choose a basis $\omega_0, ..., \omega_n$, $(n = g-1)$ for the space of holomorphic differential forms on X. For $P \in X$, we can represent each ω_i in the form $f_i \cdot dz$ in some neighbourhood of P. After possibly replacing each ω_i with $f\omega_i$, f a meromorphic function defined near P, the f_i's will all be defined at P, and at least one will be nonzero at P. Thus $(f_0(P) : \ldots : f_n(P))$ is a well-defined point of $\mathbb{P}^n(\mathbb{C})$, independent of the choice of f. It is known that the map φ

$$P \mapsto (f_0(P) : ... : f_n(P)) : X \to \mathbb{P}^n(\mathbb{C})$$

is a homeomorphism of X onto a closed subset of $\mathbb{P}^n(\mathbb{C})$, and that there is a finite set of homogeneous polynomials in $n+1$ variables whose zero set is precisely $\varphi(X)$. Moreover, the image is a nonsingular curve in $\mathbb{P}^n(\mathbb{C})$ (Griffiths 1989, IV.3). If X has genus < 2, or is hyperelliptic, a slight modification of this method again realizes X as an algebraic curve in $\mathbb{P}^n$ for some n.

2. Elliptic Modular Curves as Riemann Surfaces

In this section, we define the Riemann surfaces $Y(N) = \Gamma(N)\backslash\mathbb{H}$ and their natural compactifications, $X(N)$. Recall that $\mathbb{H}$ is the complex upper half plane

$$\mathbb{H} = \{z \in \mathbb{C} \mid \Im(z) > 0\}.$$

The upper-half plane as a quotient of $\mathrm{SL}_2(\mathbb{R})$. We saw in the Introduction that there is an action of $\mathrm{SL}_2(\mathbb{R})$ on $\mathbb{H}$ as follows:

$$\mathrm{SL}_2(\mathbb{R}) \times \mathbb{H} \to \mathbb{H}, \quad (\alpha, z) \mapsto \alpha(z) = \frac{az+b}{cz+d}, \quad \alpha = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Because $\Im(\alpha z) = \Im(z)/|cz+d|^2$, $\Im(z) > 0 \implies \Im(\alpha z) > 0$. When we give $\mathrm{SL}_2(\mathbb{R})$ and $\mathbb{H}$ their natural topologies, this action is continuous.

The *special orthogonal group* (or "circle group") is defined to be

$$\mathrm{SO}_2(\mathbb{R}) = \left\{ \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \mid \theta \in \mathbb{R} \right\}.$$

Note that $\mathrm{SO}_2(\mathbb{R})$ is a closed subgroup of $\mathrm{SL}_2(\mathbb{R})$, and so $\mathrm{SL}_2(\mathbb{R})/\mathrm{SO}_2(\mathbb{R})$ is a Hausdorff topological space (by 1.1).

Proposition 2.1. (a) *The group $\mathrm{SL}_2(\mathbb{R})$ acts transitively on $\mathbb{H}$, i.e., for any elements $z, z' \in \mathbb{H}$, there exists an $\alpha \in \mathrm{SL}_2(\mathbb{R})$ such that $\alpha z = z'$.*

(b) *The action of $\mathrm{SL}_2(\mathbb{R})$ on $\mathbb{H}$ induces an isomorphism*

$$\mathrm{SL}_2(\mathbb{R})/\{\pm I\} \to \mathrm{Aut}(\mathbb{H}) \text{ (biholomorphic automorphisms of } \mathbb{H})$$

(c) *The stabilizer of i is $\mathrm{SO}_2(\mathbb{R})$.*

(d) *The map*

$$\mathrm{SL}_2(\mathbb{R})/\mathrm{SO}_2(R) \to H, \quad \alpha \cdot \mathrm{SO}_2(R) \mapsto \alpha(i)$$

is a homeomorphism.

Proof. (a) Let $z \in \mathbb{H}$; it suffices to show that there exists an $\alpha \in \mathrm{SL}_2(\mathbb{R})$ such that $\alpha(i) = z$—if z' is a second point, then $\alpha'(i) = z'$ for some $\alpha' \in \mathrm{SL}_2(\mathbb{R})$, and $\alpha'\alpha^{-1}(z) = z'$. Write $z = x + iy$; then $\alpha \overset{\text{df}}{=} \sqrt{y}^{-1} \begin{pmatrix} y & x \\ 0 & 1 \end{pmatrix} \in$
$\mathrm{SL}_2(\mathbb{R})$, and $\alpha(i) = z$.

(b) If $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot z = z$ then $cz^2 + (d-a)z - b = 0$. If this is true for all $z \in \mathbb{H}$ (any three z's would do), then the polynomial must have zero coefficients, and so $c = 0$, $d = a$, and $b = 0$. Thus $\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}$, and this has determinant 1 if and only if $a = \pm 1$. Thus only $\pm I$ act trivially on $\mathbb{H}$.

Let γ be an automorphism $\mathbb{H}$. We know from (a) that there is an $\alpha \in \mathrm{SL}_2(\mathbb{R})$ such that $\alpha(i) = \gamma(i)$. After replacing γ with $\alpha^{-1} \circ \gamma$, we can assume that $\gamma(i) = i$. Recall that the map $\rho : \mathbb{H} \to D$, $z \mapsto \frac{z-i}{z+i}$ is an isomorphism from $\mathbb{H}$ onto the open unit disk, and it maps i to 0. Use ρ to transfer γ into an automorphism γ' of D fixing 0. □

LEMMA 2.2. *The automorphisms of D fixing* 0 *are the maps of the form $z \mapsto \lambda z$, $|\lambda| = 1$.*

PROOF. This is an easy consequence of the Schwarz Lemma (Cartan 1963, III.3), which says the following:

Let $f(z)$ be a holomorphic function on the disk $|z| < 1$ and suppose that

$$f(0) = 0, \quad |f(z)| < 1 \text{ for } |z| < 1.$$

Then
(i) $|f(z)| \leq |z|$ for $|z| < 1$;
(ii) if $|f(z_0)| = |z_0|$ for some $z_0 \neq 0$, then there is a λ such that $f(z) = \lambda z$ (and $|\lambda| = 1$).

Let γ be an automorphism of D fixing 0. When we apply (i) to γ and γ^{-1}, we find that $|\gamma(z)| = |z|$ for all z in the disk, and so we can apply (ii) to find that f is of the required form.

The lemma tells us that there is a $\theta \in \mathbb{R}$ such that $\rho \circ \gamma \circ \rho^{-1}(z) = e^{2\theta i} \cdot z$ for all z, and the following exercise shows that $\gamma(z) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \cdot z$. Thus $\gamma \in \mathrm{SO}_2(\mathbb{R}) \subset \mathrm{SL}_2(R)$.

(c) We have already proved this, but it is easy to give a direct proof. We have

$$\frac{ai+b}{ci+d} = i \iff ai + b = -c + di \iff a = d, \quad b = -c.$$

Therefore the matrix is of the form $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ with $a^2 + b^2 = 1$, and so is in $\mathrm{SO}_2(\mathbb{R})$.

(d) This is a consequence of the general result (1.2). □

EXERCISE 2.3. Let $\psi : \mathbb{C}^2 \times \mathbb{C}^2 \to \mathbb{C}$ be the Hermitian form

$$\left(\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \right) \mapsto \bar{z}_1 w_1 - z_2 \bar{w}_2.$$

and let $\mathrm{SU}(1,1)$ (*special unitary group*) be the subgroup of elements $\alpha \in \mathrm{SL}_2(\mathbb{C})$ such that $\psi(\alpha(z), \alpha(w)) = \psi(z, w)$.

(a) Show that $\mathrm{SU}(1,1) = \left\{ \begin{pmatrix} u & v \\ \bar{v} & \bar{u} \end{pmatrix} \mid u, v \in \mathbb{C}, \quad |u|^2 - |v|^2 = 1 \right\}$.
(b) Define an action of $\mathrm{SU}(1,1)$ on the unit disk as follows:

$$\begin{pmatrix} u & v \\ \bar{v} & \bar{u} \end{pmatrix} \cdot z = \frac{uz + v}{\bar{v}z + \bar{u}}.$$

Show that this defines an isomorphism $\mathrm{SU}(1,1)/\{\pm I\} \to \mathrm{Aut}(D)$.
(c) Show that, under the standard isomorphism $\rho : \mathbb{H} \to D$, the action of the element $\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$ of $\mathrm{SL}_2(\mathbb{R})$ on $\mathbb{H}$ corresponds to the action of $\begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix}$ on D.

Quotients of $\mathbb{H}$. Let Γ be a group acting on a topological space X. If $\Gamma\backslash X$ is Hausdorff, then the orbits are closed, but this condition is not sufficient to ensure that the quotient space is Hausdorff. The action is said to be *discontinuous* if for every $x \in X$ and infinite sequence (γ_i) of distinct elements of Γ, the set $\{\gamma_i x\}$ has no cluster point; it is said to be *properly discontinuous* if, for any pair of points x and y of X, there exist neighbourhoods U_x and U_y of x and y such that the set $\{\gamma \in \Gamma \mid \gamma U_x \cap U_y \neq \emptyset\}$ is finite.

PROPOSITION 2.4. *Let G be a locally compact group acting on a topological space X such that for one (hence every) point $x_0 \in X$, the stabilizer K of x_0 in G is compact and $gK \mapsto gx_0 : G/K \to X$ is a homeomorphism. The following conditions on a subgroup Γ of G are equivalent:*

(a) *Γ acts discontinuously on X;*
(b) *Γ acts properly discontinuously on X;*
(c) *for any compact subsets A and B of X, $\{\gamma \in \Gamma \mid \gamma A \cap B \neq \emptyset\}$ is finite;*
(d) *Γ is a discrete subgroup of G.*

PROOF. (d) $\implies$ (c) (This is the only implication we shall use.) Write p for the map, $gK \mapsto gx_0 : G \to X$. Let A be a compact subset of X. I claim that $p^{-1}(A)$ is compact. Write $G = \bigcup V_i$ where the V_i are open with compact closures $\bar{V}_i$. Then $A \subset \bigcup p(V_i)$, and in fact we need only finitely many $p(V_i)$'sto cover A. Then $p^{-1}(A) \subset \bigcup V_i K \subset \bigcup \bar{V}_i K$ (finite union), and each $\bar{V}_i K$ is compact (it is the image of $\bar{V}_i \times K$ under the multiplication map $G \times G \to G$). Thus $p^{-1}(A)$ is a closed subset of a compact set, and so is compact. Similarly, $p^{-1}(B)$ is compact.

Suppose $\gamma A \cap B \neq \emptyset$ and $\gamma \in \Gamma$. Then $\gamma(p^{-1}A) \cap p^{-1}B \neq \emptyset$, and so $\gamma \in \Gamma \cap (p^{-1}B) \cdot (p^{-1}A)^{-1}$. But this last set is the intersection of a discrete set with a compact set and so is finite.

(The implications (c) $\implies$ (b) $\implies$ (a) are trivial, and (b) $\implies$ (c) is easy. For (c) $\implies$ (d), let V be any neighbourhood of 1 in G whose closure $\bar{V}$ is compact. For any $x \in X$, $\Gamma \cap V \subset \{\gamma \in \Gamma \mid \gamma x \in \bar{V} \cdot x\}$, which is finite, because both $\{x\}$ and $\bar{V} \cdot x$ are compact. Thus $\Gamma \cap V$ is discrete, which shows that e is an isolated point of Γ.) □

The next result makes statement (c) more precise.

PROPOSITION 2.5. *Let G, K, X be as in (2.4), and let Γ be a discrete subgroup of G.*

(a) *For any $x \in X$, $\{g \in \Gamma \mid gx = x\}$ is finite.*
(b) *For any $x \in X$, there is a neighbourhood U of x with the following property: if $\gamma \in \Gamma$ and $U \cap \gamma U \neq \emptyset$, then $\gamma x = x$.*
(c) *For any points x and $y \in X$ that are not in the same Γ-orbit, there exist neighbourhoods U of x and V of y such that $\gamma U \cap V = \emptyset$ for all $\gamma \in \Gamma$.*

PROOF. (a) We saw in the proof of (2.4) that $p^{-1}(\text{compact})$ is compact, where $p(g) = gx$. Therefore $p^{-1}(x)$ is compact, and the set we are interested in is $p^{-1}(x) \cap \Gamma$.

(b) Let V be a compact neighbourhood of x. Because (2.4c) holds, there is a finite set $\{\gamma_1, ..., \gamma_n\}$ of elements of Γ such that $V \cap \gamma_i V \neq \emptyset$. Let $\gamma_1, ..., \gamma_s$ be the γ_i's fixing

x. For each $i > s$, choose disjoint neighbourhoods V_i of x and W_i of $\gamma_i x$, and put

$$U = V \cap (\cap_{i>s} V_i \cap \gamma_i^{-1} W_i).$$

For $i > s$, $\gamma_i U \subset W_i$ which is disjoint from V_i, which contains U.

(c) Choose compact neighbourhoods A of x and B of y, and let $\gamma_1, ..., \gamma_n$ be the elements of Γ such that $\gamma_i A \cap B \neq \emptyset$. We know $\gamma_i x \neq y$, and so we can find disjoint neighbourhoods U_i and V_i of $\gamma_i x$ and y. Take

$$U = A \cap \gamma_1^{-1} U_1 \cap ... \cap \gamma_n^{-1} U_n, \quad V = B \cap V_1 \cap ... \cap V_n.$$

□

COROLLARY 2.6. *Under the hypotheses of (2.5), the space* $\Gamma \backslash X$ *is Hausdorff.*

PROOF. Let x and y be points of X not in the same Γ-orbit, and choose neighbourhoods U and V as in (2.5). Then the images of U and V in $\Gamma \backslash X$ are disjoint neighbourhoods of Γx and Γy. □

A group Γ is said to act *freely* on a set X if $\mathrm{Stab}(x) = e$ for all $x \in X$.

PROPOSITION 2.7. *Let* Γ *be a discrete subgroup of* $\mathrm{SL}_2(\mathbb{R})$ *such that* Γ *(or* $\Gamma/\{\pm I\}$ *if* $-I \in \Gamma$*) acts freely on* $\mathbb{H}$. *Then there is a unique complex structure on* $\Gamma \backslash \mathbb{H}$ *with the following property: a function* f *on an open subset* U *of* $\Gamma \backslash \mathbb{H}$ *is holomorphic if and only if* $f \circ p$ *is holomorphic.*

PROOF. The uniqueness follows from the fact (see 1.8) that the sheaf of holomorphic functions on a Riemann surface determines the complex structure. Let $z \in \Gamma \backslash \mathbb{H}$, and choose an $x \in p^{-1}(z)$. According to (2.5b), there is a neighbourhood U of x such that γU is disjoint from U for all $\gamma \in \Gamma$, $\gamma \neq e$. The map $p|U : U \to p(U)$ is a homeomorphism, and we take all pairs of the form $(p(U), (p|U)^{-1})$ to be coordinate neighbourhoods. It is easy to check that they are all compatible, and that the holomorphic functions are as described. (Alternatively, one can define $\mathcal{O}(\mathcal{U})$ as in the statement of the proposition, and verify that $U \mapsto \mathcal{O}(\mathcal{U})$ is a sheaf of $\mathbb{C}$-algebras satisfying (1.9(*).) □

Unfortunately $\mathrm{SL}_2(\mathbb{Z})/\{\pm I\}$ doesn't act freely.

Discrete subgroups of $\mathrm{SL}_2(\mathbb{R})$**.** To check that a subgroup Γ of $\mathrm{SL}_2(\mathbb{R})$ is discrete, it suffices to check that e is isolated in Γ. A discrete subgroup of $\mathrm{SL}_2(\mathbb{R})$ (or $\mathrm{SL}_2(\mathbb{R})$) is called a *Fuchsian group.* Discrete subgroups of $\mathrm{SL}_2(\mathbb{R})$ abound, but those of interest to number theorists are rather special.

Congruence subgroups of the elliptic modular group. Clearly $\mathrm{SL}_2(\mathbb{Z})$ is discrete, and *a fortiori*, $\Gamma(N)$ is discrete. A *congruence subgroup* of $\mathrm{SL}_2(\mathbb{Z})$ is a subgroup containing $\Gamma(N)$ for some N. For example,

$$\Gamma_0(N) \stackrel{\mathrm{df}}{=} \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \mid c \equiv 0 \mod N) \right\}$$

is a congruence subgroup of $\mathrm{SL}_2(\mathbb{Z})$. By definition, the sequence

$$1 \to \Gamma(N) \to \mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$$

is exact ($\mathrm{SL}_2(A)$ makes sense for any commutative ring—it is the group of 2×2 matrices with coefficients in A having determinant 1). I claim that the map $\mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(Z/NZ)$ is surjective. To prove this, we have to show that if $A \in M_2(\mathbb{Z})$ and $\det(A) \equiv 1 \mod N$, then there is a $B \in M_2(\mathbb{Z})$ such that $B \equiv A \mod N$ and $\det(B) = 1$. Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$; the condition on A is that

$$ab - cd - Nm = 1$$

for some $m \in \mathbb{Z}$. Hence $\gcd(c, d, N) = 1$, and we can find an integer n such that $\gcd(c, d + nN) = 1$ (apply the Chinese Remainder Theorem to find an n such that $d+nN \equiv 1 \mod p$ for every prime p dividing c but not dividing N and $n \equiv 0 \mod p$ for every prime p dividing both c and N). We can replace d with $d + nN$, and so assume that $gcd(c, d) = 1$. Consider the matrix

$$B = \begin{pmatrix} a + eN & b + fN \\ c & d \end{pmatrix}$$

for some integers e, f. Its determinant is $ad - bc + N(ed - fc) = 1 + (m + ed - fc)N$. Since $\gcd(c, d) = 1$, there exist integers e, f such that $m = fc + ed$, and with this choice, B is the required matrix.

Note that the surjectivity of $\mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$ is implies that $\mathrm{SL}_2(\mathbb{Z})$ is dense in $\mathrm{SL}_2(\hat{\mathbb{Z}})$, where $\hat{\mathbb{Z}} = \varprojlim_N \mathbb{Z}/NZ =$completion of $\mathbb{Z}$ for the topology of subgroups of finite index $= \prod \mathbb{Z}_\ell$.

Discrete groups coming from quaternion algebras. For any rational numbers a, b, let $B = B_{a,b}$ be the $\mathbb{Q}$-algebra with basis $\{1, i, j, k\}$ and multiplication given by

$$i^2 = a,\ j^2 = b,\ ij = k = -ji.$$

Then $B \otimes \mathbb{R}$ is an algebra over $\mathbb{R}$ with the same basis and multiplication table, and it is isomorphic either to $M_2(\mathbb{R})$ or the usual (Hamiltonian) quaternion algebra—we suppose the former.

For $\alpha = c + di + ej + fk \in B$, define $\mathrm{Nm}(\alpha) = c^2 - d^2 - e^2 - f^2 \in \mathbb{Q}$. Under the isomorphism $B \otimes \mathbb{R} \to M_2(\mathbb{R})$, the norm corresponds to the determinant, and so the isomorphism induces an isomorphism

$$\{\alpha \in B \otimes \mathbb{R} \mid \mathrm{Nm}(\alpha) = 1\} \xrightarrow{\approx} \mathrm{SL}_2(\mathbb{R}).$$

An *order* in B is a subring $\mathcal{O}$ that is finitely generated over $\mathbb{Z}$ (hence free of rank 4). Define

$$\Gamma_{a,b} = \{\alpha \in \mathcal{O} \mid \mathrm{Nm}(\alpha) = 1\}.$$

Under the above isomorphism this is mapped to a discrete subgroup of $\mathrm{SL}_2(\mathbb{R})$, and we can define congruence subgroups of $\Gamma_{a,b}$ as for $\mathrm{SL}_2(\mathbb{Z})$.

For a suitable choice of (a, b), $B = M_2(\mathbb{Q})$ (ring of 2×2 matrices with coefficients in $\mathbb{Q}$), and if we choose $\mathcal{O}$ to be $M_2(\mathbb{Z})$, then we recover the elliptic modular groups.

If B is not isomorphic to $M_2(\mathbb{Q})$, then the families of discrete groups that we get are quite different from the congruence subgroups of $\mathrm{SL}_2(\mathbb{Z})$: they have the property that $\Gamma\backslash\mathbb{H}$ is compact.

There are infinitely many nonisomorphic quaternion algebras over $\mathbb{Q}$, and so the congruence subgroups of $\mathrm{SL}_2(\mathbb{Z})$ form just one among an infinite sequence of families of discrete subgroups of $\mathrm{SL}_2(\mathbb{R})$.

[These groups were found by Poincaré in the 1880's, but he regarded them as automorphism groups of the quadratic forms $\Phi_{a,b} = -aX^2 - bY^2 + abZ^2$. For a description of how he found them, see p52, of his book, Science and Method.]

EXERCISE 2.8. Two subgroups Γ and Γ' of a group are said to be *commensurable* if $\Gamma \cap \Gamma'$ is of finite index in both Γ and Γ'.

(a) Commensurability is an equivalence relation (only transitivity is nonobvious).
(b) If Γ and Γ' are commensurable subgroups of a topological group G, and Γ is discrete, then so also is Γ'.
(c) If Γ and Γ' are commensurable subgroups of $\mathrm{SL}_2(\mathbb{R})$ and $\Gamma\backslash\mathbb{H}$ is compact, so also is $\Gamma'\backslash\mathbb{H}$.

Arithmetic subgroups of the elliptic modular group. A subgroup of $\mathrm{SL}_2(\mathbb{Q})$ is *arithmetic* if it is commensurable with $\mathrm{SL}_2(\mathbb{Z})$. For example, every subgroup of finite index in $\mathrm{SL}_2(\mathbb{Z})$, hence every congruence subgroup, is arithmetic. The congruence subgroups are sparse among the arithmetic subgroups: if we let $N(m)$ be the number of congruence subgroups of $\mathrm{SL}_2(\mathbb{Z})$ of index $< m$, and let $N'(m)$ be the number of subgroups of index $< m$, then $N(m)/N'(m) \to 0$ as $m \to \infty$.

REMARK 2.9. This course will be concerned with quotients of $\mathbb{H}$ by congruence groups in the elliptic modular group $\mathrm{SL}_2(\mathbb{Z})$, although the congruence groups arising from quaternion algebras are of (almost) equal interest to number theorists. There is some tantalizing evidence that modular forms relative to other arithmetic groups may also have interesting arithmetic properties, but we shall ignore this.

There are many nonarithmetic discrete subgroups of $\mathrm{SL}_2(\mathbb{R})$. The ones of most interest (to analysts) are those of the "first kind"—they are "large" in the sense that $\Gamma\backslash\mathrm{SL}_2(\mathbb{R})$ (hence $\Gamma\backslash\mathbb{H}$) has finite volume relative to a Haar measure on $\mathrm{SL}_2(\mathbb{R})$.

Among matrix groups, SL_2 is anomalous in having so many discrete subgroups. For other groups there is a wonderful theorem of Margulis (for which he got the Fields medal), which says that, under some mild hypotheses (which exclude SL_2), any discrete subgroup Γ of $G(\mathbb{R})$ such that $\Gamma\backslash G(\mathbb{R})$ has finite measure is arithmetic, and for many groups one knows that all arithmetic subgroups are congruence (see Prasad's talk at the International Congress in Kyoto).

Classification of linear fractional transformations. The group $\mathrm{SL}_2(\mathbb{C})$ acts on $\mathbb{C}^2$, and hence on the set $\mathbb{P}^1(C)$ of lines through the origin in $\mathbb{C}^2$. When we identify a line with its slope, $\mathbb{P}^1(\mathbb{C})$ becomes identified with $\mathbb{C}\cup\{\infty\}$, and we get an action of $\mathrm{GL}_2(\mathbb{C})$ on $\mathbb{C}\cup\{\infty\}$:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} z = \frac{az+b}{cz+d}, \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \infty = \frac{a}{c}.$$

These mappings are called the *linear fractional transformations* of $\mathbb{P}^1(\mathbb{C}) = \mathbb{C}\cup\{\infty\}$. They map circles and lines in $\mathbb{C}$ into circles or lines in $\mathbb{C}$. The scalar matrices $\begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}$

act as the identity transformation. By the theory of Jordan canonical forms, any nonscalar α is conjugate to a matrix of the following type:

$$\text{(i)} \quad \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} \quad \text{(ii)} \quad \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}, \quad \lambda \neq \mu.$$

according as it has repeated eigenvalues or distinct eigenvalues. In the first case, α is conjugate to a transformation $z \mapsto z + \lambda^{-1}$, and in the second to $z \mapsto cz$, $c \neq 1$. In case (i), α is called *parabolic,* and case (ii), it is called *elliptic* if $|c| = 1$, *hyperbolic* if c is real and positive, and *loxodromic* otherwise.

When $\alpha \in \mathrm{SL}_2(\mathbb{C})$, the four cases can be distinguished by the trace of α :

$$\begin{aligned} \alpha \text{ is parabolic} &\iff \mathrm{Tr}(\alpha) = \pm 2; \\ \alpha \text{ is elliptic} &\iff \mathrm{Tr}(\alpha) \text{ is real and } |\mathrm{Tr}(\alpha)| < 2; \\ \alpha \text{ is hyperbolic} &\iff \mathrm{Tr}(\alpha) \text{ is real and } |\mathrm{Tr}(\alpha)| > 2; \\ \alpha \text{ is loxodromic} &\iff \mathrm{Tr}(\alpha) \text{ is not real.} \end{aligned}$$

We now investigate the elements of these types in $\mathrm{SL}_2(\mathbb{R})$.

Parabolic transformations. Suppose $\alpha \in \mathrm{SL}_2(\mathbb{R})$, $\alpha \neq \pm I$, is parabolic. Then it has exactly one eigenvector, and that eigenvector is real. Suppose the eigenvector is $\begin{pmatrix} e \\ f \end{pmatrix}$; if $f \neq 0$, then α has a fixed point in $\mathbb{R}$; if $f = 0$, then ∞ is a fixed point (the transformation is then of the form $z \mapsto z + c$). Thus α has exactly one fixed point in $\mathbb{R} \cup \{\infty\}$.

Elliptic transformations. Suppose $\alpha \in \mathrm{SL}_2(\mathbb{R})$, $\alpha \neq \pm I$, is elliptic. Its characteristic polynomial is $X^2 + bX + 1$ with $|b| < 2$; hence $\Delta = b^2 - 4 < 0$, and so α has two complex conjugate eigenvectors. Thus α has exactly one fixed point z in $\mathbb{H}$ and a second fixed point, namely, $\bar{z}$, in the lower half plane.

Hyperbolic transformations. Suppose $\alpha \in \mathrm{SL}_2(\mathbb{R})$ and α is hyperbolic. Its characteristic polynomial is $X^2 + bX + 1$ with $|b| > 2$; hence $\Delta = b^2 - 4 > 0$, and so α has two distinct real eigenvectors. Thus α has two distinct fixed points in $\mathbb{R} \cup \{\infty\}$.

Let Γ be a discrete subgroup of $\mathrm{SL}_2(\mathbb{R})$. A point $z \in \mathbb{H}$ is called an *elliptic point* if it is the fixed point of an elliptic element γ of Γ; a point $s \in \mathbb{R} \cup \{\infty\}$ is called a *cusp* if there exists a parabolic element $\gamma \in \Gamma$ with s as its fixed point.

PROPOSITION 2.10. *If z is an elliptic point of Γ, then $\{\gamma \in \Gamma \mid \gamma z = z\}$ is a finite cyclic group.*

PROOF. There exists an $\alpha \in \mathrm{SL}_2(\mathbb{R})$ such that $\alpha(i) = z$, and $\gamma \mapsto \alpha^{-1}\gamma\alpha$ defines an isomorphism

$$\{\gamma \in \Gamma \mid \gamma z = z\} \approx \mathrm{SO}_2(\mathbb{R}) \cap (\alpha^{-1}\Gamma\alpha),$$

and this last group is finite. The correspondences $\theta \leftrightarrow e^{2\pi i\theta} \leftrightarrow \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ are isomorphisms

$$\mathbb{R}/\mathbb{Z} \leftrightarrow \{z \in \mathbb{C} \mid |z| = 1\} \leftrightarrow \mathrm{SO}_2(\mathbb{R}).$$

Therefore $\mathrm{SO}_2(\mathbb{R})_{\mathrm{tors}} \approx \mathbb{Q}/\mathbb{Z}$, and every finite subgroup of $\mathbb{Q}/\mathbb{Z}$ is cyclic (each is of the form $n^{-1}\mathbb{Z}/\mathbb{Z}$ where n is the least common denominator of the elements of the group). $\square$

REMARK 2.11. Let $\Gamma(1)$ be the full modular group $\mathrm{SL}_2(\mathbb{Z})$. I claim the cusps of $\Gamma(1)$ are exactly the points of $\mathbb{Q}\cup\{\infty\}$, and each is $\Gamma(1)$-equivalent to ∞. Certainly ∞ is the fixed point of the parabolic matrix $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. Suppose $m/n \in \mathbb{Q}$; we can assume m and n to be relatively prime, and so there are integers r and s such that $rm - sn = 1$; let $\gamma = \begin{pmatrix} m & s \\ n & r \end{pmatrix}$; then $\gamma(\infty) = m/n$, and m/n is fixed by the parabolic element $\gamma T \gamma^{-1}$. Conversely, every parabolic element α of $\Gamma(1)$ is conjugate to $\pm T$, say $\alpha = \pm\gamma T\gamma^{-1}$, $\gamma \in \mathrm{GL}_2(\mathbb{Q})$. The point fixed by α is $\gamma\infty$, which belongs to $\mathbb{Q}\cup\{\infty\}$.

We now find the elliptic points of $\Gamma(1)$. Let γ be an elliptic element in $\Gamma(1)$. The characteristic polynomial of γ is of degree 2, and its roots are roots of 1 (because γ has finite order). The only roots of 1 lying in a quadratic field have order dividing 4 or 6. From this, it easy to see that every elliptic point of $\mathbb{H}$ relative to $\Gamma(1)$ is $\Gamma(1)$-equivalent to exactly one of i or $\rho = (1+i\sqrt{3})/2$. (See also 2.12 below.)

Now let Γ be a subgroup of $\Gamma(1)$ of finite index. The cusps of Γ are the cusps of $\Gamma(1)$, namely, the elements of $\mathbb{Q}\cup\{\infty\} = \mathbb{P}^1(\mathbb{Q})$, but in general they will fall into more than one Γ-orbit. Every elliptic point of Γ is an elliptic point of $\Gamma(1)$; conversely, an elliptic point of $\Gamma(1)$ is an elliptic point of Γ if an only if it is fixed by an element of Γ other than $\pm I$.

Fundamental domains. Let Γ be a discrete subgroup of $\mathrm{SL}_2(\mathbb{R})$. A *fundamental domain* for Γ is a connected open subset D of $\mathbb{H}$ such that no two points of D are equivalent under Γ and $\mathbb{H} = \bigcup \gamma\bar{D}$, where $\bar{D}$ is the closure of D . These conditions are equivalent respectively, to the statements: the map $D \to \Gamma\backslash\mathbb{H}$ is injective; the map $\bar{D} \to \Gamma\backslash\mathbb{H}$ is surjective. Every Γ has a fundamental domain, but we shall prove this only for the subgroups of finite index in $\Gamma(1)$.

Let $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ and $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. Thus

$$\begin{array}{ll} Sz = \frac{-1}{z}, & Tz = z+1 \\ S^2 \equiv 1 \mod \pm I, & (ST)^3 \equiv 1 \mod \pm I. \end{array}$$

To apply S to a z with $|z| = 1$, first reflect in the x-axis, and then reflect through the origin (because $S(e^{i\theta}) = -(e^{-i\theta})$).

THEOREM 2.12. *Let* $D = \{z \in \mathbb{H} \mid |z| > 1,\ |\Re(z)| < 1/2\}$.

(a) *D is a fundamental domain for $\Gamma(1) = \mathrm{SL}_2(\mathbb{Z})$; moreover, two elements z and z' of $\bar{D}$ are equivalent under $\Gamma(1)$ if and only if*
  (i) *$\Re(z) = \pm 1/2$ and $z' = z \pm 1$, (then $z' = Tz$ or $z = Tz'$), or*
  (ii) *$|z| = 1$ and $z' = -1/z = Sz$.*

(b) *Let $z \in \bar{D}$; if the stabilizer of $z \neq \{\pm I\}$, then*
  (i) *$z = i$, and $Stab(i) = \langle S \rangle$, which has order 2 in $\Gamma(1)/\{\pm I\}$), or*
  (ii) *$z = \rho = \exp(2\pi i/6)$, and $Stab(\rho) = \langle TS \rangle$, which has order 3 in $\Gamma(1)/\{\pm I\}$), or*
  (iii) *$z = \rho^2$, and $Stab(\rho^2) = \langle ST \rangle$, which has order 3 in $\Gamma(1)/\{\pm I\}$).*

(c) *The group $\Gamma(1)/\{\pm I\}$ is generated by S and T.*

PROOF. Let Γ' be the subgroup of $\Gamma(1)$ generated by S and T. We shall show that $\Gamma' \cdot D = \mathbb{H}$.

LEMMA 2.13. *For a fixed $z \in \mathbb{H}$ and $N \in \mathbb{N}$, there are only finitely many pairs of integers (c, d) such that*

$$|cz + d| \leq N.$$

PROOF. Write $z = x + iy$. If (c, d) is such a pair, then

$$|cz + d|^2 = (cx + d)^2 + c^2y^2,$$

so that

$$c^2y^2 \leq (cx + d)^2 + c^2y^2 \leq N.$$

As $z \in \mathbb{H}$, $y > 0$, and so $|c| \leq N/y$, which implies that there are only finitely many possibilities for c. For any such c, the equation

$$(cx + d)^2 + c^2y^2 \leq N$$

shows that there are only finitely many possible values of d. □

Recall that, if $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then $\Im(\gamma z) = \Im(z)/|cz + d|^2$. Fix a $z \in \mathbb{H}$, and choose $\gamma \in \Gamma'$ such that $|cz + d|$ is a minimum—the lemma implies that such a γ exits. Then $\Im(\gamma z)$ is a maximum among elements in the orbit of z.

For some n, $z' \stackrel{\text{df}}{=} T^n(\gamma z)$ will have

$$-1/2 \leq \Re(z') \leq 1/2.$$

I claim that $|z'| \geq 1$. If not, then

$$\Im(Sz') = \Im(-1/z') = \Im\left(\frac{-x' + iy'}{|z'|^2}\right) = \frac{\Im(z')}{|z'|^2} > \Im(z') = \Im(\gamma z),$$

which contradicts our choice of γz. We have shown that $\Gamma' \cdot \bar{D} = \mathbb{H}$.

Suppose $z, z' \in \bar{D}$ are Γ-conjugate. Then either $\Im(z) \geq \Im(z')$ or $\Im(z) \leq \Im(z')$, and we shall assume the latter. Suppose $z' = \gamma z$ with $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, and let $z = x + iy$. Then our assumption implies that

$$(cx + d)^2 + cy^2 = |cz + d|^2 \leq 1.$$

This is impossible if $c \geq 2$ (because $y \geq 1/2$), and so we need only consider the cases $c = 0, 1, -1$.

$c = 0$: Then $d = \pm 1$, $\gamma = \pm\begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix}$, and γ is translation by b. Because z and $\gamma z \in \bar{D}$, this implies that $b = \pm 1$, and we are in case (a(i)).

$c = 1$: As $|z + d| \leq 1$ we must have $d = 0$, unless $z = \rho = \frac{1}{2} + i\frac{\sqrt{3}}{2}$, in which case $d = 0$ or -1, or $z = \rho^2$, in which case $d = 0$ or 1. If $d = 0$, then $\gamma = \pm\begin{pmatrix} a & -1 \\ 1 & 0 \end{pmatrix}$, and $\gamma z = a - \frac{1}{z}$. If $a = 0$, then we are in case (a(ii)). If $a \neq 0$, then $a = 1$ and $z = \rho^2$, or $a = -1$ and $z = \rho$.

$c = -1$: This case can be treated similarly (or simply change the signs of a, b, c, d in the last case).

This completes the proof of (a) and (b) of the theorem.

We now prove (c). Let $\gamma \in \Gamma$. Choose a point $z_0 \in D$. Because $\Gamma \cdot \bar{D} = \mathbb{H}$, there is an element $\gamma' \in \Gamma'$ and a point $z \in \bar{D}$ such that $\gamma' z = \gamma z_0 \in \bar{D}$. Then z_0 is $\Gamma(1)$-equivalent to $(\gamma'^{-1}\gamma)z_0 \in \bar{D}$; because $z_0 \in D$, part (a) shows that $z_0 = (\gamma'^{-1}\gamma)z_0$. Hence $\gamma'^{-1}\gamma \in \mathrm{Stab}(z_0) \cap \Gamma(1) = \{\pm I\}$, and so γ' and γ are equal in $\Gamma(1)/\{\pm 1\}$. □

Remark 2.14. We showed that the group $\Gamma(1)/\{\pm I\}$ has generators S and T with relations $S^2 = 1$ and $(ST)^3 = 1$. One can show that this is a full set of relations, and that $\Gamma(1)/\{\pm I\}$ is the free product of the cyclic group of order 2 generated by S and the cyclic group of order 3 generated by ST.

Aside 2.15. Our computation of the fundamental domain has applications for quadratic forms and sphere packings.

Consider a binary quadratic form:

$$q(x, y) = ax^2 + bxy + cy^2, \quad a, b, c \in \mathbb{R}.$$

Assume q is definite, i.e., its discriminant $\Delta = b^2 - 4ac < 0$. Two forms q and q' are *equivalent* if there is a matrix $A \in \mathrm{SL}_2(\mathbb{Z})$ taking q into q' by the change of variables,

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = A \begin{pmatrix} x \\ y \end{pmatrix}.$$

In other words, the forms

$$q(x, y) = (x, y) \cdot Q \cdot \begin{pmatrix} x \\ y \end{pmatrix}, \qquad q'(x, y) = (x, y) \cdot Q' \cdot \begin{pmatrix} x \\ y \end{pmatrix}$$

are equivalent if $Q = A^{tr} \cdot Q' \cdot A$.

Every definite binary quadratic form can be written $q(x, y) = a(x - \omega y)(x - \bar{\omega} y)$ with $\omega \in \mathbb{H}$. The association $q \leftrightarrow \omega$ is a one-to-one correspondence between the definite binary quadratic forms with a fixed discriminant Δ and the points of $\mathbb{H}$. Moreover, two forms are equivalent if and only if the points lie in the same $\mathrm{SL}_2(\mathbb{Z})$-orbit. A definite binary quadratic form is said to be *reduced* if ω is in

$$\{z \in \mathbb{H} \mid -\frac{1}{2} \le \Re(z) < 1 \text{ and } |z| > 1, \text{ or } |z| = 1 \text{ and } -\frac{1}{2} \le \Re(z) \le 0\}.$$

More explicitly, $q(x, y) = ax^2 + bxy + cy^2$ is reduced if and only if either

$$-a < b \le a < c \text{ or } 0 \le b \le a = c.$$

Theorem 2.12 implies:

> Every definite binary quadratic form is equivalent to a reduced form; two reduced forms are equivalent if and only if they are equal.

We say that a quadratic form is *integral* if it has integral coefficients.

> There are only finitely many equivalence classes of integral definite binary quadratic forms with a given discriminant.

Each equivalence class contains exactly one reduced form $ax^2+bxy+cy^2$. Since

$$4a^2 \leq 4ac = b^2 - \Delta \leq a^2 - \Delta$$

we see that there are only finitely many values of a for a fixed Δ. Since $|b| \leq a$, the same is true of b, and for each pair (a,b) there is at most one integer c such that $b^2 - 4ac = \Delta$.

For more details, see W. LeVeque, Topics in Number Theory, II, Addison-Wesley, 1956, Chapter 1.

We can apply this to lattice sphere packings in $\mathbb{R}^2$. Such a packing is determined by the lattice of centres of the spheres (here disks). The object, of course, is to make the packing as dense as possible. With a lattice Λ in $\mathbb{R}^2$ and a choice of a basis $\{f_1, f_2\}$ for Λ, we can associate the quadratic form

$$q(x_1,x_2) = \|f_1x_1 + f_2x_2\|^2.$$

The problem of finding dense sphere packings translates into finding quadratic forms q with

$$\gamma(q) \stackrel{\text{df}}{=} \min\{q(x) \mid x \in \mathbb{Z}^2, \quad x \neq 0\}^2/\operatorname{disc}(q)$$

as large as possible. Note that changing the choice of basis for Λ translates into acting on q with an element of $\mathrm{SL}_2(\mathbb{Z})$, and so we can confine our attention to reduced quadratic forms. It is then easy to show that the quadratic form with $\gamma(q)$ minimum is that corresponding to ρ. The corresponding lattice has basis $\begin{pmatrix} 2 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ \sqrt{3} \end{pmatrix}$ (just as you would expect), and the quadratic form is $4(x^2+xy+y^2)$.

For more on sphere packings, see Math 679, Section 21.

Fundamental domains for congruence subgroups. First we have the following general result.

Proposition 2.16. *Let Γ be a discrete subgroup of $\mathrm{SL}_2(\mathbb{R})$, and let D be a fundamental domain for Γ. Let Γ' be a subgroup of Γ of finite index, and write Γ as a disjoint union of right cosets of Γ' :*

$$\Gamma = \Gamma'\gamma_1 \cup \ldots \cup \Gamma'\gamma_m$$

Then $D' \stackrel{\text{df}}{=} \bigcup \gamma_i D$ is a fundamental domain for Γ' (possibly nonconnected).

Proof. Let $z \in \mathbb{H}$. Then $z = \gamma z'$ for some $z' \in \bar{D}$, $\gamma \in \Gamma$, and $\gamma = \gamma'\gamma_i$ for some $\gamma' \in \Gamma'$. Thus $z = \gamma'\gamma_i z \in \Gamma' \cdot (\gamma_i \bar{D})$.

If $\gamma D' \cap D' \neq \emptyset$, then it would contain a transform of D. But then $\gamma\gamma_i D = \gamma_j D$ for some $i \neq j$, which would imply that $\gamma\gamma_i = \gamma_j$, and this is a contradiction. □

Proposition 2.17. *It is possible to choose the γ_i so that the closure of D' is connected; the interior of the closure of D' is then a connected fundamental domain for Γ.*

Proof. Omit. □

Remark 2.18. Once one has obtained a fundamental domain for Γ, as in (2.16), it is possible to read off a system of generators and relations for Γ.

Defining complex structures on quotients. Before defining $\mathbb{H}^*$ and the complex structure on the quotient $\Gamma\backslash\mathbb{H}^*$ we discuss two simple examples.

EXAMPLE 2.19. Let D be the open unit disk, and let Δ be a finite group acting on D. The Schwarz lemma implies that $\mathrm{Aut}(D) = \{z \in \mathbb{C} \mid |z| = 1\} \approx \mathbb{R}/\mathbb{Z}$, and it follows that Δ is a finite cyclic group. Let $z \mapsto \zeta z$ be its generator and suppose that $\zeta^m = 1$. Then z^m is invariant under Δ, and so defines a function on $\Delta\backslash D$. It is a homeomorphism from $\Delta\backslash D$ onto D, and therefore defines a complex structure on $\Delta\backslash D$.

Let p be the quotient map $D \to \Delta\backslash D$. The map $f \mapsto f \circ p$ is a bijection from the holomorphic functions on $U \subset \Delta\backslash D$ to the holomorphic functions of z^m on $p^{-1}(U) \subset D$; but these are precisely the holomorphic functions on $p^{-1}(U)$ invariant under the action of Δ.

EXAMPLE 2.20. Let $X = \{z \in \mathbb{C} \mid \Im(z) > c\}$ (some c). Fix an integer h, and let $n \in \mathbb{Z}$ act on X as $z \mapsto z + nh$. Add a point "∞" and define a topology on $X^* = X \cup \{\infty\}$ as follows: a fundamental system of neighbourhoods of a point in X is as before; a fundamental system of neighbourhoods for ∞ is formed of sets of the form $\{z \in \mathbb{C} \mid \Im(z) > N\}$. We can extend the action of $\mathbb{Z}$ on X to a continuous action on X^* by requiring $\infty + nh = \infty$ for all $n \in \mathbb{Z}$. Consider the quotient space $\Gamma\backslash X^*$. The function

$$q(z) = \begin{cases} e^{2\pi i z/h} & z \neq \infty, \\ 0 & z = \infty \end{cases}$$

is a homeomorphism $\Gamma\backslash X^* \to D$ from $\Gamma\backslash X^*$ onto the open disk of radius $e^{-2\pi c/h}$ and centre 0. It therefore defines a complex structure on $\Gamma\backslash X^*$.

The complex structure on $\Gamma(1)\backslash\mathbb{H}^*$. We first define the complex structure on $\Gamma(1)\backslash\mathbb{H}$. Write p for the quotient map $\mathbb{H} \to \Gamma(1)\backslash H$. Let P be a point of $\Gamma(1)\backslash\mathbb{H}$, and let Q be a point of $\mathbb{H}$ mapping to it.

If Q is not an elliptic point, we can choose a neighbourhood U of Q such that p is a homeomorphism $U \to p(U)$. We define $(p(U), p^{-1})$ to be a coordinate neighbourhood of P.

If Q is equivalent to i, we may as well take it to equal i. The map $z \mapsto \frac{z-i}{z+i}$ defines an isomorphism of an open disk D with centre i onto an open disk D' with centre 0, and the action of S on D is transformed into the automorphism $\sigma : z \mapsto -z$ of D' (because it fixes i and has order 2). Thus $<S>\backslash D$ is homeomorphic to $<\sigma>\backslash D'$, and we give $<S>\backslash D$ the complex structure making this a bi-holomorphic isomorphism. More explicitly, $\frac{z-i}{z+i}$ is a holomorphic function defined in a neighbourhood of i, and $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ maps it to

$$\frac{-z^{-1} - i}{-z^{-1} + i} = \frac{-1 - iz}{-1 + iz} = \frac{-i + z}{-i - z} = -\frac{z - i}{z + i}$$

Thus $z \mapsto (\frac{z-i}{z+i})^2$ is a holomorphic function defined in a neighbourhood of i which is invariant under the action of S; it therefore defines a holomorphic function in a neighbourhood of $p(i)$, and we take this to the coordinate function near $p(i)$.

The point $Q = \rho^2$ can be treated similarly. Apply a linear fractional transformation that maps Q to zero, and then take the cube of the map. Explicitly, ρ^2 is fixed by ST, which has order 3 (as a transformation). The function $z \mapsto \frac{z-\rho^2}{z-\bar{\rho}^2}$ defines an isomorphism from a disk with centre ρ^2 onto a disk with centre 0, and $(\frac{z-\rho^2}{z-\bar{\rho}^2})^3$ is invariant under ST. It therefore defines a function on a neighbourhood of $p(\rho^2)$, and we take this to be the coordinate function near $p(\rho^2)$.

The Riemann surface $\Gamma(1)\backslash\mathbb{H}$ we obtain is not compact—to compactify it, we need to add a point. The simplest way to do this is to add a point ∞ to $\mathbb{H}$, as in (2.20), and use the function $q(z) = \exp(2\pi i z)$ to map some neighbourhood $U = \{z \in \mathbb{H} \mid \Im(z) > N\}$ of ∞ onto an open disk V with centre 0. The function q is invariant under the action of the stabilizer of $<T>$ of ∞, and so defines a holomorphic function $q : <T> \backslash U \to V$, which we take to be the coordinate function near $p(\infty)$.

Alternatively, we can consider $\mathbb{H}^* = H \cup \mathbb{P}^1(\mathbb{Q})$, i.e., $\mathbb{H}^*$ is the union of $\mathbb{H}$ with the set of cusps for $\Gamma(1)$. Each cusp other than[2] ∞ is a rational point on the real axis, and is of the form $\sigma\infty$ for some $\sigma \in \Gamma(1)$ (see 2.11). Give $\sigma\infty$ the fundamental system of neighbourhoods for which σ is a homeomorphism. Then $\Gamma(1)$ acts continuously on $\mathbb{H}^*$, and we can consider the quotient space $\Gamma(1)\backslash\mathbb{H}^*$. Clearly, $\Gamma(1)\backslash\mathbb{H}^* = (\Gamma(1)\backslash H) \cup \{\infty\}$, and we can endow it with the same complex structure as before.

PROPOSITION 2.21. *The Riemann surface $\Gamma(1)\backslash\mathbb{H}^*$ is a compact and of genus zero; it is therefore isomorphic to the Riemann sphere.*

PROOF. It is compact because $\bar{D} \cup \{\infty\}$ is compact. We sketch four proofs that it has genus 0. First, by examining carefully how the points of $\bar{D}$ are identified, one can see that it must be homeomorphic to a sphere. Second, show that it is simply connected (loops can be contracted), and the Riemann sphere is the only simply connected compact Riemann surface (Riemann Mapping Theorem 0.1). Third, triangulate it by taking ρ, i, and ∞ as the vertices of the obvious triangle, add a fourth vertex not on any side of the triangle, and join it to the first three vertices; then $2 - 2g = 4 - 6 + 4 = 2$. Finally, there is a direct proof that there is a function j holomorphic on $\Gamma\backslash\mathbb{H}$ and having a simple pole at ∞—it is therefore of valence one, and so defines an isomorphism of $\Gamma\backslash\mathbb{H}^*$ onto the Riemann sphere. □

The complex structure on $\Gamma\backslash\mathbb{H}^*$. Let $\Gamma \subset \Gamma(1)$ of finite index. We can define a compact Riemann surface $\Gamma\backslash\mathbb{H}^*$ in much the same way as for $\Gamma(1)$. The complement of $\Gamma\backslash\mathbb{H}$ in $\Gamma\backslash\mathbb{H}^*$ is the set of equivalence classes of cusps for Γ.

First $\Gamma\backslash\mathbb{H}$ is given a complex structure in exactly the same way as in the case $\Gamma = \Gamma(1)$. The point ∞ will always be cusp (Γ must contain T^h for some h, and T^h is a parabolic element fixing ∞). If h is the smallest power of T in Γ, then the function $q = \exp(2\pi i z/h)$ is a coordinate function near ∞. Any other cusp for Γ is of the form $\sigma\infty$ for $\sigma \in \Gamma(1)$, and $z \mapsto q(\sigma^{-1}(z))$ is a coordinate function near $\sigma\infty$.

We write $Y(\Gamma) = \Gamma\backslash\mathbb{H}$ and $X(\Gamma) = \Gamma\backslash\mathbb{H}^*$. We abbreviate $Y(\Gamma(N))$ to $Y(N)$, $X(\Gamma(N))$ to $X(N)$, $Y(\Gamma_0(N))$ to $Y_0(N)$, $X(\Gamma_0(N))$ to $X_0(N)$ and so on.

[2]We sometimes denote ∞ by $i\infty$ and imagine it to be at the end of the imaginary axis.

The genus of $X(\Gamma)$**.** We now compute the genus of $X(\Gamma)$ by considering it as a covering of $X(\Gamma(1))$. According to (1.27)

$$2g-2=-2m+\sum(e_P-1)$$

or

$$g=1-m+\sum(e_P-1)/2.$$

where m is the degree of the covering $X(\Gamma)\to X(\Gamma(1))$ and e_P is the ramification index at the point P. The ramification points are the images of elliptic points on $\mathbb{H}^*$ and the cusps.

THEOREM 2.22. *Let* Γ *be a subgroup of* $\Gamma(1)$ *of finite index, and let* ν_2 *=the number of inequivalent elliptic points of order* 2*;* ν_3 *=the number of inequivalent elliptic points of order* 3*;* ν_∞ *= the number of inequivalent cusps. Then the genus of* $X(\Gamma)$ *is*

$$g=1+m/12-\nu_2/4-\nu_3/3-\nu_\infty/2.$$

PROOF. Let p be the quotient map $\mathbb{H}^*\to\Gamma(1)\backslash\mathbb{H}^*$, and let φ be the map $\Gamma\backslash\mathbb{H}^*\to\Gamma(1)\backslash H^*$. If Q is a point of $\mathbb{H}^*$ and P' and P are its images in $\Gamma\backslash\mathbb{H}^*$ and $\Gamma(1)\backslash\mathbb{H}^*$ then the ramification indices multiply:

$$e(Q/P)=e(Q/P')\cdot e(P'/P).$$

If Q is a cusp, then this formula is not useful, as $e(Q/P)=\infty=e(Q/P')$ (the map p is $\infty:1$ on every neighbourhood of ∞). For $Q\in\mathbb{H}$ and not an elliptic point it tells us P' is not ramified.

Suppose that $P=p(i)$, so that Q is $\Gamma(1)$-equivalent to i. Then either $e(Q/P')=2$ or $e(P'/P)=2$. In the first case, Q is an elliptic point for Γ and P' is unramified over P; in the second, Q is not an elliptic point for Γ, and the ramification index of P' over P is 2. There are ν_2 points P' of the first type, and $(m-\nu_2)/2$ points of the second. Hence $\sum e_{P'}-1=(m-\nu_2)/2$.

Suppose that $P=p(\rho)$, so that Q is $\Gamma(1)$-equivalent to ρ. Then either $e(Q/P')=3$ or $e(P'/P)=3$. In the first case, Q is an elliptic point for Γ and P' is unramified over P; in the second, Q is not an elliptic point for Γ, and the ramification index of P' over is 3. There are ν_3 points P' of the first type, and $(m-\nu_3)/3$ points of the second. Hence $\sum e_{P'}-1=2(m-\nu_3)/3$.

Suppose that $P=p(\infty)$, so that Q is a cusp for Γ. There are ν_∞ points P' and $\sum e_i=m$; hence $\sum e_i-1=m-\nu_\infty$.

We conclude:

$$\begin{aligned}\sum(e_{P'}-1)&=(m-\nu_2)/2\quad(P'\text{ lying over }\varphi(i))\\ \sum(e_{P'}-1)&=2(m-\nu_3)/3\quad(P'\text{ lying over }\varphi(\rho))\\ \sum(e_{P'}-1)&=(m-\nu_\infty)\quad(P'\text{ lying over }\varphi(\infty)).\end{aligned}$$

Therefore

$$g=1-m+\sum(e_P-1)/2=1+m/12-\nu_2/4-\nu_3/3-\nu_\infty/2.$$

□

EXAMPLE 2.23. Consider the principal congruence subgroup $\Gamma(N)$. We have to compute the index of $\Gamma(N)$ in Γ, i.e., the order of $\mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$. One sees easily that:

(a) $\mathrm{GL}_2(\mathbb{Z}/N\mathbb{Z}) \approx \prod \mathrm{GL}_2(\mathbb{Z}/p_i^{r_i}\mathbb{Z})$ if $N = \prod p_i^{r_i}$ (because $\mathbb{Z}/N\mathbb{Z} \approx \prod \mathbb{Z}/p_i^{r_i}\mathbb{Z}$).
(b) The order of $\mathrm{GL}_2(\mathbb{F}_p) = (p^2-1)(p^2-p)$ (because the top row of a matrix in $\mathrm{GL}_2(\mathbb{F}_p)$ can be any nonzero element of k^2, and the second row can then be any element of k^2 not on the line spanned by the first row).
(c) The kernel of $\mathrm{GL}_2(\mathbb{Z}/p^r\mathbb{Z}) \to \mathrm{GL}_2(\mathbb{F}_p)$ consists of all matrices of the form $I + p\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with $a, b, c, d \in \mathbb{Z}/p^{r-1}\mathbb{Z}$, and so the order of $\mathrm{GL}_2(\mathbb{Z}/p^r\mathbb{Z})$ is $(p^{r-1})^4 \cdot (p^2-1)(p^2-p)$.
(d) $\#\,\mathrm{GL}_2(\mathbb{Z}/p^r\mathbb{Z}) = \varphi(p^r) \cdot \#\,\mathrm{SL}_2(\mathbb{Z}/p^r\mathbb{Z})$, where $\varphi(p^r) = \#(\mathbb{Z}/p^r\mathbb{Z})^\times = (p-1)p^{r-1}$.

On putting these statements together, one finds that

$$(\Gamma(1) : \Gamma(N)) = N^3 \cdot \prod_{p|N}(1-p^{-2}).$$

Write $\bar{\Gamma}(N)$ for the image of $\Gamma(N)$ in $\Gamma(1)/\{\pm I\}$. Then

$$(\bar{\Gamma}(1) : \bar{\Gamma}(N)) = (\Gamma(1) : \Gamma(N))/2,$$

unless $N = 2$, in which case it $= 6$.

What are ν_2, ν_3, and ν_∞? Assume $N > 1$. Then $\Gamma(N)$ has no elliptic points—the only torsion elements in $\bar{\Gamma}(1)$ are $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, $ST = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}$, $(ST)^2$, and their conjugates; none of these three elements is in $\Gamma(N)$ for any $N > 1$, and because $\Gamma(N)$ is a normal subgroup, their conjugates can't be either. The number of inequivalent cusps is μ_N/N where $\mu_N = (\bar{\Gamma}(1) : \bar{\Gamma}(N)$ (see 2.24). We conclude that the genus of $\Gamma(N)\backslash\mathbb{H}^*$ is

$$g(N) = 1 + \mu_N \cdot (N-6)/12N \quad (N > 1).$$

For example,

$$\begin{array}{lclllllllll} N &=& 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ \mu &=& 6 & 12 & 24 & 60 & 72 & 168 & 192 & 324 & 360 & 660 \\ g &=& 0 & 0 & 0 & 0 & 1 & 3 & 5 & 10 & 13 & 26. \end{array}$$

There are similarly explicit formulas for the genus of $X_0(N)$—see Shimura 1971, p25.

EXERCISE 2.24. Let G be a group (possibly infinite) acting transitively on a set X, and let H be a subgroup of finite index in G. Fix a point x_0 in X and let G_0 be the stabilizer of x_0 in G, and let H_0 be the stabilizer of x_0 in H. Prove that the number of orbits of H acting on X is

$$(G : H)/(G_0 : H_0).$$

Deduce that the number of inequivalent cusps for $\Gamma(N)$ is μ_N/N.

REMARK 2.25. The Taniyama conjecture says that, for any elliptic curve E over $\mathbb{Q}$, there exists a surjective map $X_0(N) \to E$, where N is the conductor of E (the conductor of E is divisible only by the primes where E has bad reduction). The conjecture is suggested by studying zeta functions (see later). For any particular N, it is possible to verify the conjecture by listing all elliptic curves over $\mathbb{Q}$ with conductor N, and checking that there is a map $X_0(N) \to E$. (Mathematica includes a list of elliptic curves with small conductor.) It is known (Ribet) that the Taniyama conjecture implies Fermat's last theorem. The Taniyama conjecture has been proved for most elliptic curves by Wiles, Taylor, and Diamond (see Math 679).

An elliptic curve for which there is a map $X_0(N) \to E$ for some N is called a *modular elliptic curve*; contrast *elliptic modular curves* which are the curves of the form $\Gamma\backslash\mathbb{H}^*$ for Γ a subgroup of finite index in $\Gamma(1)$.

ASIDE 2.26. A *bounded symmetric domain* X is a bounded open connected subset of $\mathbb{C}^n$, some n, that is symmetric in the following sense: for each point $x \in X$, there is an involution s_x of X having x as an isolated fixed point. A complex manifold isomorphic to a bounded symmetric domain is also (loosely) referred to as a bounded symmetric domain.

For example, the unit disk D is a bounded symmetric domain—0 is the fixed point of the involution $z \mapsto -z$, and since $\mathrm{Aut}(D)$ acts transitively on D this shows every other point must also be the fixed point of an involution. Every bounded symmetric domain is simply connected, and so (by the Riemann mapping theorem) every bounded symmetric domain of dimension one is isomorphic to the unit disk. As $\mathbb{H} \approx D$, we also refer to $\mathbb{H}$ as a bounded symmetric domain.

The bounded symmetric domains of all dimensions were classified by Elie Cartan (except for the exceptional ones). Just as for $\mathbb{H}$, the group of automorphisms $\mathrm{Aut}(X)$ of a bounded symmetric domain is a Lie group, which is simple if X is indecomposable (i.e., not equal to a product of bounded symmetric domains). There are bounded symmetric domains attached to groups of type A_n, B_n, C_n, D_n, E_6, E_7 (here n is an integer ≥ 1).

Let X be a bounded symmetric domain. One can find many semisimple algebraic groups G over $\mathbb{Q}$ for which there exists a homomorphism $G(\mathbb{R})^+ \to \mathrm{Aut}(X)$ with finite cokernel and compact kernel—the $^+$ denotes the identity component of $G(\mathbb{R})$ for the real topology. For example, we saw above that any quaternion algebra over $\mathbb{Q}$ that splits over $\mathbb{R}$ gives rise to such a group for $\mathbb{H}$. Given such a G, one defines congruence subgroups $\Gamma \subset G(\mathbb{Z})$ just as for $\mathrm{SL}_2(\mathbb{Z})$, and studies the quotients.

In 1964, Baily and Borel showed that each quotient $\Gamma\backslash X$ has a unique structure as an algebraic variety; in fact, they proved that $\Gamma\backslash X$ could be embedded in a natural way into a projective algebraic variety $\Gamma\backslash X^*$.

Various examples of these varieties were studied by Poincaré, Hilbert, Siegel, and many others, but Shimura began an intensive study of them in the 1960's, and they are now called *Shimura varieties.*

Given a Shimura variety $\Gamma\backslash X^*$, one can attach a number field E to it, and prove that the Shimura variety is defined, in a natural way, over E. Thus one obtains a vast array of varieties defined over number fields, all with very interesting arithmetic properties. In this course, we study only the simplest case.

3. Elliptic functions

In this section, we review some of the theory of elliptic functions. For more details, see Cartan 1963, V.2.5, VI.5.3, or Silverman 1986, VI, or Koblitz 1984, I. Section 10 of my notes on Elliptic Curves (Math 679) is an expanded version of this section.

Lattices and bases. Let ω_1 and ω_2 be two nonzero complex numbers such that $\tau = \omega_1/\omega_2$ is imaginary. By interchanging ω_1 and ω_2 if necessary, we can ensure that $\tau = \omega_1/\omega_2$ lies in the upper half plane. Write

$$\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2,$$

so that Λ is the lattice generated by ω_1 and ω_2. We are interested in Λ rather than the basis $\{\omega_1, \omega_2\}$. If $\{\omega_1', \omega_2'\}$ is a second pair of elements of Λ, so that

$$\omega_1' = a\omega_1 + b\omega_2, \quad \omega_2' = c\omega_1 + d\omega_2$$

for some $a, b, c, d \in \mathbb{Z}$, then under what conditions on a, b, c, d does $\{\omega_1', \omega_2'\}$ form another basis for Λ with $\tau' = \omega_1'/\omega_2' \in \mathbb{H}$? Clearly ω_1' and ω_2' generate Λ if and only if $\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \pm 1$. We have $\tau' = \frac{a\tau+b}{c\tau+d}$, and the calculation on p2 shows that $\Im(\tau') = \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \Im(\tau)/|cz+d|^2$, and so we need that $\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} > 1$. Therefore, the bases (ω_1', ω_2') of Λ with $\Im(\omega_1'/\omega_2') > 0$ are those of the form

$$\begin{pmatrix} \omega_1' \\ \omega_2' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} \text{ with } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}).$$

Any parallelogram with vertices z_0, $z_0+\omega_1$, $z_0+\omega_1+\omega_2$, $z_0+\omega_2$, where $\{\omega_1, \omega_2\}$ is a basis for Λ, is called a *fundamental parallelogram* for Λ.

Quotients of $\mathbb{C}$ by lattices. Let Λ be a lattice in $\mathbb{C}$ (by which I always mean a full lattice, i.e., a set of the form $\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$ with ω_1 and ω_2 linearly independent over $\mathbb{R}$). We can make the quotient space $\mathbb{C}/\Lambda$ into a Riemann surface as follows: let Q be a point in $\mathbb{C}$ and let P be its image $\mathbb{C}/\Lambda$; then there exist neighbourhoods V of Q and U of P such that the quotient map $p : \mathbb{C} \to \mathbb{C}/\Lambda$ defines a homeomorphism $V \to U$; we take every such pair $(U, p^{-1} : V \to U)$ to be a coordinate neighbourhood. In this way we get a complex structure on $\mathbb{C}/\Lambda$ having the following property: the map $p : \mathbb{C} \to \mathbb{C}/\Lambda$ is holomorphic, and for any open subset U of $\mathbb{C}/\Lambda$, a function $f : U \to \mathbb{C}$ is holomorphic if and only if $f \circ p$ is holomorphic on $p^{-1}(U)$.

Topologically, $\mathbb{C}/\Lambda \approx (\mathbb{R}/\mathbb{Z})^2$, which is a single-holed torus. Thus $\mathbb{C}/\Lambda$ has genus 1. All spaces $\mathbb{C}/\Lambda$ are homeomorphic, but, as we shall see, they are *not* all isomorphic as Riemann surfaces.

Doubly periodic functions. Let Λ be a lattice in $\mathbb{C}$. A meromorphic function $f(z)$ on the complex plane is said to be *doubly periodic with respect to* Λ if it satisfies the functional equation:

$$f(z+\omega) = f(z) \text{ for each } \omega \in \Lambda.$$

Equivalently,

$$f(z+\omega_1) = f(z),\ f(z+\omega_2) = f(z)$$

for $\{\omega_1, \omega_2\}$ a basis for Λ.

PROPOSITION 3.1. *Let $f(z)$ be a doubly periodic function for Λ, not identically zero, and let D be a fundamental parallelogram for Λ such that f has no zeros or poles on the boundary of D. Then*

(a) $\sum_{P \in D} \mathrm{Res}_P(f) = 0$;
(b) $\sum_{P \in D} \mathrm{ord}_P(f) = 0$;
(c) $\sum_{P \in D} \mathrm{ord}_P(f) \cdot P \equiv 0 \pmod{\Lambda}$.

The first sum is over the points of D where f has a pole, and the other sums are over the points where it has a zero or pole. Each sum is finite.

PROOF. Regard f as a function on $\mathbb{C}/\Lambda$, and apply Proposition 1.12 to get (a) and (b). To get (c) apply (1.12b) to $z \cdot f'(z)/f(z)$. □

COROLLARY 3.2. *A nonconstant doubly periodic function has at least two poles.*

PROOF. A doubly periodic function that is holomorphic is bounded in a closed period parallelogram (by compactness), and hence on the entire plane (by periodicity); so it is constant, by Liouville's theorem. A doubly periodic function with a simple pole in a period parallelogram is impossible, because by (3.1a) the residue at the pole would be zero, and so the function would be holomorphic. □

Endomorphisms of $\mathbb{C}/\Lambda$. Note that $\mathbb{C}/\Lambda$ has a natural group structure.

PROPOSITION 3.3. *Let Λ and Λ' be two lattices in $\mathbb{C}$. An element $\alpha \in \mathbb{C}$ such that $\alpha\Lambda \subset \Lambda'$ defines a holomorphic map*

$$\varphi_\alpha : \mathbb{C}/\Lambda \to \mathbb{C}/\Lambda', \quad [z] \mapsto [\alpha z],$$

sending $[0]$ to $[0]$, and every such map is of this form (for a unique α).

PROOF. It is obvious that α defines such a map. Conversely, let $\varphi : \mathbb{C}/\Lambda \to \mathbb{C}/\Lambda'$ be a holomorphic map such that $\varphi([0]) = [0]$. Then $\mathbb{C}$ is the universal covering space of both $\mathbb{C}/\Lambda$ and $\mathbb{C}/\Lambda'$, and a standard result in topology shows that φ lifts to a continuous map $\tilde{\varphi} : \mathbb{C} \to \mathbb{C}$ such that $\tilde{\varphi}(0) = 0$:

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

Because the vertical maps are local isomorphisms, $\tilde{\varphi}$ is automatically holomorphic. For any $\omega \in \Lambda$, the map $z \mapsto \tilde{\varphi}(z+\omega) - \tilde{\varphi}(z)$ takes values in $\Lambda' \subset \mathbb{C}$. As it is continuous, $\mathbb{C}$ is connected, and Λ' is discrete, it must be constant. Therefore $\tilde{\varphi}'$ must be a holomorphic doubly periodic function, and so it is constant, say $\tilde{\varphi}'(z) = \alpha$. Then $\tilde{\varphi}(z) = \alpha z + \beta$, and the fact that $\tilde{\varphi}(0) = 0$ implies that $\beta = 0$. □

COROLLARY 3.4. *Any holomorphic map $\varphi : \mathbb{C}/\Lambda \to \mathbb{C}/\Lambda'$ such that $\varphi(0) = 0$ is a homomorphism.*

PROOF. Clearly $[z] \mapsto [\alpha z]$ is a homomorphism. □

Compare this with the result (Math 631, 5.37): any regular map $\varphi : A \to A'$ from an abelian variety A to an abelian variety A' such that $\varphi(0) = 0$ is a homomorphism.

COROLLARY 3.5. *The Riemann surfaces $\mathbb{C}/\Lambda$ and $\mathbb{C}/\Lambda'$ are isomorphic if and only if $\Lambda' = \alpha\Lambda$ for some $\alpha \in \mathbb{C}^{\times}$.*

COROLLARY 3.6. *For any lattice Λ, $\mathrm{End}(\mathbb{C}/\Lambda)$ is either $\mathbb{Z}$ or a subring R of the ring of integers in a quadratic imaginary number field K.*

PROOF. Write $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$ with $\tau \overset{\mathrm{df}}{=} \omega_1/\omega_2 \in \mathbb{H}$, and suppose that there exists an $\alpha \in \mathbb{C}$, $\alpha \notin \mathbb{Z}$, such that $\alpha\Lambda \subset \Lambda$. Then

$$\begin{aligned} \alpha\omega_1 &= a\omega_1 + b\omega_2 \\ \alpha\omega_2 &= c\omega_1 + d\omega_2, \end{aligned}$$

with $a, b, c, d \in \mathbb{Z}$. On dividing through by ω_2 we obtain the equations

$$\begin{aligned} \alpha\tau &= a\tau + b \\ \alpha &= c\tau + d. \end{aligned}$$

As $\alpha \notin \mathbb{Z}$, $c \neq 0$. On eliminating α from the between the two equations, we find that

$$c\tau^2 + (d - a)\tau + b = 0.$$

Therefore $\mathbb{Q}[\tau]$ is of degree 2 over $\mathbb{Q}$. On eliminating τ from between the two equations, we find that

$$\alpha^2 - (a + d)\alpha + bc = 0.$$

Therefore α is integral over $\mathbb{Z}$, and hence is contained in the ring of integers of $\mathbb{Q}[\tau]$. $\square$

The Weierstrass $\wp$-function. . We want to construct some doubly periodic functions. Note that when G is a *finite* group acting on a set S, then it is easy to construct functions invariant under the action of G : take h to be any function $h : S \to \mathbb{C}$, and define

$$f(s) = \sum_{g \in G} h(gs);$$

then $f(g's) = \sum_{g \in G} h(g'gs) = f(s)$, and so f is invariant (and all invariant functions are of this form, obviously). When G is not finite, one has to verify that the series converges—in fact, in order to be able to change the order of summation, one needs absolute convergence. Moreover, when S is a Riemann surface and h is holomorphic, to ensure that f is holomorphic, one needs that the series converges absolutely uniformly on compact sets.

Now let $\varphi(z)$ be a holomorphic function $\mathbb{C}$ and write

$$\Phi(z) = \sum_{\omega \in \Lambda} \varphi(z + \omega).$$

Assume that as $|z| \to \infty$, $\varphi(z) \to 0$ so fast that the series for $\Phi(z)$ is absolutely convergent for all z for which none of the terms in the series has a pole. Then $\Phi(z)$ is doubly periodic with respect to Λ; for replacing z by $z + \omega_0$ for some $\omega_0 \in \Lambda$ merely rearranges the terms in the sum. This is the most obvious way to construct doubly periodic functions; similar methods can be used to construct functions on other quotients of domains.

To prove the absolute uniform convergence on compact subsets of such series, the following test is useful.

LEMMA 3.7. *Let D be a bounded open set in the complex plane and let $c > 1$ be constant. Suppose that $\psi(z,\omega)$, $\omega \in \Lambda$, is a function that is meromorphic in z for each ω and which satisfies* [3] *the condition* $(*)$

$$\psi(z, m\omega_1 + n\omega_2) = O((m^2+n^2)^{-c}) \text{ as } m^2+n^2 \to \infty$$

uniformly in z for z in D. Then the series $\sum_{\omega\in\Lambda}\psi(z,\omega)$, with finitely many terms which have poles in D deleted, is uniformly absolutely convergent in D.

PROOF. That only finitely many terms can have poles in D follows from (*). This condition on ψ means that there are constants A and B such that

$$|\psi(z, m\omega_1 + n\omega_2)| < B(m^2+n^2)^{-c}$$

whenever $m^2+n^2 > A$. To prove the lemma it suffices to show that, given any $\varepsilon > 0$, there is an integer N such that $S < \varepsilon$ for every finite sum $S = \sum |\psi(z, m\omega_1 + n\omega_2)|$ in which all the terms are distinct and each one of them has $m^2+n^2 \geq 2N^2$. Now S consists of eight subsums, a typical member of which consists of the terms for which $m \geq n \geq 0$. (There is some overlap between these sums, but that is harmless.) In this subsum we have $m \geq N$ and $\psi < Bm^{-2c}$, assuming as we may that $2N^2 > A$; and there are at most $m+1$ possible values of n for a given m. Thus

$$S \leq \sum_{m=N}^{\infty} Bm^{-2c}(m+1) < B_1 N^{2-2c}$$

for a suitable constant B_1, and this proves the lemma. □

We know from (3.1) that the simplest possible nonconstant doubly periodic function is one with a double pole at each point of Λ and no other poles. Suppose $f(z)$ is such a function. Then $f(z) - f(-z)$ is a doubly periodic function with no poles except perhaps simple ones at the points of Λ. Hence by the argument above, it must be constant, and since it is an odd function it must vanish. Thus $f(z)$ is even, and we can make it unique by imposing the normalization condition $f(z) = z^{-2} + O(z^2)$ near $z = 0$—it turns out to be convenient to force the constant term in this expansion to vanish rather than to assign the zeros of $f(z)$. There is such an $f(z)$—indeed it is the Weierstrass function $\wp(z)$—but we can't define it by the method at the start of this subsection because if $\varphi(z) = z^{-2}$, the series $\Phi(z)$ is not absolutely convergent. However, if $\varphi(z) = -2z^{-3}$, we can apply this method, and it gives $\wp'$, the derivative of the Weierstrass $\wp$-function. Define

$$\wp'(z;\Lambda) = \wp'(z;\omega_1,\omega_2) = -2\sum_{\omega\in\Lambda}\frac{1}{(z-\omega)^3}.$$

[3] The expression $f(z) = O(\varphi(z))$ means that $|f(z)| < C\varphi(z)$ for some constant C (independent of z) for all values of z in question.

Hence

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

THEOREM 3.8. *Let $P_1, ..., P_n$ and $Q_1, ..., Q_n$ be two sets of $n \geq 2$ points in the complex plane, possibly with repetitions, but such that no P_i is congruent to a Q_j modulo Λ. If $\sum P_i \equiv \sum Q_j \mod \Lambda$, then there exists a doubly periodic function $f(z)$ whose poles are the P_i and whose zeros are the Q_j with correct multiplicty, and $f(z)$ is unique up to multiplication by a nonzero constant.*

PROOF. There is an elementary (constructive) proof. Alternatively, one can apply the Riemann-Roch theorem to $\mathbb{C}/\Lambda$. □

The addition formula. Consider $\wp(z+z')$. It is a doubly periodic function of z, and therefore it is a rational function of $\wp$ and $\wp'$.

PROPOSITION 3.9. *There is the following formula:*

$$\wp(z+z') = \frac{1}{4}\left\{ \frac{\wp'(z) - \wp'(z')}{\wp(z) - \wp(z')} \right\}^2 - \wp(z) - \wp(z').$$

PROOF. Let $f(z)$ denote the difference between the left and the right sides. Its only possible poles (in D) are at 0, or $\pm z'$, and by examining the Laurent expansion of $f(z)$ near these points one sees that it has no pole at 0 or z, and at worst a simple pole at z'. Since it is doubly periodic, it must be constant, and since $f(0) = 0$, it must be identically zero. □

Eisenstein series. Write

$$G_k(\Lambda) = \sum_{\omega \in \Lambda, \omega \neq 0} \omega^{-2k}$$

and define $G_k(z) = G_k(z\mathbb{Z} + \mathbb{Z})$.

PROPOSITION 3.10. *The Eisenstein series $G_k(z)$, $k > 1$, converges to a holomorphic function on $\mathbb{H}$; it takes the value $2\zeta(2k)$ at infinity. (Here $\zeta(s) = \sum n^{-s}$, the usual zeta function.)*

PROOF. Apply Lemma 3.7 to see that $G_k(z)$ is a holomorphic function on $\mathbb{H}$. It remains to consider $G_k(z)$ as $z \to i\infty$ (remaining in D, the fundamental domain for $\Gamma(1)$). Because the series for $G_k(z)$ converges uniformly absolutely on D, $\lim_{z\to i\infty} G_k(z) = \sum \lim_{z \to i\infty} 1/(mz+n)^{2k}$. But $\lim_{z\to i\infty} 1/(mz+n)^{2k} = 0$ unless $m = 0$, and so

$$\lim_{z \to i\infty} G_k(z) = \sum_{n \in \mathbb{Z}, n \neq 0} 1/n^{2k} = 2\sum_{n \geq 1} 1/n^{2k} = 2\zeta(2k).$$

□

The field of doubly periodic functions.

PROPOSITION 3.11. *The field of doubly periodic functions is just* $\mathbb{C}(\wp(z), \wp'(z))$, *and*

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

where $g_2 = 60G_2$ *and* $g_3 = 140G_3$.

PROOF. To prove the second statement, define $f(z)$ to be the difference of the left and the right hand sides, and show (from its Laurent expansion) that it is holomorphic near 0 and take the value 0 there. Since it is doubly periodic and holomorphic elsewhere, this implies that it is zero. The proof of the first statement is omitted. □

Elliptic curves. Let k be a field of characteristic $\neq 2, 3$. By an *elliptic curve* over k, I shall mean a nonsingular projective curve E of genus one together with a point $0 \in E(k)$. From the Riemann-Roch theorem, one obtains regular functions x and y on E such that x has a double pole at 0 and y a triple pole at 0, and neither has any other poles. Again from the Riemann-Roch theorem applied to the divisor $6 \cdot 0$, one finds that there is a relation between 1, x, x^2, x^3, y, y^2, xy, which can be put in the form

$$y^2 = 4x^3 - ax - b.$$

The fact that E is nonsingular implies that $\Delta \stackrel{\text{df}}{=} a^3 - 27b^2 \neq 0$. Thus E is isomorphic to the projective curve defined by the equation,

$$Y^2Z = 4X^3 - aXZ^2 - bZ^3,$$

and every equation of this form (with $\Delta \neq 0$) defines an elliptic curve. Define

$$j(E) = 1728a^3/\Delta.$$

Then two elliptic curves E and E' are isomorphic if and only if $j(E) = j(E')$. If E is an elliptic curve over $\mathbb{C}$, then $E(\mathbb{C})$ has a natural complex structure—it is a Riemann surface. (See Math 679 for proofs of these, and other statements, about elliptic curves.)

An elliptic curve has a unique group structure (defined by regular maps) having 0 as its zero.

The elliptic curve $E(\Lambda)$. Let Λ be a lattice in $\mathbb{C}$. We have seen that

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

Let $E(\Lambda)$ be the projective curve defined by the equation:

$$Y^2Z = 4X^3 - g_2XZ^2 - g_3Z^3.$$

PROPOSITION 3.12. *The curve* $E(\Lambda)$ *is an elliptic curve (i.e.,* $\Delta \neq 0$*), and the map*

$$\mathbb{C}/\Lambda \to E(\Lambda), \quad z \mapsto (\wp(z) : \wp'(z) : 1), \quad 0 \mapsto (0 : 1 : 0)$$

is an isomorphism of Riemann surfaces. Every elliptic curve E *is isomorphic to* $E(\Lambda)$ *for some* Λ.

PROOF. There are direct proofs of this result, but we shall see in the next section that $z \mapsto \Delta(z, \mathbb{Z}+\mathbb{Z})$ is a modular function for $\Gamma(1)$ with weight 12 having no zeros in $\mathbb{H}$, and that $z \mapsto j(z\mathbb{Z}+\mathbb{Z})$ is a modular function and defines a bijection $\Gamma(1)\backslash\mathbb{H} \to \mathbb{C}$ (therefore every j equals $j(\Lambda)$ for some lattice $\mathbb{Z}z + \mathbb{Z}$, $z \in \mathbb{H}$). □

The addition formula shows that the map in the proposition is a homomorphism.

PROPOSITION 3.13. *There are natural equivalences between the following categories:*

(a) *Objects: Elliptic curves E over $\mathbb{C}$.*
Morphisms: Regular maps $E \to E'$ that are homomorphisms.
(b) *Objects: Riemann surfaces E of genus* 1 *together with a point 0.*
Morphisms: Holomorphic maps $E \to E'$ sending 0 to $0'$.
(c) *Objects: Lattices $\Lambda \subset \mathbb{C}$.*
Morphisms: $\mathrm{Hom}(\Lambda, \Lambda') = \{\alpha \in \mathbb{C} \mid \alpha\Lambda \subset \Lambda'\}$.

PROOF. The functor $c \to b$ is $\Lambda \mapsto \mathbb{C}/\Lambda$. The functor $a \to b$ is $(E, 0) \mapsto (E(\mathbb{C}), 0)$, regarded as a pointed Riemann surface. □

4. Modular Functions and Modular Forms

Modular functions. Let Γ be a subgroup of finite index in $\Gamma(1)$. A *modular function for* Γ is a meromorphic function on the compact Riemann surface $\Gamma\backslash\mathbb{H}^*$. We often regard it as a meromorphic function on $\mathbb{H}^*$ invariant under Γ. Thus, from this point of view, a modular function f for Γ is a function on $\mathbb{H}$ satisfying the following conditions:

(a) $f(z)$ is invariant under Γ, i.e., $f(\gamma z) = f(z)$ for all $\gamma \in \Gamma$;
(b) $f(z)$ is meromorphic in $\mathbb{H}$;
(c) $f(z)$ is meromorphic at the cusps.

For the cusp $i\infty$, the last condition means the following: the subgroup of $\Gamma(1)$ fixing $i\infty$ is generated by $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$—it is free abelian group of rank 1; the subgroup of Γ fixing $i\infty$ is a subgroup of finite index in $<T>$, and it therefore is generated by $\begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}$ for some $h \in \mathbb{N}$, (h is called the *width* of the cusp); as $f(z)$ is invariant under $\begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}$, $f(z+h) = f(z)$, and so $f(z)$ can be expressed as a function $f^*(q)$ of the variable $q = \exp(2\pi i z/h)$; this function $f^*(q)$ is defined on a punctured disk, $0 < |q| < \varepsilon$, and for f to be meromorphic at $i\infty$ means f^* is meromorphic at $q = 0$.

For a cusp $\tau \neq i\infty$, the condition means the following: we know there is an element $\sigma \in \Gamma(1)$ such that $\tau = \sigma(i\infty)$; the function $z \mapsto f(\sigma z)$ is invariant under $\sigma\Gamma\sigma^{-1}$, and $f(\sigma z)$ is required to be meromorphic at $i\infty$ in the above sense.

Of course (c) has to be checked only for a set of representatives of the Γ-equivalence classes of cusps (which will be finite).

Recall that a function $f(z)$ that is holomorphic in a neighbourhood of a point $a \in \mathbb{C}$ (except possibly at a) is holomorphic at a if and only if $f(z)$ is bounded in a neighbourhood of a. It follows that $f(z)$ has a pole at a, and therefore defines a meromorphic function in a neighbourhood of a, if and only if $(z-a)^n f(z)$ is bounded near a for some n, i.e., if $f(z) = O((z-a)^{-n})$ near a. When we apply this remark to a modular function, we see that $f(z)$ is meromorphic at $i\infty$ if and only if $f^*(q) = O(q^{-n})$ for some n as $q \to 0$, i.e., if and only if, for some $A > 0$, $e^{Aiz} \cdot f(z)$ is bounded as $z \to i\infty$.

Example 4.1. As $\Gamma(1)$ is generated by S and T, to check condition (a) it suffices to verify that

$$f(-1/z) = f(z), \quad f(z+1) = f(z).$$

The second equation implies that $f = f^*(q)$, $q = \exp(2\pi i z)$, and condition (c) says that

$$f^*(q) = \sum_{n \geq -N_0} a_i q^i.$$

Example 4.2. Consider $\Gamma(2)$. Then $\Gamma(2)$ is of index 6 in $\Gamma(1)$. It is possible to find a set of generators for $\Gamma(2)$ just as we found a set of generators for $\Gamma(1)$, and again it suffices to check condition (a) for the generators. There are three inequivalent cusps, namely, $i\infty$, $S(i\infty) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} = 0$, and $TS(i\infty) = 1$. Note that $S(0) = i\infty$. The stabilizer of $i\infty$ in $\Gamma(2)$ is generated by $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$, and so $f(z) = f^*(q)$, $q = \exp(2\pi iz/2)$, and for $f(z)$ to be meromorphic at $i\infty$ means f^* is meromorphic at 0. For $f(z)$ to be meromorphic at 0 means that $f(Sz) = f(-1/z)$ is meromorphic at $i\infty$, and for $f(z)$ to be meromorphic at 1 means that $f(1-\frac{1}{z})$ is meromorphic at $i\infty$.

PROPOSITION 4.3. *There exists a unique modular function J for $\Gamma(1)$ which is holomorphic except at $i\infty$, where it has a simple pole, and which takes the values*

$$J(i) = 1,\ J(\rho) = 0.$$

PROOF. From Proposition 2.21 we know there is an isomorphism of Riemann surfaces $f : \Gamma(1)\backslash\mathbb{H}^* \to S$ (Riemann sphere). Write a, b, c for the images of ρ, i, ∞. Then there exists a (unique) linear fractional transformation $S \to S$ sending a, b, c to 0, 1, ∞, and on composing f with it we obtain a function J satisfying the correct conditions.

If g is a second function satisfying the same conditions, then $g \circ f^{-1}$ is an automorphism of the Riemann sphere, and so it is a linear fractional transformation. Since it fixes 0, 1, ∞ it must be the identity map. □

REMARK 4.4. Let $j(z) = 1728g_2^3/\Delta$, as on Section 3. Then $j(z)$ is invariant under $\Gamma(1)$ because g_2^3 and Δ are both modular forms of weight 12 (we give all the details for this example later). It is holomorphic on $\mathbb{H}$ because both of g_2^3 and Δ are holomorphic on $\mathbb{H}$, and Δ has no zeros on $\mathbb{H}$. Because Δ has a simple zero at ∞, j has a simple pole at ∞ . Therefore $j(z)$ has valence one, and it defines an isomorphism from $\Gamma\backslash\mathbb{H}^*$ onto S (the Riemann sphere). In fact, $j(z) = 1728J(z)$.

Modular forms. Let Γ be a subgroup of finite index in $\Gamma(1)$.

DEFINITION 4.5. A *modular form for* Γ *of weight* 2k is a function on $\mathbb{H}$ such that:

(a) $f(\gamma z) = (cz+d)^{2k} \cdot f(z)$, all $z \in \mathbb{H}$;
(b) $f(z)$ is holomorphic in $\mathbb{H}$;
(c) $f(z)$ is holomorphic at the cusps of Γ.

A modular form is a *cusp form* if it is zero at the cusps.

For example, for the cusp $i\infty$, this last condition means the following: let h be the width of $i\infty$ as a cusp for Γ; then implies that $f(z+h) = f(z)$, and so $f(z) = f^*(q)$ for some function f^* on a punctured disk; f^* is required to be holomorphic at $q = 0$.

When $f(z)$ is zero at every cusp, it is called a *cusp form*. Occasionally we shall refer to a function satisfying only (4.5a) as being *weakly modular of weight* 2k, and a function satisfying (4.5a,b,c) with "holomorphic" replaced by "meromorphic" as being a *meromorphic modular form of weight* 2k. Thus a meromorphic modular form of weight 0 is a modular function.

As our first examples of modular forms, we have the Eisenstein series. Let $\mathcal{L}$ be the set of lattices in $\mathbb{C}$, and write $\Lambda(\omega_1, \omega_2)$ for the lattice $\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$ generated by independent elements ω_1, ω_2 with $\Im(\omega_1/\omega_2) > 0$. Note that $\Lambda(\omega_1', \omega_2') = \Lambda(\omega_1, \omega_2)$ if and only if

$$\begin{pmatrix} \omega_1' \\ \omega_2' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}, \text{ some } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) = \Gamma(1).$$

LEMMA 4.6. *Let $F : \mathcal{L} \to \mathbb{C}$ be a function of weight $2k$, i.e., such that $F(\lambda\Lambda) = \lambda^{-2k} \cdot F(\Lambda)$ for $\lambda \in \mathbb{C}^\times$. Then $f(z) \stackrel{\mathrm{df}}{=} F(\Lambda(z,1))$ is a weakly modular form on $\mathbb{H}$ of weight $2k$ and $F \mapsto f$ is a bijection from the functions of weight $2k$ on $\mathcal{L}$ to the weakly modular forms of weight $2k$ on $\mathbb{H}$.*

PROOF. Write $F(\omega_1, \omega_2)$ for the value of F at the lattice $\Lambda(\omega_1, \omega_2)$. Then because F is of weight $2k$, we have

$$F(\lambda\omega_1, \lambda\omega_2) = \lambda^{-2k} \cdot F(\omega_1, \omega_2),\ \lambda \in \mathbb{C}^\times,$$

and, because $F(\omega_1, \omega_2)$ depends only on $\Lambda(\omega_1, \omega_2)$, it is invariant under the action of $\mathrm{SL}_2(\mathbb{Z})$:

$$F(a\omega_1 + b\omega_2, c\omega_1 + d\omega_2) = F(\omega_1, \omega_2), \text{ all } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z}), \quad (*).$$

The first equation shows that $\omega_2^{2k} \cdot F(\omega_1, \omega_2)$ is invariant under $(\omega_1, \omega_2) \mapsto (\lambda\omega_1, \lambda\omega_2)$, $\lambda \in \mathbb{C}^\times$, and so depends only on the ratio ω_1/ω_2; thus there is a function $f(z)$ such that

$$F(\omega_1, \omega_2) = \omega_2^{-2k} \cdot f(\omega_1/\omega_2), \quad (**).$$

When expressed in terms of f, (*) becomes

$$(c\omega_1 + d\omega_2)^{-2k} \cdot f(a\omega_1 + b\omega_2/c\omega_1 + d\omega_2) = \omega_2^{-2k} \cdot f(\omega_1/\omega_2),$$

or

$$(cz + d)^{-2k} \cdot f(az + b/cz + d) = f(z).$$

This shows that f is weakly modular. Conversely, given a weakly modular f, define F by the formula (**). □

PROPOSITION 4.7. *The Eisenstein series $G_k(z)$, $k > 1$, is a modular form of weight $2k$ for $\Gamma(1)$ which takes the value $2\zeta(2k)$ at infinity.*

PROOF. Recall that we defined $G_k(\Lambda) = \sum_{\omega\in\Lambda,\omega\neq 0} 1/\omega^{2k}$. Clearly, $G_k(\lambda\Lambda) = \lambda^{-2k} G_k(\Lambda)$, and therefore $G_k(z) =_{df} G_k(\Lambda(z,1)) = \sum_{(m,n)\neq(0,0)} 1/(mz+n)^{2k}$ is weakly modular. That it is holomorphic on $\mathbb{H}$ and takes the value $2\zeta(2k)$ at $i\infty$ is proved in Proposition 3.10. □

Modular forms as k-fold differentials. The definition of modular form may seem strange, but we have seen that such functions arise naturally in the theory of elliptic functions. Here we give another explanation of the definition. For the experts, we shall show later that the modular forms of a fixed weight $2k$ are the sections of a line bundle on $\Gamma\backslash\mathbb{H}^*$.

REMARK 4.8. Consider a differential $\omega = f(z) \cdot dz$ on $\mathbb{H}$, where $f(z)$ is a meromorphic function. Under what conditions on f is ω invariant under the action of Γ? Let $\gamma(z) = \frac{az+b}{cz+d}$; then

$$\gamma^*\omega = f(\gamma z) \cdot d\frac{az+b}{cz+d} = f(\gamma z) \cdot \frac{(a(cz+d) - c(az+b))}{(cz+d)^2} \cdot dz = f(\gamma z) \cdot (cz+d)^{-2} \cdot dz.$$

Thus ω is invariant if and only if $f(z)$ is a meromorphic differential form of weight 2. We have one-to-one correspondences between the following sets:

{meromorphic modular forms of weight 2 on $\mathbb{H}$ for Γ}

{meromorphic differential forms on $\mathbb{H}^*$ invariant under the action of Γ}

{meromorphic differential forms on $\Gamma\backslash\mathbb{H}^*$}.

There is a notion of a *k-fold differential form* on a Riemann surface. Locally it can be written $\omega = f(z) \cdot (dz)^k$, and if $w = w(z)$, then

$$w^*\omega = f(w(z)) \cdot (dw(z))^k = f(w(z)) \cdot w'(z)^k \cdot (dz)^k.$$

Then modular forms of weight 2k correspond to Γ-invariant k-fold differential forms on $\mathbb{H}^*$, and hence to meromorphic k-fold differential forms on $\Gamma\backslash\mathbb{H}^*$. Warning: these statements don't (quite) hold with meromorphic replaced with holomorphic.

We say that $\omega = f(z) \cdot (dz)^k$ has a *zero or pole of order m* at $z = 0$ according as $f(z)$ has a zero or pole of order m at $z = 0$. This definition is independent of the choice of the local coordinate near the point in question on the Riemann surface.

The dimension of the space of modular forms. For a subgroup of finite index in $\Gamma(1)$, we write $\mathcal{M}_k(\Gamma)$ for the space of modular forms of weight $2k$ for Γ, and $\mathcal{S}_k(\Gamma)$ for the subspace of cusp forms of weight $2k$. They are vector spaces over $\mathbb{C}$, and we shall use the Riemann-Roch theorem to compute their dimensions.

Note that $\mathcal{M}_0(\Gamma)$ consists of modular functions that are holomorphic on $\mathbb{H}$ and at the cusps, and therefore define holomorphic functions on $\Gamma\backslash\mathbb{H}^*$. Because $\Gamma\backslash\mathbb{H}^*$ is compact, such a function is constant, and so $\mathcal{M}_0(\Gamma) = \mathbb{C}$. The product of a modular form of weight k with a modular form of weight ℓ is a modular form of weight $k + \ell$. Therefore,

$$\mathcal{M}(\Gamma) \stackrel{\text{df}}{=} \oplus \mathcal{M}_k(\Gamma)$$

is a graded ring. The next theorem gives us the dimensions of the homogeneous pieces.

THEOREM 4.9. *The dimension of $\mathcal{M}_k(\Gamma)$ is given by:*

$$\dim(\mathcal{M}_k(\Gamma)) = \begin{cases} 0 & \text{if } k \leq -1 \\ 1 & \text{if } k = 0 \\ (2k-1)(g-1) + \nu_\infty k + \sum_P [k(1 - \frac{1}{e_P})] & \text{if } k \geq 1 \end{cases}$$

where g is the genus of $X(\Gamma)$ ($\stackrel{\text{df}}{=} \Gamma\backslash\mathbb{H}^*$);

ν_∞ is the number of inequivalent cusps;

the last sum is over a set of representatives for the the elliptic points P of Γ;

e_P is the order of the stabilizer of P in the image $\bar{\Gamma}$ of Γ in $\Gamma(1)/\{\pm I\}$);

$[k(1 - 1/e_P)]$ is the integer part of $k(1 - 1/e_P)$.

We prove the result by applying the Riemann-Roch theorem to the compact Riemann surface $\Gamma\backslash\mathbb{H}^*$, but first we need to examine the relation between the zeros and poles of a Γ-invariant k -fold differential form on $\mathbb{H}^*$ and the zeros and poles of the corresponding modular form on $\Gamma\backslash\mathbb{H}^*$. It will be helpful to consider first a simple example.

EXAMPLE 4.10. Let D be the unit disk, and consider the map $w : D \to D$, $z \mapsto z^e$. Let $Q \mapsto P$. If $Q \neq 0$, then the map is a local isomorphism, and so there is no difficulty. Thus we suppose that P and Q are both zero.

First suppose that f is a function on D (the target disk), and let $f^* = f \circ w$. If f has a zero of order m (regarded as function of w), then f^* has a zero of order em, for if $f(w) = aw^m +$ terms of higher degree, then $f(z^e) = az^{em} +$ terms of higher degree. Thus

$$\mathrm{ord}_Q(f^*) = e \cdot \mathrm{ord}_P(f).$$

Now consider a k-fold differential form ω on D, and let $\omega^* = w^*(\omega)$. Then $\omega = f(z) \cdot (dz)^k$ for some $f(z)$, and

$$\omega^* = f(z^e) \cdot (dz^e)^k = f(z^e) \cdot (ez^{e-1} \cdot dz)^k = e^k \cdot f(z^e) \cdot z^{k(e-1)} \cdot (dz)^k.$$

Thus

$$\mathrm{ord}_Q(\omega^*) = e\,\mathrm{ord}_P(\omega) + k(e-1).$$

LEMMA 4.11. *Let f be a (meromorphic) modular form of weight $2k$, and let ω be the corresponding k-fold differential form on $\Gamma\backslash\mathbb{H}^*$. Let $Q \in \mathbb{H}^*$ map to $P \in \Gamma\backslash\mathbb{H}^*$.*

(a) *If Q is an elliptic point with multiplicity e, then*

$$\mathrm{ord}_Q(f) = e\,\mathrm{ord}_P(\omega) + k(e-1).$$

(b) *If Q is a cusp, then*

$$\mathrm{ord}_Q(f) = \mathrm{ord}_P(\omega) + k.$$

(c) *For the remaining points,*

$$\mathrm{ord}_Q(f) = \mathrm{ord}_P(\omega).$$

PROOF. Let p be the quotient map $\mathbb{H} \to \Gamma\backslash\mathbb{H}$.

(a) We defined the complex structure near P so that, for appropriate neighbourhoods V of Q and U of P, there is a commutative diagram:

$$\begin{array}{ccc} V & \xrightarrow[\approx]{Q\mapsto 0} & D \\ \downarrow{\scriptstyle p} & & \downarrow{\scriptstyle z\mapsto z^e} \\ U & \xrightarrow[\approx]{P\mapsto 0} & D \end{array}$$

Thus this case is isomorphic to that considered in the example.

(b) Consider the map $q : \mathbb{H} \to$ (punctured disk), $q(z) = \exp(2\pi i z/h)$, and let $\omega^* = g(q) \cdot (dq)^k$ be a k-fold differential form on the punctured disk. Then $dq = (2\pi i/h) \cdot q \cdot dz$, and so the inverse image of ω^* on $\mathbb{H}$ is

$$\omega = (\mathrm{cnst}) \cdot g(q(z)) \cdot q(z)^k \cdot (dz)^k,$$

and so ω^* corresponds to the modular form $f(z) = (\mathrm{cnst}) \cdot g(q(z)) \cdot q(z)^k$. Thus $f^*(q) = g(q) \cdot q^k$, which gives our formula.

(iii) In this case, p is a local isomorphism near Q and P, and so there is nothing to prove. $\square$

We now prove the theorem. Let $f \in \mathcal{M}_k(\Gamma)$, and let ω be the corresponding k-fold differential on $\Gamma\backslash\mathbb{H}^*$. Because f is holomorphic, we must have

$$\begin{aligned} e\,\mathrm{ord}_P(\omega) + k(e-1) &= \mathrm{ord}_Q(f) \geq 0 \text{ at the image of an elliptic point;} \\ \mathrm{ord}_P(\omega) + k &\geq \mathrm{ord}_Q(f) \geq 0 \text{ at the image of a cusp;} \\ \mathrm{ord}_P(\omega) &= \mathrm{ord}_Q(f) \geq 0 \text{ at the remaining cusps.} \end{aligned}$$

Fix a k-fold differential ω_0, and write $\omega = h \cdot \omega_0$. Then

$$\begin{aligned} \mathrm{ord}_P(h) + ord_P(\omega_0) + k(1-1/e) &\geq 0 \quad \text{at the image of an elliptic point;} \\ \mathrm{ord}_P(h) + ord_P(\omega_0) + k &\geq 0 \quad \text{at the image of a cusp;} \\ \mathrm{ord}_P(h) + ord_P(\omega_0) &\geq 0 \quad \text{at the remaining points.} \end{aligned}$$

On combining these inequalities, we find that

$$div(h) + D \geq 0,$$

where

$$D = div(\omega_0) + \sum k \cdot P_i + \sum [k(1-1/e_i)] \cdot P_i$$

(the first sum is over the images of the cusps, and the second sum is over the images of the elliptic points). As we noted in Corollary 1.21, the degree of the divisor of a 1-fold differential form is $2g-2$; hence that of a k-fold differential form is $k(2g-2)$. Thus the degree of D is

$$k(2g-2) + \nu_\infty \cdot k + \sum_P [k(1-1/e_P)].$$

Now the Riemann-Roch Theorem (1.22) tells us that the space of h's has dimension

$$1 - g + k(2g-2) + \nu_\infty \cdot k + \sum_P [k(1-1/e_P)]$$

for $k \geq 1$. As the h's are in one-to-one correspondence with the holomorphic modular forms of weight $2k$, this proves the theorem in this case. For $k = 0$, we have already noted that modular forms are constant, and for $k < 0$ it is easy to see that there can be no modular forms.

Zeros of modular forms. Lemma 4.11 allows us to count the number of zero and poles of a meromorphic differential form.

PROPOSITION 4.12. *Let f be a (meromorphic) modular form of weight $2k$; then*

$$\sum (\mathrm{ord}_Q(f)/e_Q - k(1-1/e_Q)) = k(2g-2) + k \cdot \nu_\infty$$

where the sum is over a set of representatives for the points in $\Gamma\backslash\mathbb{H}^*$, ν_∞ is the number of inequivalent cusps, and e_Q is the ramification index of Q over $p(Q)$ if $Q \in \mathbb{H}$ and it is 1 if Q is a cusp.

PROOF. Let ω be the associated k-fold differential form on $\Gamma\backslash\mathbb{H}^*$. We showed above that: $\mathrm{ord}_Q(f)/e_Q = \mathrm{ord}_P(\omega) + k(1 - 1/e_Q)$ for Q an elliptic point for Γ; $\mathrm{ord}_Q(f) = \mathrm{ord}_P(\omega) - k$ for Q a cusp; $\mathrm{ord}_Q(f) = \mathrm{ord}_P(\omega)$ at the remaining points. On summing these equations, we find that

$$\sum \mathrm{ord}_Q(f)/e_Q - k(1 - 1/e_Q) = \deg(div(\omega)) + k \cdot \nu_\infty,$$

and we noted above that $\deg(div(\omega)) = k(2g - 2)$. □

EXAMPLE 4.13. When $\Gamma = \Gamma(1)$, this becomes

$$\mathrm{ord}_{i\infty}(f) + \frac{1}{2}\mathrm{ord}_i(f) + \frac{1}{3}\mathrm{ord}_\rho(f) + \sum \mathrm{ord}_Q(f) = -2k + k + \frac{1}{2}k + \frac{2}{3}k = \frac{k}{6}.$$

Here $i\infty$, i, ρ are points in $\mathbb{H}^*$, and the sum $\sum$ is over the remaining points in a fundamental domain.

Modular forms for $\Gamma(1)$. We now describe all the modular forms for $\Gamma(1)$.

EXAMPLE 4.14. On applying Theorem 4.9 to the full modular group $\Gamma(1)$, we obtain the following result: $\mathcal{M}_k = 0$ for $k < 0$, dim $\mathcal{M}_0 = 1$, and

$$\dim\ \mathcal{M}_k = 1 - k + [\frac{k}{2}] + [\frac{2k}{3}], \quad k > 1.$$

Thus

$$\begin{array}{rcllllllll} k & = & 1 & 2 & 3 & 4 & 5 & 6 & 7 & \ldots; \\ \dim \mathcal{M}_k & = & 0 & 1 & 1 & 1 & 1 & 2 & 1 & \ldots. \end{array}$$

In fact, when k is increased by 6, the dimension increases by 1. Thus we have

(a) $\mathcal{M}_k = 0$ for $k < 0$;
(b) $\dim(\mathcal{M}_k) = [k/6]$ if $k \equiv 1 \bmod 6$; $[k/6] + 1$ otherwise; $k \geq 0$.

EXAMPLE 4.15. On applying (4.13) to the Eisenstein series G_k, $k > 1$, one obtains the following result:

$k = 2$: G_2 has a simple zero at $z = \rho$, and no other zeros.

$k = 3$: G_3 has a simple zero at $z = i$, and no other zeros.

$k = 6$: because Δ has no zeros in $\mathbb{H}$, it has a simple zero at ∞.

There is a geometric explanation for these statements. Let $\Lambda = \Lambda(i, 1)$; then the torus $\mathbb{C}/\Lambda$ has complex multiplication by i, and so the elliptic curve

$$Y^2 = 4X^3 - g_2(\Lambda)X - g_3(\Lambda)$$

has complex multiplication by i; clearly the curve

$$Y^2 = X^3 + X$$

has complex multiplication by i (and up to isogeny, it is the only such curve); this suggests that $g_3(\Lambda) = 0$. Similarly, $G_2(\Lambda) = 0$ "because" $Y^2 = X^3 + 1$ has complex multiplication by $\sqrt[3]{1}$. Finally, if Δ had no zero at ∞, the family of elliptic curves

$$Y^2 = 4X^3 - g_2(\Lambda)X - g_3(\Lambda)$$

over $\Gamma(1)\backslash\mathbb{H}$ would extend to a smooth family over $\Gamma(1)\backslash\mathbb{H}^*$, and this is not possible for topological reasons (its cohomology groups would give a nonconstant local system on $\Gamma(1)\backslash\mathbb{H}^*$, but the Riemann sphere is simply connected, and so admits no such system).

PROPOSITION 4.16. (a) *For $k < 0$, and $k = 1$, $\mathcal{M}_k = 0$.*

(b) *For $k = 0, 2, 3, 4, 5$, $\mathcal{M}_k$ is a space of dimension 1, admitting as basis 1, G_2, G_3, G_4, G_5 respectively; moreover $\mathcal{S}_k(\Gamma) = 0$ for $0 \leq k \leq 5$.*

(c) *Multiplication by Δ defines an isomorphism of $\mathcal{M}_{k-6}$ onto $\mathcal{S}_k$.*

(d) *The graded k-algebra $\oplus\mathcal{M}_k = \mathbb{C}[G_2, G_3]$ with G_2 and G_3 of weights 2 and 3 respectively.*

PROOF. (a) See (4.14).

(b) Since the spaces are one-dimensional, and no G_k is identically zero, this is obvious.

(c) Certainly $f \mapsto f\Delta$ is a homomorphism $\mathcal{M}_{k-6} \to \mathcal{S}_k$. But if $f \in \mathcal{S}_k$, then $f/\Delta \in \mathcal{M}_{k-6}$ because Δ has only a simple zero at $i\infty$ and f has a zero there. Now $f \mapsto f/\Delta$ is inverse to $f \mapsto f\Delta$.

(d) We have to show that $\{G_2^m \cdot G_3^n \mid 2m + 3n = k,\ m \in \mathbb{N},\ n \in \mathbb{N}\}$ forms a basis for $\mathcal{M}_k$. We first show, by induction on k, that this set generates $\mathcal{M}_k$. For $k \leq 3$, we have already noted it. Choose a pair $m \geq 0$ and $n \geq 0$ such that $2m + 3n = k$ (this is always possible for $k \geq 2$). The modular form $g = G_2^m \cdot G_3^n$ is not zero at infinity. If $f \in \mathcal{M}_k$, then $f - \frac{f(\infty)}{g(\infty)} g$ is zero at infinity, and so is a cusp form. Therefore, it can be written $\Delta \cdot h$ with $h \in \mathcal{M}_{k-6}$, and we can apply the induction hypothesis.

Thus $\mathbb{C}[G_2, G_3] \to \oplus\mathcal{M}_k$ is surjective, and we want to show that it is injective. If not, the modular function G_2^3/G_3^2 satisfies an algebraic equation over $\mathbb{C}$, and so is constant. But $G_2(\rho) = 0 \neq G_3(\rho)$ whereas $G_2(i) \neq 0 = G_3(i)$. $\square$

REMARK 4.17. We have verified all the assertions in (4.3).

The Fourier coefficients of the Eisenstein series for $\Gamma(1)$. For future use, we compute the coefficients in the expansion $G_k(z) = \sum a_n q^n$.

The Bernoulli numbers B_k. They are defined by the formal power series expansion:

$$\frac{x}{e^x - 1} = 1 - \frac{x}{2} + \sum_{k=1}^{\infty} (-1)^{k+1} B_k \frac{x^{2k}}{(2k)!}.$$

Thus $B_1 = 1/6$; $B_2 = 1/30$; ... ; $B_{14} = 23749461029/870$; ... Note that they are all rational numbers.

PROPOSITION 4.18. *For any integers $k \geq 1$,*

$$\zeta(2k) = \frac{2^{2k-1}}{(2k)!} B_k \pi^{2k}.$$

PROOF. Recall that (by definition)

$$\cos(z) = \frac{e^{iz} + e^{-iz}}{2}, \quad \sin(z) = \frac{e^{iz} - e^{-iz}}{2i}.$$

Therefore,

$$\cot(z) = i\frac{e^{iz} + e^{-iz}}{e^{iz} - e^{-iz}} = i\frac{e^{2iz} + 1}{e^{2iz} - 1} = i + \frac{2i}{e^{2iz} - 1}.$$

On replacing x with $2iz$ in the definition of the Bernoulli numbers, we find that

$$z\cot(z) = 1 - \sum_{k=1}^{\infty} B_k \frac{2^{2k}z^{2k}}{(2k)!} \qquad (*).$$

There is a standard formula

$$\sin(z) = z\prod_{n=1}^{\infty}(1-\frac{z^2}{n^2\pi^2})$$

(see Cartan 1963, V.3.3). On forming the logarithmic derivative of this (i.e., forming $d\log(f) = f'/f$) and multiplying by z, we find that

$$\begin{aligned} z\cot z &= 1-\sum_{n=1}^{\infty}\frac{2z^2/n^2\pi^2}{1-z^2/n^2\pi^2} \\ &= 1+2\sum_{n=1}^{\infty}\frac{1}{1-n^2\pi^2/z^2} \\ &= 1+2\sum_{n=1}^{\infty}\sum_{k=1}^{\infty}\frac{z^{2k}}{n^{2k}\pi^{2k}} \\ &= 1+2\sum_{k=1}^{\infty}\left(\sum_{n=1}^{\infty}n^{-2k}\right)\frac{z^{2k}}{\pi^{2k}}. \end{aligned}$$

On comparing this formula with $(*)$, we obtain the result. $\square$

For example, $\zeta(2) = \frac{\pi^2}{2\times 3}$, $\zeta(4) = \frac{\pi^4}{2\times 3^2\times 5}$, $\zeta(6) = \frac{\pi^6}{3^3\times 5\times 7}$, ...

REMARK 4.19. Until a few years ago, when Apery showed that $\zeta(3)$ is irrational, nothing was known about the values of ζ at the odd positive integers.

The Fourier coefficients of G_k. For any integer n and number k, we write

$$\sigma_k(n) = \sum_{d|n} d^k.$$

PROPOSITION 4.20. *For any integer $k \geq 2$,*

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

PROOF. In the above proof, we showed above that

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

and so (replace z with πz and divide by z)

$$\pi\cot(\pi z) = \frac{1}{z} + 2\sum_{n=1}^{\infty}\frac{z}{z^2-n^2} = \frac{1}{z}+\sum_{n=1}^{\infty}\left(\frac{1}{z+n}+\frac{1}{z-n}\right).$$

Moreover, we showed that

$$\cot(z) = i + \frac{2i}{e^{2iz}-1},$$

and so

$$\begin{aligned}\pi\cot(\pi z) &= \pi i - \frac{2\pi i}{1-q}\\ &= \pi i - 2\pi i\sum_{n=1}^{\infty} q^n\end{aligned}$$

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

$$\frac{1}{z}+\sum_{n=1}^{\infty}\left(\frac{1}{z+n}+\frac{1}{z-n}\right) = \pi i - 2\pi i\sum_{n=1}^{\infty} q^n.$$

The $(k-1)^{th}$ derivative of this $(k\geq 2)$ is

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

Now

$$\begin{aligned}G_k(z) &\stackrel{\text{df}}{=} \sum_{(n,m)\neq(0,0)}\frac{1}{(nz+m)^{2k}}\\ &= 2\zeta(2k)+2\sum_{n=1}^{\infty}\sum_{m\in\mathbb{Z}}\frac{1}{(nz+m)^{2k}}\\ &= 2\zeta(2k)+\frac{2(-2\pi i)^{2k}}{(2k-1)!}\sum_{n=1}^{\infty}\sum_{a=1}^{\infty} a^{2k-1}\cdot q^{an}\\ &= 2\zeta(2k)+\frac{2(2\pi i)^{2k}}{(2k-1)!}\sum_{n=1}^{\infty}\sigma_{2k-1}(a)\cdot q^n.\end{aligned}$$

□

The expansion of Δ and j. Recall that

$$\Delta \stackrel{\text{df}}{=} g_2^3 - 27g_3^2.$$

From the above expansions of G_2 and G_3, one finds that

$$\Delta = (2\pi)^{12}\cdot(q-24q^2+252q^3-1472q^4+\cdots)$$

THEOREM 4.21 (Jacobi). $\Delta = (2\pi)^{12}q\prod_{n=1}^{\infty}(1-q^n)^{24}$, $q=e^{2\pi i z}$.

PROOF. Let $f(q) = q\prod_{n=1}^{\infty}(1-q^n)^{24}$. The space of cusp forms of weight 12 has dimension 1. Therefore, if we show that $f(-1/z) = z^{12}f(z)$, then f will be a multiple of Δ. It is possible to prove by an elementary argument (due to Hurwitz), that $f(-1/z)$ and $z^{12}f(z)$ have the same logarithmic derivative; therefore

$$f(-1/z) = Cz^{12}\cdot f(z),$$

some C. Put $z=1$ to see $C=1$. See Serre 1970, VII.4.4, for the details. □

Write $q\prod(1-q^n)^{24} = \sum_{n=1}^{\infty}\tau(n)\cdot q^n$. The function $n\mapsto\tau(n)$ was studied by Ramanujan, and is called the *Ramanujan τ-function.* We have

$$\tau(1)=1,\ \tau(2)=-24,\ \ldots,\ \tau(12)=-370944,\ldots.$$

Evidently each $\tau(n) \in \mathbb{Z}$. Ramanujan made a number of interesting conjectures about $\tau(n)$, some of which, as we shall see, have been proved.

Recall that $j(z) = \frac{1728 g_2^3}{\Delta}$, $\Delta = g_2^3 - 27g_3^2$, $g_2 = 60G_2$, $g_3 = 140G_3$.

THEOREM 4.22. *The function*

$$j(z) = \frac{1}{q} + 744 + 196\,884q + 21\,493\,760q^2 + c(3)q^3 + c(4)q^4 + \cdots, \quad q = e^{2\pi i z},$$

where $c(n) \in \mathbb{Z}$ *for all* n.

PROOF. Immediate consequence of the definition and the above calculations. $\square$

The size of the coefficients of a cusp form. Let $f(z) = \sum a_n q^n$ be a cusp form of weight $2k \geq 2$ for some congruence subgroup of $\mathrm{SL}_2(\mathbb{Z})$. For various reasons, for example, in order to obtain estimates of the number of times an integer can be represented by a quadratic form, one is interested in $|a_n|$.

Hecke showed that $a_n = O(n^k)$—the proof is quite easy (see Serre 1970, VII.4.3, for the case of $\Gamma(1)$). Various authors improved on this—for example, Selberg showed in 1965 that $a_n = O(n^{k-1/4+\varepsilon})$ for all $\varepsilon > 0$. It was conjectured that $a_n = O(n^{k-1/2} \cdot \sigma_0(n))$ (for the τ-function, this goes back to Ramanujan). The usual story with such conjectures is that they prompt an infinite sequence of papers proving results converging to the conjecture, but (happily) in this case Deligne proved in 1969 that the conjecture follows from the Weil conjectures for varieties over finite fields, and he proved the Weil conjectures in 1973. I hope to return to this question.

Modular forms as sections of line bundles. Let X be a topological manifold. A line bundle on X is a map of topological spaces $\pi : L \to X$ such that, for some open covering $X = \bigcup U_i$ of X, $\pi^{-1}(U_i) \approx U_i \times \mathbb{R}$. Similarly, a line bundle on a Riemann surface is a map of complex manifolds $\pi : L \to X$ such that locally L is isomorphic to $U \times \mathbb{C}$, and a line bundle on an algebraic variety is a map of algebraic varieties $\pi : L \to X$ such that locally for the Zariski topology on X, $L \approx U \times \mathbb{A}^1$.

If L is a line bundle on X (say a Riemann surface), then for any open subset U of X, $\Gamma(U, L)$ denotes the group of sections of L over U, i.e., the set of holomorphic maps $f : U \to L$ such that $\pi \circ f =$ identity map. Note that if $L = U \times \mathbb{C}$, then $\Gamma(U, L)$ can be identified with the set of holomorphic functions on U. (The Γ in $\Gamma(U, L)$ should not be confused with a congruence group Γ.)

Now consider the following situation: Γ is a group acting freely and properly discontinuously on a Riemann surface H, and $X = \Gamma \backslash H$. Write p for the quotient map $H \to X$. Let $\pi : L \to X$ be a line bundle on X; then

$$p^*(L) \stackrel{\text{df}}{=} \{(h, l) \subset H \times L \mid p(h) = \pi(l)\}$$

is a line bundle on H (for example, $p^*(X \times L) = H \times L$), and Γ acts on $p^*(L)$ through its action on H. Suppose we are given an isomorphism $i : H \times \mathbb{C} \to p^*(L)$. Then we can transfer the action of Γ on $p^*(L)$ to an action of Γ on $H \times \mathbb{C}$ over H. For $\gamma \in \Gamma$ and $(\tau, z) \in H \times \mathbb{C}$, write

$$\gamma(\tau, z) = (\gamma\tau, j_\gamma(\tau) z),\; j_\gamma(\tau) \in \mathbb{C}^\times.$$

Then

$$\gamma\gamma'(\tau, z) = \gamma(\gamma'\tau, j_{\gamma'}(\tau)z) = (\gamma\gamma'\tau, j_\gamma(\gamma'\tau)\cdot j_{\gamma'}(z)\cdot z).$$

Hence:

$$j_{\gamma\gamma'}(\tau) = j_\gamma(\gamma'\tau)\cdot j_{\gamma'}(\tau).$$

DEFINITION 4.23. An *automorphy factor* is a map $j : \Gamma \times \mathbb{H} \to \mathbb{C}^\times$ such that

(a) for each $\gamma \in \Gamma$, $\tau \mapsto j_\gamma(\tau)$ is a holomorphic function on $\mathbb{H}$;
(b) $j_{\gamma\gamma'}(\tau) = j_\gamma(\gamma'\tau)\cdot j_{\gamma'}(\tau)$.

Condition (b) should be thought of as a cocycle condition (in fact, that's what it is). Note that if j is an automorphy factor, so also is j^k for any integer k.

EXAMPLE 4.24. For any open subset H of $\mathbb{C}$ with a group Γ acting on it, there is canonical automorphy factor $j_\gamma(\tau)$, namely,

$$\Gamma \times H \to \mathbb{C},\ (\gamma, \tau) \mapsto (d\gamma)_\tau.$$

By $(d\gamma)_\tau$ I mean the following: each γ defines a map $H \to H$, and $(d\gamma)_\tau$ is the map on the tangent space at τ defined by γ. As $H \subset \mathbb{C}$, the tangent spaces at τ and at $\gamma\tau$ are canonically isomorphic to $\mathbb{C}$, and so $(d\gamma)_\tau$ can be regarded as a complex number.

Suppose we have maps

$$M \xrightarrow{\alpha} N \xrightarrow{\beta} P$$

of (complex) manifolds, then for any point $m \in M$, $(d(\beta\circ\alpha))_m = (d\beta)_{\alpha(n)} \circ (d\alpha)_m$ (maps on tangent spaces). Therefore,

$$j_{\gamma\gamma'}(\tau) =_{df} (d\gamma\gamma')_\tau = (d\gamma)_{\gamma'\tau}\cdot(d\gamma')_\tau = j_\gamma(\gamma'\tau)\cdot j_{\gamma'}(\tau).$$

Thus $j_\gamma(\tau) =_{df} (d\gamma)_\tau$ is an automorphy factor.

For example, consider $\Gamma(1)$ acting on $\mathbb{H}$. If $\gamma = (z \mapsto \frac{az+b}{cz+d})$, then

$$d\gamma = \frac{1}{(cz+d)^2}dz,$$

and so $j_\gamma(\tau) = (cz+d)^{-2}$, and $j_\gamma(\tau)^k = (cz+d)^{-2k}$.

PROPOSITION 4.25. *There is a one-to-one correspondence between the set of pairs (L, i) where L is a line bundle on $\Gamma\backslash H$ and i is an isomorphism $H \times \mathbb{C} \approx p^*(L)$ and the set of automorphy factors.*

PROOF. We have seen how to go $(L, i) \mapsto j_\gamma(\tau)$. For the converse, use i and j to define an action of Γ on $H \times \mathbb{C}$, and define L to be $\Gamma\backslash H \times \mathbb{C}$. $\square$

REMARK 4.26. Every line bundle on $\mathbb{H}$ is trivial (i.e., isomorphic to $\mathbb{H}\times\mathbb{C}$), and so Proposition 4.25 gives us a classification of the line bundles on $\Gamma\backslash\mathbb{H}$.

Let L be a line bundle on X. Then

$$\Gamma(X, L) = \{F \in \Gamma(H, p^*L) \mid F \text{ commutes with the actions of } \Gamma\}.$$

Suppose we are given an isomorphism $p^*L \approx H \times \mathbb{C}$. We use it to identify the two line bundles on H. Then Γ acts on $H \times \mathbb{C}$ by the rule:

$$\gamma(\tau, z) = (\gamma\tau, j_\gamma(\tau)z).$$

A holomorphic section $F : H \to H \times \mathbb{C}$ can be written $F(\tau) = (\tau, f(\tau))$ with $f(\tau)$ a holomorphic map $H \to \mathbb{C}$. What does it mean for F to commute with the action of Γ? We must have

$$F(\gamma\tau) = \gamma F(\tau), \text{ i.e., } (\gamma\tau, f(\gamma\tau)) = (\gamma\tau, j_\gamma(\tau)f(\tau)).$$

Hence

$$f(\gamma\tau) = j_\gamma(\tau) \cdot f(\tau).$$

Thus, if L_k is the line bundle on $\Gamma\backslash\mathbb{H}$ corresponding to $j_\gamma(\tau)^{-k}$, where $j_\gamma(\tau)$ is the canonical automorphy factor (4.24), then the condition becomes

$$f(\gamma\tau) = (cz+d)^{2k} \cdot f(\tau),$$

i.e., condition (4.5a). Therefore the sections of L_k are in natural one-to-one correspondence with the functions on $\mathbb{H}$ satisfying (4.5a,b). The line bundle L_k extends to a line bundle L_k^* on the compactification $\Gamma\backslash\mathbb{H}^*$, and the sections of L_k^* are in natural one-to-one correspondence with the modular forms of weight $2k$.

Poincaré series. We want to construct modular forms for subgroups Γ of finite index in $\Gamma(1)$. Throughout, we write Γ' for the image of Γ in $\Gamma(1)/\{\pm I\}$.

Recall the standard way of constructing invariant functions: if h is a function on $\mathbb{H}$, then

$$f(z) \stackrel{\text{df}}{=} \sum_{\gamma\in\Gamma'} h(\gamma z)$$

is invariant under Γ, provided the series converges absolutely (which it rarely will). Poincaré found a similar argument for constructing modular forms.

Let

$$\Gamma \times \mathbb{H} \to \mathbb{C},\ (\gamma, z) \mapsto j_\gamma(z)$$

be an automorphy factor for Γ; thus

$$j_{\gamma\gamma'}(z) = j_\gamma(\gamma' z) \cdot j_{\gamma'}(z).$$

Of course, we will be particularly interested in the case

$$j_\gamma(z) = (cz+d)^{2k},\ \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

We wish to construct a function f such that $f(\gamma z) = j_\gamma(z) \cdot f(z)$.

Try

$$f(z) = \sum_{\gamma\in\Gamma'} \frac{h(\gamma z)}{j_\gamma(z)}.$$

If this series converges absolutely uniformly on compact sets, then

$$f(\gamma' z) = \sum_{\gamma\in\Gamma'} \frac{h(\gamma\gamma' z)}{j_\gamma(\gamma' z)} = \sum_{\gamma\in\Gamma'} \frac{h(\gamma\gamma' z)}{j_{\gamma\gamma'}(z)} j_{\gamma'}(z) = j_{\gamma'}(z) \cdot f(z)$$

as wished.

Unfortunately, there is little hope of convergence, for the following (main) reason: there may be infinitely many γ's for which $j_\gamma(z) = 1$ identically, and so the sum contains infinitely many redundant terms. Let

$$\Gamma_0 = \{\gamma \in \Gamma' \mid j_\gamma(z) = 1 \text{ identically}\}.$$

For example, if $j_\gamma(z) = (cz+d)^{-2k}$, then

$$\Gamma_0 = \left\{\pm\begin{pmatrix} a & b \\ c & d\end{pmatrix} \in \Gamma \mid c = 0,\ d = 1\right\} = \left\{\pm\begin{pmatrix} 1 & b \\ 0 & 1\end{pmatrix} \in \Gamma\right\} = <\pm\begin{pmatrix} 1 & h \\ 0 & 1\end{pmatrix}>$$

where h is the smallest positive integer such that $\begin{pmatrix} 1 & h \\ 0 & 1\end{pmatrix} \in \Gamma$ (thus h is the width of the cusp $i\infty$ for Γ). In particular, Γ_0 is an infinite cyclic group.

If γ, $\gamma' \in \Gamma_0$, then

$$j_{\gamma\gamma'}(z) = j_\gamma(\gamma' z) \cdot j_{\gamma'}(z) = 1 \quad (\text{all } z),$$

and so Γ_0 is closed under multiplication—in fact, it is a subgroup of Γ'.

Let h be a holomorphic function on $\mathbb{H}$ invariant under Γ_0, i.e., such that $h(\gamma_0 z) = h(z)$ for all $\gamma_0 \in \Gamma_0$. Let $\gamma \in \Gamma'$ and $\gamma_0 \in \Gamma_0$; then

$$\frac{h(\gamma_0\gamma z)}{j_{\gamma_0\gamma}(z)} = \frac{h(\gamma z)}{j_{\gamma_0}(\gamma z)\cdot j_\gamma(z)} = \frac{h(\gamma z)}{j_\gamma(z)},$$

i.e., $h(\gamma z)/j_\gamma(z)$ is constant on the coset $\Gamma_0\gamma$. Thus we can consider the series

$$f(z) = \sum_{\Gamma_0\backslash\Gamma'} \frac{h(\gamma z)}{j_\gamma(z)}$$

If the series converges absolutely uniformly on compact sets, then the previous argument shows that we obtain a holomorphic function f such that $f(\gamma z) = j_\gamma(z)\cdot f(z)$.

Apply this with $j_\gamma(z) = (cz+d)^{2k}$, $\gamma = \begin{pmatrix} a & b \\ c & d\end{pmatrix}$, and Γ a subgroup of finite index in $\Gamma(1)$. As we noted above, Γ_0 is generated by $z \mapsto z+h$ for some h, and a typical function invariant under $z \mapsto z+h$ is $\exp(2\pi i n z/h)$, $n = 0, 1, 2, \ldots$

Definition 4.27. The *Poincaré series of weight* $2k$ and *character* n for Γ is the series

$$\varphi_n(z) = \sum_{\Gamma_0\backslash\Gamma'} \frac{\exp(\frac{2\pi i n\cdot\gamma(z)}{h})}{(cz+d)^{2k}}$$

where Γ' is the image of Γ in $\Gamma(1)/\{\pm I\}$.

We need a set of representatives for $\Gamma_0\backslash\Gamma'$. Note that

$$\begin{pmatrix} 1 & m \\ 0 & 1\end{pmatrix}\cdot\begin{pmatrix} a & b \\ c & d\end{pmatrix} = \begin{pmatrix} a+mc & b+md \\ c & d\end{pmatrix}.$$

Using this, it is easily checked that $\begin{pmatrix} a & b \\ c & d\end{pmatrix}$ and $\begin{pmatrix} a' & b' \\ c' & d'\end{pmatrix}$ are in the same coset of Γ_0 if and only if $(c, d) = \pm(c', d')$ and $(a, b) \equiv \pm(a', b') \mod h$. Thus a set of representatives for $\Gamma_0\backslash\Gamma'$ can be obtained by taking one element of Γ' for each pair (c, d), $c > 0$, which is the second row of a matrix in Γ'.

THEOREM 4.28. *The Poincaré series $\varphi_n(z)$ for $2k \geq 2$, $n \geq 0$, converges absolutely uniformly on compact subsets of $\mathbb{H}$; it converges absolutely uniformly on every fundamental domain D for Γ, and hence is a modular form of weight $2k$ for Γ. Moreover,*

(a) *$\varphi_0(z)$ is zero at all finite cusps, and $\varphi_0(i\infty) = 1$;*
(b) *for all $n \geq 1$, $\varphi_n(z)$ is a cusp form.*

PROOF. To see convergence, compare the Poincaré series with

$$\sum_{m,n\in\mathbb{Z},(m,n)\neq(0,0)} \frac{1}{|mz+n|^{2k}}$$

which converges uniformly on compact subsets of $\mathbb{H}$ when $2k > 2$. For the details of the proof, which is not difficult, see Gunning 1962, III.9. □

THEOREM 4.29. *The Poincaré series $\varphi_n(z)$, $n \geq 1$, of weight $2k$ span $\mathcal{M}_k(\Gamma)$.*

Before we can prove this, we shall need some preliminaries.

The geometry of $\mathbb{H}$. As Poincaré pointed out, $\mathbb{H}$ can serve as a model for non-Euclidean hyperbolic plane geometry.

Recall that the axioms for hyperbolic geometry are the same as for Euclidean geometry, except that the axiom of parallels is replaced with the following axiom: suppose we are given a straight line and a point in the plane; if the line does not contain the point, then there exist at least two lines passing through the point and not intersecting the line.

The points of our non-Euclidean plane are the points of $\mathbb{H}$. A non-Euclidean "line" is a half-circle in $\mathbb{H}$ orthogonal to the real axis, or a vertical half-line. The angle between two lines is the usual angle. To obtain the distance $\delta(z_1, z_2)$ between two points, draw the non-Euclidean line through z_1 and z_2, let ∞_1 and ∞_2 be the points on the real axis (or $i\infty$) on the "line" labeled in such a way that $\infty_1, z_1, z_2, \infty_2$ follow one another cyclically around the circle, and define

$$\delta(z_1, z_2) = \log\ D(z_1, z_2, \infty_1, \infty_2)$$

where $D(z_1, z_2, z_3, z_4)$ is the cross-ratio.

The group $\mathrm{PSL}_2(\mathbb{R}) \stackrel{\mathrm{df}}{=} \mathrm{SL}_2(\mathbb{R})/\pm I$ plays the same role as the group of orientation preserving affine transformations in the Euclidean plane, namely, it is the group of transformations preserving distance and orientation.

The measure $\mu(U) = \iint_U \frac{dxdy}{y^2}$ plays the same role as the usual measure $dxdy$ on $\mathbb{R}^2$—it is invariant under translation by elements of $\mathrm{PSL}_2(\mathbb{R})$. This follows from the invariance of the differential $y^{-2}dxdy$. (We prove something more general below.)

Thus we can consider $\iint_D \frac{dxdy}{y^2}$ for any fundamental domain D of Γ—the invariance of the differential shows that this doesn't depend on the choice of D. One shows that the integral does converge, and in fact that

$$\int_D dx \cdot dy/y^2 = 2\pi(2g - 2 + \nu_\infty + \sum(1 - 1/e_P)).$$

See Shimura, 2.5. (There is a detailed discussion of the geometry of $\mathbb{H}$— equivalently, the open unit disk—in C. Siegel, Topics in Complex Functions II, Wiley, 1971, Chapter 3.)

Petersson inner product. Let f and g be two modular forms of weight $2k > 0$ for a subgroup Γ of finite index in $\Gamma(1)$.

LEMMA 4.30. *The differential $f(z)\cdot\overline{g(z)}\cdot y^{2k-2}dxdy$ is invariant under the action of $\mathrm{SL}_2(\mathbb{R})^+$. (Here $z = x+iy$, so the notation is mixed.)*

PROOF. Let $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then

$$\begin{aligned}
f(\gamma z) &= (cz+d)^{2k}\cdot f(z) && \text{(definition of a modular form)}\\
\overline{g(\gamma z)} &= \overline{(cz+d)}^{2k}\cdot\overline{g(z)} && \text{(the conjugate of the definition)}\\
\Im(\gamma z) &= \frac{\Im(z)}{|cz+d|^2} && \text{(see the Introduction)}\\
\gamma^*(dx\cdot dy) &= \frac{dxdy}{|cz+d|^4}.
\end{aligned}$$

The last equation follows from the next lemma and the fact (4.8) that $d\gamma/dz = 1/(cz+d)^2$. On raising the third equation to the $(2k-2)$th power, and multiplying, we obtain the result. □

LEMMA 4.31. *For any holomorphic function $w(z)$, the map $z \mapsto w(z)$ multiplies areas by $|\frac{dw}{dz}|^2$.*

PROOF. Write $w(z) = u(x,y) + iv(x,y)$, $z = x+iy$. Thus, $z \mapsto w(z)$ is the map

$$(x,y) \mapsto (u(x,y), v(x,y)),$$

and the Jacobian is

$$\begin{vmatrix} u_x & v_x \\ u_y & v_y \end{vmatrix} = u_x v_y - v_x u_y.$$

According to the Cauchy-Riemann equations, $w'(z) = u_x + iv_x$, $u_x = v_y$, $u_y = -v_x$, and so

$$|w'(z)|^2 = u_x^2 + v_x^2 = u_x v_y - v_x u_y.$$

□

LEMMA 4.32. *Let D be a fundamental domain for Γ. If f or g is a cusp form, then the integral*

$$\iint_D f(z)\cdot\overline{g(z)}\cdot y^{2k-2}dxdy$$

converges.

PROOF. Clearly the integral converges if we exclude a neighbourhood of each of the cusps. Near the cusp $i\infty$, $f(z)\cdot\overline{g(z)} = O(e^{-cy})$ for some $c > 0$, and so the integral is dominated by $\int_{y_1}^{\infty} e^{-cy}y^{k-2}dy < \infty$. The other cusps can be handled similarly. □

Let f and g be modular forms of weight 2k for some group $\Gamma \subset \Gamma(1)$, and assume that one at least is a cusp form. The *Petersson inner product* of f and g is defined to be

$$<f,g>= \iint_D f(z) \cdot \overline{g(z)} \cdot y^{2k-2} dxdy.$$

Lemma 4.30 shows that it is independent of the choice of D. It has the following properties:

- it is linear in the first variable, and semi-linear in the second;
- $<f,g>= \overline{<g,f>}$;
- $< f,f >> 0$ for all $f \neq 0$.

It is therefore a positive-definite Hermitian form on $\mathcal{S}_k(\Gamma)$, and so $\mathcal{S}_k(\Gamma)$ together with $<,>$ is a finite-dimensional Hilbert space.

Completeness of the Poincaré series. Again let Γ be a subgroup of finite index in $\Gamma(1)$.

THEOREM 4.33. *Let f be a cusp form of weight $2k \geq 2$ for Γ, and let φ_n be the Poincaré series of weight $2k$ and character $n \geq 1$ for Γ. Then*

$$<f,\varphi_n>= \frac{h^{2k}(2k-2)!}{(4\pi)^{2k-1}} \cdot n^{1-2k} \cdot a_n$$

where h is the width of $i\infty$ as a cusp for Γ and a_n is the n^{th} coefficient in the Fourier expansion of f :

$$f = \sum a_n e^{\frac{2\pi i n z}{h}}.$$

PROOF. Write φ_n as a sum, and interchange the order of the integral and the sum. Look at a typical term. Write it as an integral over a fundamental domain for Γ_0 in $\mathbb{H}$,

$$< f,\varphi_n >= \int_{x=0}^{h} \int_{y=0}^{\infty} f(z) \cdot \exp(-2\pi i n z/h) \cdot y^{2(2k-1)} \cdot dxdy.$$

Now write $f(z)$ as a sum, and interchange the order of integration and summation. Evaluate. See Gunning, III.11, for the details. □

COROLLARY 4.34. *Every cusp form is a linear combination of Poincaré series $\varphi_n(z)$, $n \geq 1$.*

PROOF. If f is orthogonal to the subspace generated by the Poincaré series, then all the coefficients of its Fourier expansion are zero. □

Eisenstein series for $\Gamma(N)$. The Poincaré series of weight $2k > 2$ and character 0 for $\Gamma(N)$ is

$$\phi_0(z) = \sum \frac{1}{(cz+d)^{2k}} \quad (\text{sum over } (c,d) \equiv (0,1) \mod N,\ \gcd(c,d)=1).$$

Recall (4.28) that this is a modular form of weight $2k$ for $\Gamma(N)$ which takes the value 1 at $i\infty$ and vanishes at all the other cusps.

For every complex-valued function ν on the (finite) set of inequivalent cusps for $\Gamma(N)$, we want to construct a modular function f of weight $2k$ such that $f|\{\text{cusps}\} = \nu$. Moreover, we would like to choose the f's to be orthogonal (for the Petersson inner product) to the space of cusp forms. To do this, we shall construct a function (restricted Eisenstein series) which takes the value 1 at a particular cusp, takes the value 0 at the remaining cusps, and is orthogonal to cusp forms.

Write $j_\gamma(z) = 1/(cz+d)^2$, so that $j_\gamma(z)$ is an automorphy factor:

$$j_{\gamma\gamma'}(z) = j_\gamma(\gamma' z) \cdot j_{\gamma'}(z).$$

Let P be a cusp for $\Gamma(N)$, $P \neq i\infty$, and let $\sigma \in \Gamma(1)$ be such that $\sigma(P) = i\infty$. Define

$$\varphi(z) = j_\sigma(z)^k \cdot \varphi_0(\sigma z).$$

LEMMA 4.35. *The function $\varphi(z)$ is a modular form of weight $2k$ for $\Gamma(N)$; moreover φ takes the value 1 at P, and it is zero at every other cusp.*

PROOF. Let $\gamma \in \Gamma(N)$. For the first statement, we have to show that $\varphi(\gamma z) = j_\gamma(z)^{-k}\varphi(z)$. From the definition of φ, we find that

$$\varphi(\gamma z) = j_\sigma(\gamma z)^k \cdot \varphi_0(\sigma\gamma z).$$

As $\Gamma(N)$ is normal, $\sigma\gamma\sigma^{-1} \in \Gamma(N)$, and so

$$\varphi_0(\sigma\gamma z) = \varphi_0(\sigma\gamma\sigma^{-1} \cdot \sigma z) = j_{\sigma\gamma\sigma^{-1}}(\sigma z)^{-k} \cdot \varphi_0(\sigma z).$$

On comparing this formula for $\varphi(\gamma z)$ with

$$j_\gamma(z)^{-k} \cdot \varphi(z) = j_\gamma(z)^{-k} \cdot j_\sigma(z)^k \cdot \varphi_0(\sigma z),$$

we see that it suffices to prove that

$$j_\sigma(\gamma z) \cdot j_{\sigma\gamma\sigma^{-1}}(\sigma z)^{-1} = j_\gamma(z)^{-1} \cdot j_\sigma(z),$$

or that

$$j_\sigma(\gamma z) \cdot j_\gamma(z) = j_{\sigma\gamma\sigma^{-1}}(\sigma z) \cdot j_\sigma(z).$$

But, because of the defining property of automorphy factors, this is just the obvious equality

$$j_{\sigma\gamma}(z) = j_{\sigma\gamma\sigma^{-1}\sigma}(z).$$

The second statement is a consequence of the definition of φ and the properties of φ_0. $\square$

We now compute $\varphi(z)$. Let T be a set of coset representatives for Γ_0 in $\Gamma(N)$. Then

$$\begin{aligned}\varphi(z) &\overset{\text{df}}{=} j_\sigma(z)^k \cdot \varphi_0(\sigma z)\\ &= j_\sigma(z)^k \cdot \textstyle\sum_{\tau\in T} j_\tau(\sigma z)^k\\ &= \textstyle\sum_{\tau\in T} j_{\tau\sigma}(z)^k\\ &= \textstyle\sum_{\gamma\in T\sigma} j_\gamma(z)^k.\end{aligned}$$

Let $\sigma = \begin{pmatrix} a_0 & b_0 \\ c_0 & d_0 \end{pmatrix}$, so that $\sigma^{-1} = \begin{pmatrix} d_0 & -b_0 \\ -c_0 & a_0 \end{pmatrix}$, and $P = \sigma^{-1}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = -d_0/c_0$. Note that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma(N) \implies \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} a_0 & b_0 \\ c_0 & d_0 \end{pmatrix} \equiv \begin{pmatrix} * & * \\ c_0 & d_0 \end{pmatrix} \mod N.$$

From this, we can deduce that $T\sigma$ contains exactly one element of $\Gamma(N)'$ for each pair (c, d) with $\gcd(c, d) = 1$ and $(c, d) \equiv (c_0, d_0)$.

DEFINITION 4.36. (a) A *restricted Eisenstein series* of weight $2k > 2$ for $\Gamma(N)$ is a series

$$G(z; c_0, d_0; N) = \sum (cz + d)^{-2k}$$

(sum over $(c, d) \equiv (c_0, d_0) \mod N$, $\gcd(c, d) = 1$). Here (c_0, d_0) is a pair such that $\gcd(c_0, d_0, N) = 1$.

(b) A *general Eisenstein series of weight* $2k > 2$ for $\Gamma(N)$ is a series

$$G(z; c_0, d_0; N) = \sum (cz + d)^{-2k}$$

(sum over $(c, d) \equiv (c_0, d_0) \mod N$, $(c, d) \neq (0, 0)$). Here it is not required that $\gcd(c_0, d_0, N) = 1$.

Consider the restricted Eisenstein series. Clearly,

$$G(z; c_0, d_0; N) = G(z; c_1, d_1; N)$$

if $(c_0, d_0) \equiv \pm(c_1, d_1) \mod N$. On the other hand, we get a restricted Eisenstein series for each cusp, and these Eisenstein series are linearly independent. On counting, we see that there is exactly one restricted Eisenstein series for each cusp, and so the distinct restricted Eisenstein series are linearly independent.

PROPOSITION 4.37. *The general Eisenstein series are the linear combinations of the restricted Eisenstein series.*

PROOF. Omitted. □

REMARK 4.38. (a) Sometimes Eisenstein series are defined to be the linear combinations of restricted Eisenstein series.

(b) The Petersson inner product $<f, g>$ is defined provided at least one of f or g is a cusp form. One finds that $<f, g> = 0$ (e.g., φ_0 gives the 0^{th} coefficient) for the restricted Eisenstein series, and hence $<f, g> = 0$ for all cusp forms f and all Eisenstein series g : the space of Eisenstein series is the orthogonal complement of $\mathcal{S}_k(\Gamma)$ in $\mathcal{M}_k(\Gamma)$.

For more details on Eisenstein series for $\Gamma(N)$, see Gunning 1962, IV.13.

ASIDE 4.39. In the one-dimensional case, compactifying $\Gamma\backslash\mathbb{H}$ presents no problem, and the Riemann-Roch theorem tells us there are many modular forms. The Poincaré series allow us to write down a set of modular forms that spans $\mathcal{S}_k(\Gamma)$. In the higher dimensional case (see 2.26), it is much more difficult to embed the quotient $\Gamma\backslash D$ of a bounded symmetric domain in a compact analytic space. Here the Poincaré series play a much more crucial role. In their famous 1964 paper, Baily and Borel showed that the Poincaré series can be used to give an embedding of the complex manifold $\Gamma\backslash D$ into projective space, and that the closure of the image is a projective algebraic variety. It follows that $\Gamma\backslash D$ has a canonical structure of an algebraic variety.

In the higher-dimensional case, the boundary of $\Gamma\backslash D$, i.e., the complement of $\Gamma\backslash D$ in its compactification, is more complicated than in the one-dimensional case. It is a union of varieties of the form $\Gamma'\backslash D'$ with D' a bounded symmetric domain of lower dimension than that of D. The Eisenstein series then attaches to a cusp form on D'

a modular form on D. (In our case, a cusp form on the zero-dimensional boundary is just a complex number.)

5. Hecke Operators

Hecke operators play a fundamental role in the theory of modular forms. After describing the problem they were first introduced to solve, we develop the theory of Hecke operators for the full modular group, and then for a congruence subgroup of the modular group.

Introduction. Recall that the cusp forms of weight 12 for $\Gamma(1)$ form a one-dimensional vector space over $\mathbb{C}$, generated by $\Delta = g_2^3 - 27g_3^2$, where $g_2 = 60G_2$ and $g_3 = 140G_3$. In more geometric terms, $\Delta(z)$ is the discriminant of the elliptic curve $\mathbb{C}/\mathbb{Z}z + \mathbb{Z}$. Jacobi showed that

$$\Delta(z) = (2\pi)^{12} \cdot q \cdot \prod_{n=1}^{\infty}(1-q^n)^{24}, \quad q = e^{2\pi i z}.$$

Write $f(z) = q \cdot \prod_{n=1}^{\infty}(1-q)^{24} = \sum \tau(n) \cdot q^n$. Then $n \mapsto \tau(n)$ is the *Ramanujan τ-function.* Ramanujan conjectured that it had the following properties:

(a) $|\tau(p)| \leq 2 \cdot p^{11/2}$,

(b) $\begin{cases} \tau(mn) &= \tau(m)\tau(n) \text{ if } \gcd(m,n) = 1 \\ \tau(p)\tau(p^n) &= \tau(p^{n+1}) + p^{11}\tau(p^{n-1}) \text{ if } p \text{ is prime and } n \geq 1. \end{cases}$

Property (b) was proved by Mordell in 1917 in a paper in which he introduced the first examples of Hecke operators. To Δ we can attach a Dirichlet series

$$L(\Delta, s) = \sum \tau(n) n^{-s}.$$

Proposition 5.1. *The Dirichlet series $L(\Delta, s)$ has an Euler product expansion of the form*

$$L(\Delta, s) = \prod_{p \text{ prime}} \frac{1}{(1 - \tau(p)p^{-s} + p^{11-2s})}$$

if and only if (b) holds.

Proof. For a prime p, define

$$L_p(s) = \sum_{m \geq 0} \tau(p^m) \cdot p^{-ms} = 1 + \tau(p) \cdot p^{-s} + \tau(p^2) \cdot (p^{-s})^2 + \cdots .$$

If $n \in \mathbb{N}$ has the factorization $n = \prod p_i^{r_i}$, then the coefficient of $(p^{-s})^n$ in $\prod L_p(s)$ is $\prod \tau(p_i^{r_i})$, which the first equation in (b) implies is equal to $\tau(n)$. Thus

$$L(\Delta, s) = \prod L_p(s).$$

Now consider

$$(1 - \tau(p)p^{-s} + p^{11-2s}) \cdot L_p.$$

By inspection, we find that the coefficient of $(p^{-s})^n$ in this product is

1 for $n = 0$;

0 for $n = 1$;

. . . .

$\tau(p^{n+1}) - \tau(p)\tau(p^n) + p^{11}\tau(p^{n-1})$ for $n + 1$.

Thus the second equation in (b) implies that $(1-\tau(p)p^{-s}+p^{11-2s})\cdot L_p=1$, and hence that

$$L(\Delta,s)=\prod_p(1-\tau(p)p^{-s}+p^{11-2s})^{-1}.$$

The argument can be run in reverse. □

PROPOSITION 5.2. *Write*

$$1-\tau(p)X+p^{11}X^2=(1-aX)(1-a'X);$$

Then the following conditions are equivalent:

(a) $|\tau(p)|\le 2\cdot p^{11/2}$;
(b) $|a|=p^{11/2}=|a'|$;
(c) *a and a' are conjugate complex numbers, i.e., $a'=\bar{a}$.*

PROOF. First note that $\tau(p)$ is real (in fact, it is an integer).

(b) $\implies$ (a): We have $\tau(p)=a+a'$, and so (a) follows from the triangle inequality.

(c) $\implies$ (b): We have that $|a|^2=a\bar{a}=aa'=p^{11}$.

(a) $\implies$ (c): The discriminant of $1-\tau(p)X+p^{11}X^2$ is $\tau(p)^2-4p^{11}$, which (a) implies is <0. □

For each $n\ge 1$, we shall define an operator:

$$T(n):\mathcal{M}_k(\Gamma(1))\to\mathcal{M}_k(\Gamma(1)).$$

These operators will have the following properties:

$T(m)\circ T(n)=T(mn)$ if $\gcd(m,n)=1$;

$T(p)\circ T(p^n)=T(p^{n+1})+p^{2k-1}T(p^{n-1})$, p prime;

$T(n)$ preserves the space of cusp forms, and is a Hermitian (self-adjoint) operator on $\mathcal{S}_k(\Gamma)$:

$$<T(n)f,g>=<f,T(n)g>,\quad f,g \text{ cusp forms.}$$

LEMMA 5.3. *Let V be a finite-dimensional vector space over $\mathbb{C}$ with a positive definite Hermitian form $<,>$.*

(a) Let $\alpha:V\to V$ be a linear map which is Hermitian (i.e., such that $<\alpha v,v'>=<v,\alpha v'>$); then V has a basis consisting of eigenvectors for α (thus α is diagonalizable).

(b) Let $\alpha_1,\alpha_2,\ldots$ be a sequence of commuting Hermitian operators; then V has a basis consisting of vectors that are eigenvectors for all α_i (thus the α_i are simultaneously diagonalizable).

PROOF. (a) Because $\mathbb{C}$ is algebraically closed, α has an eigenvector e_1. Let $V_1=(\mathbb{C}\cdot e_1)^\perp$. Because α is Hermitian, V_1 is stable under α, and so it has an eigenvector e_2. Let $V_2=(\mathbb{C}e_1+\mathbb{C}e_2)^\perp$, and continue in this manner.

(b) From (a) we know that $V=\oplus V(\lambda_i)$ where the λ_i are the distinct eigenvalues for α_1 and $V(\lambda_i)$ is the eigenspace for λ_i; thus α_1 acts as multiplication by λ_i on $V(\lambda_i)$. Because α_2 commutes with α_1 , it preserves each $V(\lambda_i)$, and we can decompose each $V(\lambda_i)$ further into a sum of eigenspaces for α_2. Continuing in this fashion, we arrive at a decomposition $V=\oplus V_j$ such that each α_i acts as a scalar on each V_j. Now choose a basis for each V_j and take the union. □

REMARK 5.4. The pair $(V, <,>)$ is a finite-dimensional Hilbert space. There is an analogous statement to the lemma for infinite-dimensional Hilbert spaces (it's called the spectral theorem).

PROPOSITION 5.5. *Let $f(z) = \sum c(n)q^n$ be a modular form of weight $2k$, $k > 0$, $f \neq 0$. If f is an eigenfunction for all $T(n)$, then $c(1) \neq 0$, and when we normalize f so that $c(1) = 1$, then*

$$T(n)f = c(n) \cdot f.$$

PROOF. See later (5.18). □

COROLLARY 5.6. *If $f(z)$ is a normalized eigenform for all $T(n)$, then $c(n)$ is real.*

PROOF. The eigenvalues of a Hermitian operator are real, because

$$<\alpha v, v>=<\lambda v, v>= \lambda <v, v>, =<v, \alpha v>=<v, \lambda v>= \bar{\lambda} <v, v>$$

for any eigenvector v. □

We deduce from these statements that if f is a normalized eigenform for all the $T(n)$, then

$c(m)c(n) = c(mn)$ if $\gcd(m, n) = 1$;

$c(p)c(p^n) = c(p^{n+1}) + p^{2k-1}c(p^{n-1})$ if p is prime $n \geq 1$.

Just as in the case of Δ, this implies that

$$L(f, s) \stackrel{\text{df}}{=} \sum c(n) \cdot n^{-s} = \prod_{p \text{ prime}} \frac{1}{(1 - c(p)p^{-s} + p^{2k-1-2s})}.$$

Write $1 - c(p)X + p^{2k-1-2s} = (1 - aX)(1 - a'X)$. As before, the following statements are equivalent:

$|c(p)| \leq 2 \cdot p^{\frac{k-1}{2}}$;

$|a| = p^{\frac{k-1}{2}} = |a'|$;

a and a' are complex conjugates.

These statements are also referred to as the Ramanujan conjecture. As we mentioned in Section 3, they have been proved by Deligne.

EXAMPLE 5.7. Because the space of cusp forms of weight 12 is one-dimensional, Δ is a simultaneous eigenform for the Hecke operators, and so Ramanujan's Conjecture (b) for $\tau(n)$ does follow from the existence of Hecke operators with the above properties.

Note the similarity of $L(f, s)$ to the L-function of an elliptic curve $E/\mathbb{Q}$, which is defined to be

$$L(E, s) = \prod_{p \text{ good}} \frac{1}{1 - a(p)p^{-s} + p^{1-2s}}.$$

Here $1 - a(p) + p = \#E(\mathbb{F}_p)$. The Riemann hypothesis for $E/\mathbb{F}_p$ is that $|a(p)| \leq 2\sqrt{p}$. The number $a(p)$ can also be realized as the trace of the Frobenius map on $V_\ell E$. Since $\tau(p)$ is the trace of $T(p)$ acting on an eigenspace, this suggests that there should be a relation of the form

$$\text{“}T(p) = \Pi_p + \bar{\Pi}_p\text{”}$$

where Π_p is the Frobenius operator at p. We shall see that there do exist relations of this form, and that this is the key to Deligne's proof that the Weil conjectures imply the (generalized) Ramanujan conjecture.

CONJECTURE 5.8 (Taniyama). *Let E be an elliptic curve over $\mathbb{Q}$. Then $L(E,s) = L(f,s)$ for some normalized eigenform of weight 2 for $\Gamma_0(N)$, where N is the conductor of E.*

This conjecture is very important. A vague statement of this form was suggested by Taniyama in the 50's, was promoted by Shimura in the 60's, and then in 1967 Weil provided some rather compelling evidence for it. We shall discuss Weil's work in Section 6. Since it is possible to list the normalized eigenforms of weight 2 for $\Gamma_0(N)$ for a fixed N, the conjecture predicts how many elliptic curves with conductor N there are over $\mathbb{Q}$. Computer searches have confirmed the number for small N. [The conjecture has been proved for most elliptic curves by Wiles, Taylor, and Diamond.]

It is known that Conjecture 5.8 implies Fermat's last theorem (which for me, is the most compelling evidence for Fermat's last theorem).

The conjecture is now subsumed by the Langlands program which (roughly speaking) predicts that all Dirichlet series arising from algebraic varieties (more generally, motives) occur among those arising from automorphic forms (better, automorphic representations) for reductive algebraic groups.

Abstract Hecke operators. Let $\mathcal{L}$ be the set of full lattices in $\mathbb{C}$. Recall (4.6) that modular forms are related to functions on $\mathcal{L}$. We first define operators on $\mathcal{L}$, which define operators on functions on $\mathcal{L}$, and then operators on modular forms.

Let $\mathcal{D}$ be the free abelian group generated by the elements of $\mathcal{L}$; thus an element of $\mathcal{D}$ is a finite sum

$$\sum n_i[\Lambda_i],\ n_i \in \mathbb{Z},\ \Lambda_i \in \mathcal{L}.$$

For $n = 1, 2, \ldots$ we define a $\mathbb{Z}$-linear operator $T(n) : \mathcal{D} \to \mathcal{D}$ by setting

$$T(n)[\Lambda] = \sum [\Lambda'] \text{ (sum over all sublattices } \Lambda' \text{of } \Lambda \text{ of index } n).$$

The sum is obviously finite because any such sublattice Λ' contains $n\Lambda$, and $\Lambda/n\Lambda$ is finite. Write $R(n)$ for the operator

$$R(n)[\Lambda] = [n\Lambda].$$

PROPOSITION 5.9. (a) *If m and n are relatively prime, then*

$$T(m) \circ T(n) = T(mn).$$

(b) *If p is prime and $n \geq 1$, then*

$$T(p^n) \circ T(p) = T(p^{n+1}) + pR(p) \circ T(p^{n-1}).$$

PROOF. (a) Note that

$T(mn)[\Lambda] = \sum [\Lambda'']$ (sum over Λ'' with $(\Lambda : \Lambda'') = mn$);

$T(m) \circ T(n)[\Lambda] = \sum [\Lambda'']$ (sum over pairs (Λ', Λ'') with $(\Lambda : \Lambda') = n$, $(\Lambda' : \Lambda'') = m$).

But, if Λ'' is a sublattice of Λ of index mn, then there is a unique chain

$$\Lambda \supset \Lambda' \supset \Lambda$$

with Λ' of index n in Λ, because $\Lambda/mn\Lambda$ is the direct sum of a group of order m and a group of order n.

(b) Let Λ be a lattice. Note that

$T(p^n)\circ T(p)[\Lambda] = \sum[\Lambda'']$ (sum over pairs (Λ', Λ'') with $(\Lambda : \Lambda') = p$, $(\Lambda' : \Lambda'') = p^n$);

$T(p^{n+1})[\Lambda] = \sum[\Lambda'']$ (sum over Λ with $(\Lambda : \Lambda'') = p^{n+1}$);

$pR(p) \circ T(p^{n-1})[\Lambda] = p \cdot \sum R(p)[\Lambda']$ (sum over $\Lambda' \subset \Lambda$ with $(\Lambda : \Lambda') = p^{n-1}$).

Hence $pR(p) \circ T(p^{n-1})[\Lambda] = p \cdot \sum[\Lambda'']$ (sum over $\Lambda'' \subset p\Lambda$ with $(p\Lambda : \Lambda'') = p^{n-1}$).

Each of these is a sum of sublattices Λ'' of index p^{n+1} in Λ. Fix such a lattice, and let a be the number of times it occurs in the first sum, and b the number of times it occurs in the last sum. It occurs exactly once in the second sum, and so we have to prove:

$$a = 1 + pb.$$

There are two cases to consider.

The lattice Λ'' is not contained in $p\Lambda$. Then $b = 0$, and a is the number of lattices Λ' containing Λ'' and of index p in Λ. Such a lattice contains $p\Lambda$, and its image in $\Lambda/p\Lambda$ is of order p and contains the image of Λ'', which is also of order p. Since the subgroups of Λ of index p are in one-to-one correspondence with the subgroups of $\Lambda/p\Lambda$ of index p, this shows that there is exactly one lattice Λ', namely $\Lambda + p\Lambda''$, and so $a = 1$.

The lattice $\Lambda'' \subset p\Lambda$. Here $b = 1$. Any lattice Λ' of index p contains $p\Lambda$, and *a fortiori* Λ. We have to count the number of subgroups of $\Lambda/p\Lambda$ of index p, and this is the number of lines through the origin in the $\mathbb{F}_p$-plane, which is $(p^2 - 1)/(p - 1) = p + 1$. □

COROLLARY 5.10. *For any m and n,*

$$T(m) \cdot T(n) = \sum_{d|\gcd(m,n), d>0} d \cdot R(d) \circ T(mn/d^2)$$

PROOF. Prove by induction on s that

$$T(p^r) \circ T(p^s) = \sum_{i \le r,s} p^i \cdot R(p^i) \circ T(p^{r+s-2i}),$$

and then apply (a) of the theorem. □

COROLLARY 5.11. *Let $\mathcal{H}$ be the $\mathbb{Z}$-subalgebra of* $\mathrm{End}(\mathcal{D})$ *generated by the $T(p)$ and $R(p)$ for p prime; then $\mathcal{H}$ is commutative, and it contains $T(n)$ for all n.*

PROOF. Obvious from 5.10. □

Let F be a function $\mathcal{L} \to \mathbb{C}$. We can extend F by linearity to a function $F : \mathcal{D} \to \mathbb{C}$,

$$F(\sum n_i[\Lambda_i]) = \sum n_i F(\Lambda_i).$$

For any operator T on $\mathcal{D}$, we define $T \cdot F$ to be the function $\mathcal{L} \to \mathbb{C}$ such that

$$(T \cdot F)([\Lambda]) = F(T[\Lambda]).$$

For example,

$$(T(n)\cdot F)([\Lambda]) = \sum F([\Lambda']) \text{ (sum over sublattices } \Lambda' \text{ of } \Lambda \text{ of index } n)$$

and if F has weight $2k$, so that $F(\lambda\Lambda) = \lambda^{-2k}F(\Lambda)$, then

$$R(n)\cdot F = n^{-2k}\cdot F.$$

PROPOSITION 5.12. *Let $F : \mathcal{L} \to \mathbb{C}$ be a homogeneous function of weight $2k$. Then $T(n)\cdot F$ is again of weight $2k$, and for any m and n,*

$$T(m)\cdot T(n)\cdot F = \sum_{d|\gcd(m,n),\ d>0} d^{1-2k}\cdot T(mn/d^2)\cdot F.$$

In particular, if m and n are relatively prime, then

$$T(m)\cdot T(n)\cdot F = T(mn)\cdot F,$$

and if p is prime and $n \geq 1$, then

$$T(p)\cdot T(p^n)\cdot F = T(p^{n+1})\cdot F + p^{1-2k}\cdot T(p^{n-1})\cdot F.$$

PROOF. Immediate from Corollary 5.10 and the definitions. □

Lemmas on 2×2 matrices. Before defining the action of Hecke operators on modular forms, we review some elementary results concerning 2×2 matrices with integer coefficients. Write $M_2(\mathbb{Z})$ for the ring of 2×2 matrices with coefficients in $\mathbb{Z}$.

LEMMA 5.13. *Let A be a 2×2 matrix with coefficients in $\mathbb{Z}$ and determinant n. Then there is an invertible matrix U in $M_2(\mathbb{Z})$ such that $U\cdot A = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$ with*

$$ad = n,\ a \geq 1,\ 0 \leq b < d. \qquad (*)$$

Moreover, the integers a, b, d are uniquely determined.

PROOF. Apply row operations to A that are invertible in the ring $M_2(\mathbb{Z})$ to get A into upper triangular form—see Math 676, 2.43 for the details. For the uniqueness, note that a is the gcd of the elements in the first column of A, d is the unique positive element such that $ad = n$, and b is obviously uniquely determined modulo d. □

REMARK 5.14. Let $M(n)$ be the set of 2×2 matrices with coefficients in $\mathbb{Z}$ and determinant n. The group $\mathrm{SL}_2(\mathbb{Z})$ acts on $M(n)$ by left multiplication, and the lemma provides us with a canonical set of representatives for the orbits:

$$M(n) = \bigcup \mathrm{SL}_2(\mathbb{Z})\cdot \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \text{ (disjoint union over } a, b, d \text{ as in the lemma).}$$

Now let Λ be a lattice in $\mathbb{C}$. Choose a basis ω_1, ω_2 for Λ, so that $\Lambda = \Lambda(\omega_1, \omega_2)$. For any $\alpha = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M(n)$, define $\alpha\Lambda = \Lambda(a\omega_1 + b\omega_2, c\omega_1 + d\omega_2)$. Then $\alpha\Lambda$ is a sublattice of Λ of index n, and every such lattice is of this form for some $\alpha \in M(n)$. Clearly $\alpha\Lambda = \beta\Lambda$ if and only if $\beta = u\alpha$ for $u \in \mathrm{SL}_2(\mathbb{Z})$. Thus we see that the sublattices of Λ of index n are precisely the lattices

$$\Lambda(a\omega_1 + b\omega_2, d\omega_2),\ a, b, d \in \mathbb{Z},\ ad = n,\ a \geq 1,\ 0 \leq b < d-1.$$

For example, consider the case $n = p$. Then the sublattices of Λ are in one-to-one correspondence with the lines through the origin in the 2-dimensional $\mathbb{F}_p$-vector space $\Lambda/p\Lambda$. Write $\Lambda/p\Lambda = \mathbb{F}_p e_1 \oplus \mathbb{F}_p e_2$ with $e_i = \omega_i \pmod{p}$. The lines through the origin are determined by their intersections (if any) with the vertical line through $(1,0)$. Therefore there are $p+1$ lines through the origin, namely,

$$\mathbb{F}_p \cdot e_1, \quad \mathbb{F}_p \cdot (e_1 + e_2), \quad \mathbb{F}_p \cdot (e_1 + (p-1)e_2), \quad \mathbb{F}_p(e_2).$$

Hence there are exactly $p+1$ sublattices of $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$ of index p, namely,

$$\Lambda(\omega_1, p\omega_2), \quad \Lambda(\omega_1 + \omega_2, p\omega_2), \quad \ldots, \Lambda(p\omega_1, \omega_2),$$

in agreement with the general result.

REMARK 5.15. Let $\alpha \in M(n)$, and let $\Lambda' = \alpha\Lambda$. According to a standard theorem, we can choose bases ω_1, ω_2 for Λ and ω_1', ω_2' for Λ' such that

$$\omega_1' = a\omega_1,\ \omega_2' = d\omega_2,\ a, d \in \mathbb{Z},\ ad = n,\ a|d,\ a \geq 1$$

and a, d are uniquely determined. In terms of matrices, this says that

$$M(n) = \bigcup \mathrm{SL}_2(\mathbb{Z}) \cdot \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \cdot \mathrm{SL}_2(\mathbb{Z})$$

—disjoint union over $a, d \in \mathbb{Z}$, $ad = n$, $a|d$, $a \geq 1$. This decomposition of $M(n)$ into a union of double cosets can also be proved directly by applying both row and column operations, invertible in $M_2(\mathbb{Z})$, to the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$.

Hecke operators for $\Gamma(1)$. Recall 4.6 that we have a one-to-one correspondence between functions F on $\mathcal{L}$ of weight $2k$ and functions f on $\mathbb{H}$ that are weakly modular of weight $2k$, under which

$$F(\Lambda(\omega_1, \omega_2)) = \omega_2{}^{-2k} \cdot f(\omega_1/\omega_2);\ f(z) = F(\Lambda(z,1)).$$

Let $f(z)$ be a modular form of weight $2k$, and let F be the associated function of weight $2k$ on $\mathcal{L}$. We define $T(n) \cdot f(z)$ to be the function on $\mathbb{H}$ associated with $n^{2k-1} \cdot T(n) \cdot F$. The factor n^{2k-1} is inserted so that some formulas have integer coefficients rather than rational coefficients. Thus

$$T(n) \cdot f(z) = n^{2k-1} \cdot (T(n) \cdot F)(\Lambda(z,1)).$$

More explicitly,

$$T(n) \cdot f(z) = n^{2k-1} \cdot \sum d^{-2k}\ f(\frac{az+b}{d})$$

where the sum is over the triples a, b, d satisfying (5.13(*)).

PROPOSITION 5.16. (a) *If f is a weakly modular form of weight $2k$ for $\Gamma(1)$, then $T(n) \cdot f$ is also weakly modular of weight $2k$, and*

$$T(m) \cdot T(n) \cdot f = T(mn) \cdot f \text{ if } m \text{ and } n \text{ are relatively prime;}$$
$$T(p) \cdot T(p^n) \cdot f = T(p^{n+1}) \cdot f + p^{2k-1} \cdot T(p^{n-1}) \cdot f \text{ if } p \text{ is prime and } n \geq 1.$$

(b) *Let f be a modular form of weight $2k$ for $\Gamma(1)$, with the Fourier expansion $f = \sum_{m \geq 0} c(m) q^m$, $q = e^{2\pi i z}$. Then $T(n) \cdot f$ is also a modular form, and*

$$T(n) \cdot f(z) = \sum_{m \geq 0} \gamma(m) \cdot q^m$$

with

$$\gamma(m) = \sum_{a \mid \gcd(m,n), a \geq 1} a^{2k-1} \cdot c(\frac{mn}{a^2}).$$

PROOF. (a) We know that

$$T(p) \cdot T(p^n) \cdot F(\Lambda(z,1)) = T(p^{n+1}) \cdot F(\Lambda(z,1)) + p^{1-2k} \cdot T(p^{n-1}) \cdot F(\Lambda(z,1)).$$

On multiplying through by $(p^{n+1})^{2k-1}$ we obtain the second equation. The first is obvious.

(b) We know that

$$T(n) \cdot f(z) = n^{2k-1} \sum_{a,b,d} d^{-2k} \ f(\frac{az+b}{d})$$

where the sum over a, b, d satisfying 5.13(*), i.e.,

$$ad = n, \quad a \geq 1, \quad 0 \leq b < d.$$

Therefore $T(n) \cdot f(z)$ is holomorphic on $\mathbb{H}$ because f is. Moreover

$$T(n) \cdot f(z) = n^{2k-1} \sum_{a,b,d} d^{-2k} \sum_{m \geq 0} c(m) q^{2\pi i \frac{az+b}{d} m}.$$

But

$$\sum_{0 \leq b < d} e^{2\pi i \frac{bm}{d}} = \begin{cases} d & \text{if } d|m \\ 0 & \text{otherwise.} \end{cases}$$

Set $m/d = m'$; then

$$T(n) \cdot f(z) = n^{2k-1} \sum_{a,d,m'} d^{-2k+1} \ c(m'd) q^{am'}$$

where the sum is over the integers a, d, m' such that $ad = n$ and $a \geq 1$. The coefficient of q^t in this is

$$\sum_{a \mid \gcd(n,t), a \geq 1} a^{2k-1} \cdot c(\frac{t}{a}\frac{n}{a}).$$

When we substitute m for t in this formula, we obtain the required formula. Because $\gamma(m) = 0$ for $m < 0$, $T(n) \cdot f$ is holomorphic at $i\infty$. □

COROLLARY 5.17. *Retain the notations of the proposition.*

(a) *The coefficients* $\gamma(0) = \sigma_{2k-1}(n) \cdot c(0)$, $\gamma(1) = c(m)$.
(b) *If* $n = p$ *is prime, then*
  $\gamma(m) = c(pm)$ *if* p *does not divide* m;
  $\gamma(m) = c(pm) + p^{2k-1} c(m/p)$ *if* $p|m$.
(c) *If* f *is a cusp form, then so also is* $T(n) \cdot f$.

PROOF. These are all obvious consequences of the proposition. □

Thus the $T(n)$'s act on the vector spaces $\mathcal{M}_k(\Gamma(1))$ and $\mathcal{S}_k(\Gamma(1))$, and satisfy the identities

$T(m) \circ T(n) = T(mn)$ if m and n relatively prime;

$T(p) \circ T(p^n) = T(p^{n+1}) + p^{2k-1} \cdot T(p^{n-1})$ if p is prime and $n \geq 1$.

PROPOSITION 5.18. *Let $f = \sum c(n)q^n$ be a nonzero modular form of weight $2k$. Assume f is a simultaneous eigenform for all the $T(n)$, say,*

$$T(n) \cdot f = \lambda(n) \cdot f, \quad \lambda(n) \in \mathbb{C}.$$

Then $c(1) \neq 0$, and if f is normalized so that $c(1) = 1$, then

$$c(n) = \lambda(n)$$

for all $n \geq 1$.

PROOF. We have seen that the coefficient of q in $T(n) \cdot f$ is $c(n)$. But, it is also $\lambda(n) \cdot c(1)$, and so $c(n) = \lambda(n) \cdot c(1)$. If $c(1)$ were zero, then all $c(n)$ would be zero, and f would be constant, which is impossible. □

COROLLARY 5.19. *Two normalized eigenforms of the same weight with the same eigenvalues are equal.*

PROOF. The proposition implies that the coefficients of their Fourier expansions are equal. □

COROLLARY 5.20. *If $f = \sum c(n)q^n$ is a normalized eigenform for the $T(n)$, then*

$c(m) \cdot c(n) = c(mn)$ if m and n are relatively prime,

$c(p) \cdot c(p^n) = c(p^{n+1}) + p^{2k-1}c(p^{n-1})$ if p is prime and $n \geq 1$.

PROOF. We know that these relations hold for the eigenvalues. □

With a modular form f, we can associate a Dirichlet series

$$L(f, s) = \sum_{n \geq 1} c(n) \cdot n^{-s}.$$

The series $\sum n^{-s}$ converges for $\Re(s) > 1$. The bounds on the values $|c(n)|$ (see Section 4) show that $L(f, s)$ converges to the right of some vertical line (if one accepts Deligne's theorem and f is a cusp form of weight $2k$, it converges for $\Re(s - k + \frac{1}{2}) > 1$, i.e., for $s > k + \frac{1}{2}$).

PROPOSITION 5.21. *For any normalized eigenform f,*

$$L(f, s) = \prod_p \frac{1}{1 - c(p)p^{-s} + p^{2k-1-2s}}.$$

PROOF. This follows from 5.20, as in the proof of (5.1). □

The Hecke operators for $\Gamma(1)$ *are Hermitian.* Before proving this, we make a small excursion.

Write $\mathrm{GL}_2(\mathbb{R})^+$ for the group of real 2×2 matrices with positive determinant. Let

$$\alpha = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}_2(\mathbb{R})^+,$$

and let f be a function on $\mathbb{H}$; we define

$$f|_k\alpha = (\det\alpha)^k \cdot (cz+d)^{-2k} \cdot f(\frac{az+b}{cz+d}).$$

For example, if $\alpha = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}$, then $f|_k\alpha = a^{2k}\cdot a^{-2k}\cdot f(z) = f(z)$; i.e., the centre of $GL_2(\mathbb{R})^+$ acts trivially. Note that f is weakly modular of weight $2k$ for $\Gamma \subset \Gamma(1)$ if and only if $f|_k\alpha = f$ for all $\alpha \in \Gamma$.

Recall that

$$T(n)\cdot f(z) = n^{2k-1}\cdot \sum d^{-2k}\cdot f(\frac{az+b}{d})$$

—sum over a, b, d, $ad = n$, $a \geq 1$, $0 \leq b < d$. We can restate this as

$$T(n)\cdot f = \sum n^{k-1}\cdot f|_k\alpha$$

where the α's run through a particular set of representatives for the orbits $\Gamma(1)\backslash M(n)$. It is clear from the above remarks, that the right hand side is independent of the choice of the set of representatives.

Recall, that the Petersson inner product of two cusp forms f and g for $\Gamma(1)$ is

$$<f,g> = \iint_D f\cdot\bar{g}\cdot y^{2k-2}\cdot dxdy$$

where $z = x+iy$ and D is any fundamental domain for $\Gamma(1)$.

LEMMA 5.22. *For any* $\alpha \in \mathrm{GL}_2(\mathbb{R})^+$,

$$<f|_k\alpha, g|_k\alpha> = <f,g>.$$

PROOF. Write $\omega(f,g) = f(z)\bar{g}(z)y^{k-2}dxdy$, where $z = x+iy$. I claim that

$$\omega(f|_k\alpha, g|_k\alpha) = \alpha^*\omega(f,g),$$

and so

$$\iint_D \omega(f|_k\alpha, g|_k\alpha) = \iint_D \alpha^*\omega(f,g) = \iint_{\alpha D}\omega(f,g),$$

which equals $<f,g>$ because αD is also a fundamental domain for $\Gamma(1)$.

Since multiplying α by a scalar changes neither $\omega(f|_k\alpha, g|_k\alpha)$ nor $\alpha^*\omega(f,g)$, we can assume (in proving the claim) that $\det\alpha = 1$. Then

$f|_k\alpha = (cz+d)^{-2k}\cdot f(\alpha z)$

$\bar{g}|_k\alpha = (c\bar{z}+d)^{-2k}\cdot \overline{g(\alpha z)}$

and so

$\omega(f|_k\alpha, g|_k\alpha) = |cz+d|^{-4k}\cdot f(\alpha z)\cdot\overline{g(\alpha z)}\cdot dx\cdot dy.$

On the other hand (see the proof of 4.30)

$\Im(\alpha z) = \Im(z)/|cz+d|^2$
$\alpha^*(dx \cdot dy) = dx \cdot dy/|cz+d|^4,$
and so
$\alpha^*(\omega(f,g)) = f(\alpha z)\cdot\overline{g(\alpha z)}\cdot|cz+d|^{4-4k}\cdot y^{2k-2}\cdot|cz+d|^{-4}\cdot dx\cdot dy = \omega(f|_k\alpha, g|_k\alpha)$. □

Note that the lemma implies that

$$<f|_k\alpha, g>=<f, g|_k\alpha^{-1}>, \text{ all } \alpha \in \mathrm{GL}_2(\mathbb{R})^+.$$

THEOREM 5.23. *For cusp forms f, g of weight $2k$*

$$<T(n)f, g>=<f, T(n)g>, \text{ all } n.$$

Because of (5.10), it suffices to prove the theorem for $T(p)$, p prime. Recall that $M(n)$ is the set of integer matrices with determinant n.

LEMMA 5.24. *There exists a common set of representatives $\{\alpha_i\}$ for the set of left orbits $\Gamma(1)\backslash M(p)$ and for the set of right orbits $M(p)/(1)$.*

PROOF. Let α, $\beta \in M(p)$; then (see 5.15)

$$\Gamma(1)\cdot\alpha\cdot\Gamma(1) = \Gamma(1)\cdot\begin{pmatrix}1 & 0\\ 0 & p\end{pmatrix}\cdot\Gamma(1) = \Gamma(1)\cdot\beta\cdot\Gamma(1).$$

Hence there exist elements $u, v, u', v' \in \Gamma(1)$ such that

$$u\alpha v = u'\beta v'$$

and so $u'^{-1}u\alpha = \beta v' v^{-1}, = \gamma$ say. Then $\Gamma(1)\cdot\alpha = \Gamma(1)\cdot\gamma$ and $\beta\cdot\Gamma(1) = \gamma\cdot\Gamma(1)$. □

For $\alpha = \begin{pmatrix}a & b\\ c & d\end{pmatrix} \in M(p)$, set $\alpha' = \begin{pmatrix}d & -b\\ -c & a\end{pmatrix} = p\cdot\alpha^{-1} \in M(p)$. Let α_i be a set of common representatives for $\Gamma(1)\backslash M(p)$ and $M(p)/\Gamma(1)$, so that

$$M(p) = \bigcup_i \Gamma(1)\cdot\alpha_i = \bigcup_i \alpha_i\cdot\Gamma(1) \text{ (disjoint unions)}.$$

Then

$$M(p) = p\cdot M(p)^{-1} = \bigcup p\cdot\Gamma(1)\cdot{\alpha_i}^{-1} = \bigcup \Gamma(1)\cdot\alpha_i'.$$

Therefore,

$$<T(p)f, g>= p^{k-1}\sum_i <f|_k\alpha_i, g>= p^{k-1}\sum_i <f, g|_k\alpha_i^{-1}>= p^{k-1}\sum_i <f, g|_k\alpha_i'>=<f, T(p)g>.$$

The $\mathbb{Z}$-structure on the space of modular forms for $\Gamma(1)$. Recall (4.20) that the Eisenstein series

$$G_k(z) \overset{\text{df}}{=} \sum_{(m,n)\neq(0,0)} \frac{1}{(mz+n)^{2k}} = 2\zeta(2k) + 2\frac{(2\pi i)^{2k}}{(2k-1)!}\sum_{n=1}^{\infty}\sigma_{2k-1}(n)q^n, \quad q = e^{2\pi i z}.$$

For $k \geq 1$, define the *normalized Eisenstein series*

$$E_k(z) = G_k(z)/2\zeta(2k).$$

Then, using that $\zeta(2k) = \frac{2^{2k-1}}{(2k)!}B_k\pi^{2k}$, one finds that

$$E_k(z) = 1 + \gamma_k \sum_{n=1}^{\infty} \sigma_{2k-1}(n)q^n, \quad \gamma_k = (-1)^k \cdot \frac{4k}{B_k} \in \mathbb{Q}.$$

For example,

$$\begin{aligned} E_2(z) &= 1 + 240 \sum_{n=1}^{\infty} \sigma_3(n)q^n, \\ E_3(z) &= 1 - 504 \sum_{n=1}^{\infty} \sigma_5(n)q^n, \\ &\dots \\ E_6(z) &= 1 + \frac{54600}{691} \sum_{n=1}^{\infty} \sigma_{11}(n)q^n. \end{aligned}$$

Note that $E_2(z)$ and $E_3(z)$ have integer coefficients.

LEMMA 5.25. *The Eisenstein series G_k, $k \geq 2$, is an eigenform of the $T(n)$, with eigenvalue is $\sigma_{2k-1}(n)$. The normalized eigenform is ${\gamma_k}^{-1} \cdot E_k$. The corresponding Dirichlet series is*

$$\zeta(s) \cdot \zeta(s - 2k + 1).$$

PROOF. The short proof that G_k is an eigenform, is to observe that $\mathcal{M}_k = \mathcal{S}_k \oplus < G_k >$, and that $T(n) \cdot G_k$ is orthogonal to $\mathcal{S}_k$ (because G_k is, $T(n)$ is Hermitian, and $T(n)$ preserves $\mathcal{S}_k$). Therefore $T(n) \cdot G_k$ is a multiple of G_k.

The computational proof starts from the definition

$$G_k(\Lambda) = \sum_{\lambda \in \Lambda, \lambda \neq 0} \frac{1}{\lambda^{2k}}.$$

Therefore

$$T(p) \cdot G_k(\Lambda) = \sum_{\Lambda'} \sum_{\lambda \in \Lambda', \lambda \neq 0} \frac{1}{\lambda^{2k}}$$

where the outer sum is over the lattices Λ' of index p in Λ. If $\lambda \in p\Lambda$, it lies in all Λ', and so contributes $(p+1)/\lambda^{2k}$ to the sum. Otherwise, it lies in only one lattice Λ', namely $p\Lambda + \mathbb{Z}\lambda$, and so it contributes $1/\lambda^{2k}$. Hence

$$(T(p) \cdot G_k)(\Lambda) = G_k(\Lambda) + p \sum_{\lambda \in p\Lambda, \lambda \neq 0} \frac{1}{\lambda^{2k}} = G_k(\Lambda) + p^{1-2k}G_k(\Lambda) = (1 + p^{1-2k})G_k(\Lambda).$$

Therefore $G_k(\Lambda)$, as a function on $\mathcal{L}$, is an eigenform of $T(p)$, with eigenvalue $1+p^{1-2k}$. As a function on $\mathbb{H}$ it is an eigenform with eigenvalue $p^{2k-1}(1 + p^{1-2k}) = p^{2k-1} + 1 = \sigma_{2k-1}(p)$.

The normalized eigenform is

$$\gamma_k^{-1} + \sum_{n=1}^{\infty} \sigma_{2k-1}(n)q^n, \quad \gamma_k = (-1)^k \cdot \frac{4k}{B_k},$$

and the associated Dirichlet series is

$$\sum_{n=1}^{\infty} \frac{\sigma_{2k-1}(n)}{n^s} = \sum_{a,d\geq 1} \frac{a^{2k-1}}{a^s d^s} = (\sum_{d\geq 1} \frac{1}{d^s})(\sum_{a\geq 1} \frac{1}{a^{s+1-2k}}) = \zeta(s) \cdot \zeta(s-2k+1).$$

$\square$

Let V be a vector space over $\mathbb{C}$. By a $\mathbb{Z}$-*structure* on V, I mean a $\mathbb{Z}$-module $V_0 \subset V$ which is free of rank equal to the dimension of V. Equivalently, it is a $\mathbb{Z}$-submodule that is freely generated by a $\mathbb{C}$-basis for V, or a $\mathbb{Z}$-submodule such that the natural map $V_0 \otimes_{\mathbb{Z}} \mathbb{C} \to V$ is an isomorphism (or a full lattice in V).

Let $\mathcal{M}_k(\mathbb{Z})$ be the $\mathbb{Z}$-submodule of $\mathcal{M}_k(\Gamma(1))$ consisting of modular forms $f = \sum_{n=0}^{\infty} a_n q^n$ with the $a_n \in \mathbb{Z}$.

PROPOSITION 5.26. *The module $\mathcal{M}_k(\mathbb{Z})$ is a $\mathbb{Z}$-structure on $\mathcal{M}_k(\Gamma(1))$.*

PROOF. Recall that $\oplus \mathcal{M}_k(\mathbb{C}) = \mathbb{C}[G_2, G_3] = \mathbb{C}[\mathbb{E}_2, \mathbb{E}_3]$. It suffices to show that

$$\oplus_k \ \mathcal{M}_k(\mathbb{Z}) = \mathbb{Z}[E_2, E_3].$$

Note that $E_2(z)$, $E_3(z)$, and $\Delta' = q\prod(1-q^n)^{24}$ all have integer coefficients. We prove by induction on k that $\mathcal{M}_k(\mathbb{Z})$ is the $2k^{th}$-graded piece of $\mathbb{Z}[E_2, E_3]$ (here E_2 has degree 4 and E_3 has degree 6). Given $f(z) = \sum a_n q^n$, $a_n \in \mathbb{Z}$, write

$$f = a_0 E_2^a \cdot E_3^b + \Delta \cdot g$$

with $4a + 6b = 2k$, and $g \in \mathcal{M}_{k-12}$. Then $a_0 \in \mathbb{Z}$, and one checks by explicit calculation that $g \in \mathcal{M}_{k-12}(\mathbb{Z})$. $\square$

PROPOSITION 5.27. *The eigenvalues of the Hecke operators are algebraic integers.*

PROOF. Let $\mathcal{M}_k(\mathbb{Z})$ be the $\mathbb{Z}$-module of modular forms with integer Fourier coefficients. It is stabilized by $T(n)$, because.

$$T(n) \cdot f(z) = \sum_{m\geq 0} \gamma(m) \cdot q^m$$

with

$$\gamma(m) = \sum a^{2k-1} \cdot c(\frac{mn}{a^2}) \text{ (sum over } a|m, a \geq 1).$$

The matrix of $T(n)$ with respect to a basis for $\mathcal{M}_k(\mathbb{Z})$ integer coefficients, and this shows that the eigenvalues of $T(n)$ are algebraic integers. $\square$

ASIDE 5.28. For Siegel modular forms of all levels, the analogous result was only proved fairly recently (by Chai and Faltings), using difficult algebraic geometry. See Section 7.

Geometric interpretation of Hecke operators. Before discussing Hecke operators for a general group, we explain the geometric significance of Hecke operators. Fix a subgroup Γ of finite index in $\Gamma(1)$.

Let $\alpha \in \mathrm{GL}_2(\mathbb{R})^+$. Then α defines a map $x \mapsto \alpha x : \mathbb{H} \to \mathbb{H}$, and we would like to define a map $\alpha : \Gamma\backslash\mathbb{H} \to \Gamma\backslash\mathbb{H}$, $\Gamma z \mapsto$ "$\alpha\Gamma z$". Unfortunately, far from being normal in $\mathrm{GL}_2(\mathbb{R})^+$. If we try defining $\alpha(\Gamma z) = \Gamma\alpha z$ we run into the problem that the orbit $\Gamma\alpha z$ depends on the choice of z (because $\alpha^{-1}\Gamma\alpha \neq \Gamma$ in general, even if α has integer coefficients and $\Gamma = \Gamma(N)$).

In fact, $\alpha\Gamma z$ is not even a Γ-orbit. Instead, we need to consider the union of the orbits meeting $\alpha\Gamma z$, i.e., we need to look at $\Gamma\alpha\Gamma z$. Any coset (right or left) of Γ in $\mathrm{GL}_2(\mathbb{R})^+$ that meets $\Gamma\alpha\Gamma$ is contained in it, and so we can we can write

$$\Gamma\alpha\Gamma = \bigcup \Gamma\alpha_i \text{ (disjoint union)},$$

and then $\Gamma\alpha\Gamma z = \bigcup \Gamma\alpha_i z$ (disjoint union). Thus α, or better, the double coset $\Gamma\alpha$, defines a "many-valued map"

$$\Gamma\backslash\mathbb{H} \to \Gamma\backslash\mathbb{H}, \quad \Gamma z \mapsto \{\Gamma\alpha_i z\}.$$

Since "many-valued maps" don't exist in my lexicon, we shall have to see how to write this in terms of honest maps. First we give a condition on α that ensures that the "map" is at least finitely-valued.

LEMMA 5.29. *Let $\alpha \in \mathrm{GL}_2(\mathbb{R})^+$. Then $\Gamma\alpha\Gamma$ is a finite union of right (and of left) cosets if and only if α is a scalar multiple of a matrix with integer coefficients.*

PROOF. Omit. [Note that the next lemma shows that this is equivalent to $\alpha^{-1}\Gamma\alpha$ being commensurable with Γ.] □

LEMMA 5.30. *Let $\alpha \in \mathrm{GL}_2(\mathbb{R})^+$. Write*

$$\Gamma = \bigcup (\Gamma \cap \alpha^{-1}\Gamma\alpha)\beta_i \text{ (disjoint union)};$$

then

$$\Gamma\alpha = \bigcup \Gamma\alpha_i \text{ (disjoint union)}$$

with $\alpha_i = \alpha \cdot \beta_i$.

PROOF. We are given that $\Gamma = \bigcup(\Gamma \cap \alpha^{-1}\Gamma\alpha)\beta_i$. Therefore

$$\alpha^{-1}\Gamma\alpha\Gamma = \bigcup_i \alpha^{-1}\Gamma\alpha \cdot (\Gamma \cap \alpha^{-1}\Gamma\alpha) \cdot \beta_i = \bigcup_i (\alpha^{-1}\Gamma\alpha\Gamma \cap \alpha^{-1}\Gamma\alpha) \cdot \beta_i.$$

But $\alpha^{-1}\Gamma\alpha\Gamma \supset \alpha^{-1}\Gamma\alpha$, and so we can drop it from the right hand term. Therefore

$$\alpha^{-1}\Gamma\alpha\Gamma = \bigcup_i \alpha^{-1}\Gamma\alpha\beta_i.$$

On multiplying by α, we find that $\Gamma\alpha\Gamma = \bigcup_i \Gamma\alpha\beta_i$, as claimed.

If $\Gamma\alpha\beta_i = \Gamma\alpha\beta_j$, then $\beta_i{\beta_j}^{-1} \in \alpha^{-1}\Gamma\alpha$; since it also lies in Γ, this implies that $i = j$. □

Now let $\Gamma_\alpha = \Gamma \cap \alpha^{-1}\Gamma\alpha$, and write $\Gamma = \bigcup \Gamma_\alpha \cdot \beta_i$ (disjoint union). Consider

$$\begin{array}{ccc} & \Gamma_\alpha\backslash\mathbb{H} & \\ \swarrow & & \searrow \alpha \\ \Gamma\backslash\mathbb{H} & & \Gamma\backslash\mathbb{H}. \end{array}$$

The map α sends an orbit $\Gamma_\alpha \cdot x$ to $\Gamma \cdot \alpha x$—this is now well-defined—and the left hand arrow sends an orbit $\Gamma_\alpha \cdot x$ to $\Gamma \cdot x$.

Let f be a modular function, regarded as a function on $\Gamma\backslash\mathbb{H}$. Then $f \circ \alpha$ is a function on $\Gamma_\alpha\backslash\mathbb{H}$, and its "trace" $\sum f \circ \alpha \circ \beta_i$ is invariant under Γ, and is therefore a function on $\Gamma\backslash\mathbb{H}$. This function is $\sum f \circ \alpha_i = T(p) \cdot f$. Similarly, a (meromorphic) modular form can be thought of as a k-fold differential form on $\Gamma\backslash\mathbb{H}$, and $T(p)$ can be interpreted as the pull-back followed by the trace in the above diagram.

Remark 5.31. In general a diagram of finite-to-one maps

$$\begin{array}{ccc} & Y & \\ \swarrow & & \searrow \\ X & & Z. \end{array}$$

is called a *correspondence* on $X \times Z$. The simplest example is obtained by taking Y to be the graph of a map $\varphi : X \to Z$; then the projection $Y \to X$ is a bijection. A correspondence is a is a "many-valued mapping", correctly interpreted: an element $x \in X$ is "mapped" to the images in Z of its inverse images in Y. The above observation shows the Hecke operator on modular functions and forms is defined by a correspondence, which we call the *Hecke correspondence*.

The Hecke algebra. The above discussion suggests that we should define an action of double cosets $\Gamma\alpha\Gamma$ on modular forms. It is convenient first to define an abstract algebra, $\mathcal{H}(\Gamma, \Delta)$, called the *Hecke algebra*.

Let Γ be a subgroup of $\Gamma(1)$ of finite index, and let Δ be a set of real matrices with positive determinant, closed under multiplication, and such that for $\alpha \in \Delta$, the double coset $\Gamma\alpha\Gamma$ contains only finitely many left and right cosets for Γ. Define $\mathcal{H}(\Gamma, \Delta)$ to be the free $\mathbb{Z}$-module generated by the double cosets $\Gamma\alpha\Gamma$, $\alpha \in \Delta$. Thus an element of $\mathcal{H}(\Gamma, \Delta)$ is a finite sum,

$$\sum n_\alpha \Gamma\alpha\Gamma,\ \alpha \in \Delta,\ n_\alpha \in \mathbb{Z}.$$

Write $[\alpha]$ for $\Gamma\alpha\Gamma$ when it is regarded as an element of $\mathcal{H}(\Gamma, \Delta)$.

We define a multiplication on $\mathcal{H}(\Gamma, \Delta)$ as follows. Note that if $\Gamma\alpha\Gamma$ meets a right coset $\Gamma\alpha'$, then it contains it. Therefore, we can write $\Gamma\alpha\Gamma = \cup\Gamma\alpha_i$, $\Gamma\beta\Gamma = \cup\Gamma\beta_i$ (finite disjoint unions). Then $\Gamma\alpha\Gamma \cdot \Gamma\beta\Gamma = \Gamma\alpha\Gamma\beta\Gamma = \cup\Gamma\alpha\Gamma\beta_j = \cup_{i,j}\Gamma\alpha_i\beta_j$; therefore $\Gamma\alpha\Gamma\beta\Gamma$ is a finite union of double cosets. Define

$$[\alpha] \cdot [\beta] = \sum c^\gamma_{\alpha,\beta} \cdot [\gamma]$$

where the union is over the $\gamma \in \Delta$ such that $\Gamma\gamma\Gamma \subset \Gamma\alpha\Gamma\beta\Gamma$, and $c^\gamma_{\alpha,\beta}$ is the number of pairs (i, j) with $\Gamma\alpha_i\beta_j = \Gamma\gamma$.

Example 5.32. Let $\Gamma = \Gamma(1)$, and let Δ be the set of matrices with integer coefficients and positive determinant. Then $\mathcal{H}(\Gamma, \Delta)$ is the free abelian group on the generators

$$\Gamma(1)\begin{pmatrix} a & b \\ 0 & d \end{pmatrix}\Gamma(1),\ a|d,\ ad>0,\ a\geq 1,\ a,d\in\mathbb{Z}.$$

Write $T(a,d)$ for the element $\Gamma(1)\begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}\Gamma(1)$ of $\mathcal{H}(\Gamma,\Delta)$. Thus $\mathcal{H}(\Gamma,\Delta)$ has a quite explicit set of free generators, and it is possible to write down (complicated) formulas for the multiplication.

For a prime p, we define $T(p)$ to be the element $T(1,p)$ of $\mathcal{H}(\Gamma,\Delta)$. We would like to define

$$T(n)=M(n)\overset{\text{df}}{=}\{\text{matrices with integer coefficients and determinant } n\}.$$

We can't do this because $M(n)$ is not a double coset, but it is a finite union of double cosets (see 5.15), namely,

$$M(n)=\bigcup\Gamma(1)\cdot\begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}\cdot\Gamma(1),\quad a|d,\quad ad=n,\quad a\geq 1,\quad a,d\in\mathbb{Z}.$$

This suggests defining

$$T(n)=\sum T(a,d),\quad a|d,\quad ad=n,\quad a\geq 1,\quad a,d\in\mathbb{Z}.$$

As before, we let $\mathcal{D}$ be the free abelian group on the set of lattices $\mathcal{L}$ in $\mathbb{C}$. A double coset $[\alpha]$ acts on $\mathcal{D}$ according to the rule:

$$[\alpha]\cdot\Lambda=\alpha\Lambda.$$

(To compute $\alpha\Lambda$, choose a basis $\begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}$ for Λ, and let $\alpha\Lambda$ be the lattice with basis $\alpha\cdot\begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}$; this is independent of the choice of the basis, and of the choice of a representative for the double coset $\Gamma\alpha\Gamma$.) We extend this by linearity to an action of $\mathcal{H}(\Gamma,\Delta)$ on $\mathcal{D}$. It is immediate from the various definitions that $T(n)$ (element of $\mathcal{H}(\Gamma,\Delta)$) acts on $\mathcal{D}$ as the $T(n)$ on defined at the start of this section. The relation in (5.10) implies that the following relation holds in the ring $\mathcal{H}(\Gamma,\Delta)$:

$$T(n)T(m)=\sum_{d|\gcd(m,n)} d\cdot T(d,d)\cdot T(nm/d^2)\qquad (*)$$

In particular, for relatively prime integers m and n,

$$T(n)T(m)=T(nm),$$

and for a prime p,

$$T(p)\cdot T(p^n)=T(p^{n+1})+p\cdot T(p,p)\cdot T(p^{n-1}).$$

The ring $\mathcal{H}(\Gamma,\Delta)$ acts on the set of functions on $\mathcal{L}$:

$$[\alpha]\cdot F=\sum F(\alpha_i\Lambda)\text{ if }\Gamma\alpha=\cup\Gamma\alpha_i.$$

The relation $(*)$ implies that

$$T(n)\cdot T(m)\cdot F=\sum_{d|\gcd(m,n)} d^{1-2k}\cdot T(mn/d^2)\cdot F$$

for F a function on $\mathcal{L}$ of weight $2k$.

Finally, we make $\mathcal{H}(\Gamma,\Delta)$ act on $\mathcal{M}_k(\Gamma(1))$ by

$$[\alpha]\cdot f=\det(\alpha)^{k-1}\cdot\sum f|_k\alpha_i \qquad (**)$$

if $\Gamma\alpha\Gamma=\cup\Gamma(1)\cdot\alpha_i$. Recall that $f|_k\alpha=(\det\alpha)^k\cdot(cz+d)^k\cdot f(\frac{az+b}{cz+d})$, if $\alpha=\begin{pmatrix} a & b\\ c & d\end{pmatrix}$. The element $T(n)\in\mathcal{H}(\Gamma,\Delta)$ acts on $\mathcal{M}_k(\Gamma(1))$ as in the old definition, and (*) implies that

$$T(n)\cdot T(m)\cdot f=\sum_{d|\gcd(m,n)} d^{2k-1}\cdot T(\frac{mn}{d^2})\cdot f.$$

We now define a Hecke algebra for $\Gamma(N)$. For this we take $\Delta(N)$ to be the set of integer matrices α such that $n\overset{\mathrm{df}}{=}\det(\alpha)$ is positive and prime to N, and $\alpha\equiv\begin{pmatrix}1&0\\0&n\end{pmatrix}$ mod N.

LEMMA 5.33. *Let $\Delta'(N)$ be the set of integer matrices with positive determinant prime to N. Then the map*

$$\Gamma(N)\cdot\alpha\cdot\Gamma(N)\mapsto\Gamma(1)\cdot\alpha\cdot\Gamma(1):\mathcal{H}(\Gamma(N),\Delta(N))\to\mathcal{H}(\Gamma(1),\Delta'(N))$$

is an isomorphism.

PROOF. Elementary. (See Ogg 1969, pIV-10.) □

Let $T^N(a,d)$ and $T^N(n)$ be the elements of $\mathcal{H}(\Gamma(N),\Delta(N))$ corresponding to $T(a,d)$ and $T(n)$ in $\mathcal{H}(\Gamma(1),\Delta'(N))$ under the isomorphism in the lemma. Note that $\mathcal{H}(\Gamma(1),\Delta'(N))$ is a subring of $\mathcal{H}(\Gamma(1),\Delta)$. From the identity $(*)$, we obtain the identity

$$T^N(n)T^N(m)=\sum_{d|\gcd(n,m)} d\cdot T^N(d,d)\cdot T^N(mn/d^2) \qquad (***)$$

for $(mn,N)=1$.

When we let $\mathcal{H}(\Gamma(N),\Delta(N))$ act on $\mathcal{M}_k(\Gamma(N))$ by the rule $(**)$, the identity $(*)$ translates into a slightly different identity for operators on $\mathcal{M}_k(\Gamma(N))$. (The key point is that $\begin{pmatrix}d&0\\0&d\end{pmatrix}\in\Delta'(N)$ if $\gcd(d,N)=1$ but not $\Delta(N)$—see Ogg 1969, pIV-12). For $f\in\mathcal{M}_k(\Gamma(N))$, we have the identity is

$$T^N(n)\cdot T^N(m)\cdot f=\sum_{d|m,n} d^{2k-1}\cdot R_d\cdot T^N(mn/d^2)\cdot f$$

for $(mn,N)=1$. Here R_d is a matrix in $\Gamma(1)$ such that $R_d\equiv\begin{pmatrix}d^{-1}&0\\0&d\end{pmatrix}$ mod N.

The term R_d causes problems. Let $V=\mathcal{M}_k(\Gamma(N))$. If $d\equiv 1$ mod N, then $R_d\in\Gamma(N)$, and so it acts as the identity map on V. Therefore $d\mapsto R_d$ defines an action of $(\mathbb{Z}/N\mathbb{Z})^\times$ on V, and so we can decompose V into a direct sum,

$$V=\oplus V(\varepsilon),$$

over the characters ε of $(\mathbb{Z}/N\mathbb{Z})^\times$, where

$$V(\varepsilon) = \{f \in V \mid f|R_d = \varepsilon(d) \cdot f\}.$$

LEMMA 5.34. *The operators R_n and $T^N(m)$ on V commute for $(nm, N) = 1$. Hence $V(\varepsilon)$ is invariant under $T^N(m)$.*

PROOF. See Ogg 1969, pIV-13. □

Let $\mathcal{M}_k(\Gamma(N), \varepsilon) = V(\varepsilon)$. Then $T^N(n)$ acts on $\mathcal{M}_k(\Gamma(N), \varepsilon)$ with the basic identity:

$$T^N(n) \cdot T^N(m) = \sum_{d|\gcd(n,m)} d^{2k-1} \cdot \varepsilon(d) \cdot T^N(nm/d^2),$$

for $(nm, N) = 1$.

PROPOSITION 5.35. *Let $f \in \mathcal{M}_k(\Gamma(N), \varepsilon)$ have the Fourier expansion $f = \sum a_n q^n$. Assume that f is an eigenform for all $T^N(n)$, and normalize it so that $a_1 = 1$. Then*

$$L_N(f, s) \overset{\text{df}}{=} \sum_{\gcd(n,N)=1} a_n n^{-s} = \prod_{\gcd(p,N)=1} \frac{1}{(1 - a_p p^{-s} + \varepsilon(p) p^{2k-1-2s})}.$$

PROOF. Essentially the same as the proof of Proposition 5.1. □

Let $U = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ (it would be too confusing to continue denoting it as T). Then $U^N \in \Gamma(N)$, and so

$$f \mapsto f|_k U^m = f(z+m)$$

defines an action of $\mathbb{Z}/N\mathbb{Z}$ on $V \overset{\text{df}}{=} \mathcal{M}_k(\Gamma(N))$. We can decompose V into a direct sum over the characters of $\mathbb{Z}/N\mathbb{Z}$. But the characters of $\mathbb{Z}/N\mathbb{Z}$ are parametrized by the N^{th} roots of one in $\mathbb{C}$—the character corresponding to ζ is $m \mod N \mapsto \zeta^m$. Thus

$$V = \oplus V(\zeta),\ \zeta \text{ an } N^{th} \text{ root of } 1,$$

where $V(\zeta) = \{f \in V \mid U^m \cdot f = \zeta^m f\}$. Alas $V(\zeta)$ is not invariant under $T^N(n)$. To remedy this, we have to consider, for each $t|N$,

$$V(t) = \oplus V(\zeta),\ \zeta \text{ a primitive } (N/t)^{th} \text{ root of } 1.$$

Let m be an integer divisible only by the primes dividing N; we define

$$T^t(m) = m^{k-1} \cdot \sum_{0 \le b < m} f|_k \begin{pmatrix} 1 & bN/t \\ 0 & m \end{pmatrix}.$$

For a general $n > 1$, we write $n = mn_0$ with $\gcd(n_0, N) = 1$, and set

$$T^t(n) = T(n_0) \cdot T^t(m).$$

We then have the relation:

$$T^t(n) \cdot T^t(m) \cdot f = \sum_{d|\gcd(n,m)} \varepsilon(d) d^{2k-1} T^t(nm/d^2) \cdot f$$

for $f \in V(,t) \overset{\text{df}}{=} V(\varepsilon) \cap V(t)$.

THEOREM 5.36. *Let $f \in V(t, \varepsilon)$ have the Fourier expansion $f(z) = \sum a_n q^n$. If $a_1 = 1$ and f is an eigenform for all the $T^t(n)$ with* $\gcd(n, N) = 1$*, then the associated Dirichlet series has the Euler product expansion*

$$\sum a_n n^{-s} = \prod_p \frac{1}{1 - a_p p^{-s} + \varepsilon(p) p^{2k-1-2s}}.$$

PROOF. See Ogg 1969, pIV-10. □

In the statement of the theorem, we have extended ε from $(\mathbb{Z}/N\mathbb{Z})^\times$ to $\mathbb{Z}/N\mathbb{Z}$ by setting $\varepsilon(p) = 0$ for $p|N$. Thus $\varepsilon(p) = 0$ if $p|N$, and $a_p = 0$ if $p|\frac{N}{t}$. This should be compared with the L-series of an elliptic curve E with conductor N, where the p-factor of the L-series is $(1 \pm p^{-s})^{-1}$ if $p|N$ but p^2 does not divide N, and is 1 if $p^2|N$.

PROPOSITION 5.37. *Let f and g be cusp forms for $\Gamma(N)$ of weight $2k$ and character ε. Then*

$$<T(n) \cdot f, g> = \varepsilon(n) <f, T(n) \cdot g>$$

PROOF. See Ogg 1969, pIV-24. □

Unlike the case of forms for $\Gamma(1)$, this does not imply that the eigenvalues are real. It does imply that $\mathcal{M}_k(\Gamma(N), \varepsilon, t)$ has a basis of eigenforms for the $T(n)$ with $\gcd(n, N) = 1$ (but not for all $T(n)$'s).

For a summary of the theory of Hecke operators for $\Gamma_0(N)$, see Math 679, especially Section 26.

Part II: Applications to Arithmetic Geometry

In this part we apply the preceding theory, first to obtain elliptic modular curves defined over number fields, and then to study the zeta functions of modular curves and of elliptic curves. There is considerable overlap between this part and Math 679.

6. The Modular Equation for $\Gamma_0(N)$

For any congruence subgroup Γ of $\Gamma(1)$, the algebraic curve $\Gamma\backslash\mathbb{H}^*$ is defined over a specific number field. As a first step toward proving this general statement, we find in this section a canonical polynomial $F(X,Y)$ with coefficients in $\mathbb{Q}$ such that the curve $F(X,Y)=0$ is birationally equivalent to $X_0(N) \stackrel{\mathrm{df}}{=} \Gamma_0(N)\backslash\mathbb{H}^*$.

Recall that

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

If $\gamma = \begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix}$, then

$$\begin{pmatrix} N^{-1} & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a & N^{-1}b \\ Nc & d \end{pmatrix} \text{ for } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma(1),$$

and so

$$\Gamma_0(N) = \Gamma(1) \cap \gamma^{-1}\Gamma(1)\gamma.$$

Note that $-I \in \Gamma_0(N)$. In the map

$$\mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$$

the image of $\Gamma_0(N)$ is the group of all matrices of the form $\begin{pmatrix} a & b \\ 0 & a^{-1} \end{pmatrix}$ in $\mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$. This group obviously has order $N \cdot \varphi(N)$, and so (cf. 2.23),

$$\mu \stackrel{\mathrm{df}}{=} (\bar{\Gamma}(1) : \bar{\Gamma}_0(N)) = (\Gamma(1) : \Gamma_0(N)) = N \cdot \prod_{p|N}(1+\frac{1}{p}).$$

(Henceforth, $\bar{\Gamma}$ denotes the image of Γ in $\mathrm{SL}_2(\mathbb{Z})/\{\pm I\}$.) Consider the set of pairs (c,d) of positive integers satisfying:

$$\gcd(c,d) = 1, \quad d|N, \quad 0 \le c < N/d. \qquad (*)$$

For each such pair, we choose a pair (a,b) of integers such that $ad - bc = 1$. Then the matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ form a set of representatives for $\Gamma_0(N)\backslash\Gamma(1)$. (Check that they are not equivalent under left multiplication by elements of $\Gamma_0(N)$, and that there is the correct number.)

If $4|N$ then $\bar{\Gamma}_0(N)$ contains no elliptic elements of order 2, and if $9|N$ then it contains no elliptic elements of order 3. The cusps for $\Gamma_0(N)$ are represented by the pairs (c,d) satisfying (*), modulo the equivalence relation:

$$(c,d) \sim (c',d') \text{ if } d = d' \text{ and } c' = c + m, \text{ some } m \in \mathbb{Z}.$$

For each d, there are exactly $\varphi(\gcd(d, N/d))$ inequivalent pairs, and so the number of cusps is

$$\sum_{d|N, d>0} \varphi(\gcd(d, \frac{N}{d})).$$

It is now possible to use Theorem 2.22 to compute the genus of $X_0(N)$. (See Shimura 1971, p25, for more details on the above material.)

THEOREM 6.1. *The field $\mathbb{C}(X_0(N))$ of modular functions for $\Gamma_0(N)$ is generated (over $\mathbb{C}$) by $j(z)$ and $j(Nz)$. The minimum polynomial $F(j,Y) \in \mathbb{C}(j)[Y]$ of $j(Nz)$ over $\mathbb{C}(j)$ has degree μ. Moreover, $F(j,Y)$ is a polynomial in j and has coefficients in $\mathbb{Z}$, i.e., $F(X,Y) \in \mathbb{Z}[X,Y]$. When $N > 1$, $F(X,Y)$ is symmetric in X and Y, and when $N = p$ is prime,*

$$F(X,Y) \equiv X^{p+1} + Y^{p+1} - X^pY^p - XY \mod p.$$

PROOF. Let $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ be an element of $\Gamma_0(N)$ with $c = Nc'$, $c' \in \mathbb{Z}$. Then

$$j(N\gamma z) = j\left(\frac{Naz + Nb}{cz + d}\right) = j\left(\frac{Naz + Nb}{Nc'z + d}\right) = j\left(\frac{a(Nz) + Nb}{c'(Nz) + d}\right) = j(Nz)$$

because $\begin{pmatrix} a & Nb \\ c' & d \end{pmatrix} \in \Gamma(1)$. Therefore $\mathbb{C}(j(z), j(Nz))$ is contained in the field of modular functions for $\Gamma_0(N)$.

The curve $X_0(N)$ is a covering of $X(1)$ of degree $\mu = (\Gamma(1) : \Gamma_0(N))$. From Proposition 1.16 we know that the field of meromorphic functions $\mathbb{C}(X_0(N))$ on $X_0(N)$ has degree μ over $\mathbb{C}(X(1)) = \mathbb{C}(j)$, but we shall prove this again. Let $\{\gamma_1 = 1, ..., \gamma_\mu\}$ be a set of representatives for the right cosets of $\Gamma_0(N)$ in $\Gamma(1)$, so that,

$$\Gamma(1) = \cup \Gamma_0(N)\gamma_i \qquad \text{(disjoint union)}.$$

For any $\gamma \in \Gamma(1)$, $\{\gamma_1\gamma, ..., \gamma_\mu\gamma\}$ is also a set of representatives for the right cosets of $\Gamma_0(N)$ in $\Gamma(1)$—the set $\{\Gamma_0(N)\gamma_i\gamma\}$ is just a permutation of the set $\{\Gamma_0(N)\gamma_i\}$.

If $f(z)$ is a modular function for $\Gamma_0(N)$, then $f(\gamma_i z)$ depends only on the coset $\Gamma_0(N)\gamma_i$. Hence the functions $\{f(\gamma_i\gamma z)\}$ are a permutation of the functions $\{f(\gamma_i z)\}$, and any symmetric polynomial in the $f(\gamma_i z)$ is invariant under $\Gamma(1)$; since such a polynomial obviously satisfies the other conditions, it is a modular function for $\Gamma(1)$, and hence a rational function of j. We have shown that $f(z)$ satisfies a polynomial of degree μ with coefficients in $\mathbb{C}(j)$, namely, $\prod(Y - f(\gamma_i z))$. Since this holds for every $f \in \mathbb{C}(X_0(N))$, we see that $\mathbb{C}(X_0(N))$ has degree at most μ over $\mathbb{C}(j)$.

Next I claim that all the $f(\gamma_i z)$ are conjugate to $f(z)$ over $\mathbb{C}(j)$: for let $F(j,Y)$ be the minimum polynomial of $f(z)$ over $\mathbb{C}(j)$; in particular, $F(j,Y)$ is monic and irreducible when regarded as a polynomial in Y with coefficients in $\mathbb{C}(j)$; on replacing z with $\gamma_i z$ and remembering that $j(\gamma_i z) = j(z)$, we find that $F(j(z), f(\gamma_i z)) = 0$, which proves the claim.

If we can show that the functions $j(N\gamma_i z)$ are distinct, then it will follow that the minimum polynomial of $j(Nz)$ over $\mathbb{C}(j)$ has degree μ; hence $[\mathbb{C}(X_0(N)) : \mathbb{C}(j)] = \mu$ and $\mathbb{C}(X_0(N)) = \mathbb{C}(j(z))[j(Nz)]$.

Suppose $j(N\gamma_i z) = j(N\gamma_{i'} z)$ for some $i \neq i'$. Recall that j defines an isomorphism $\Gamma(1)\backslash\mathbb{H}^* \to$ (Riemann sphere), and so $j(N\gamma_i z) = j(N\gamma_{i'} z)$ all z implies that there exists a $\gamma \in \Gamma(1)$ such that $N\gamma_i z = \gamma N\gamma_{i'} z$ all z, and this implies that

$$\begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix} \gamma_i = \pm\gamma \begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix} \gamma_{i'}.$$

Hence $\gamma_i \gamma_{i'}^{-1} \in \Gamma(1) \cap \begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix}^{-1} \Gamma(1) \begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix} = \Gamma_0(N)$, and this contradicts the fact that γ_i and $\gamma_{i'}$ lie in different cosets.

The minimum polynomial of $j(Nz)$ over $\mathbb{C}(j)$ is $F(j, Y) = \prod(Y - j(N\gamma_i z))$. The symmetric polynomials in the $j(N\gamma_i z)$ are holomorphic on $\mathbb{H}$. As they are rational functions of $j(z)$, they must in fact be polynomials in $j(z)$, and so $F(X, Y) \in \mathbb{C}[X, Y]$ (rather than $\mathbb{C}(X)[Y]$).

But we know (4.22) that

$$j(z) = q^{-1} + \sum_{n=0}^{\infty} c_n q^n \qquad (*).$$

with the $c_n \in \mathbb{Z}$. Consider $j(N\gamma z)$ for some $\gamma = \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} \in \Gamma(1)$. Then $N\gamma z = \begin{pmatrix} Na' & Nb' \\ c' & d' \end{pmatrix} z$, and $j(N\gamma z)$ is unchanged when we act on the matrix on the left by an element of $\Gamma(1)$. Therefore (see 5.15)

$$j(N\gamma z) = j(\frac{az+b}{d})$$

for some integers a, b, d with $ad = N$. On substituting $\frac{az+b}{d}$ for z in (*) and noting that $e^{2\pi i(az+b)/d} = e^{2\pi i b/d} \cdot e^{2\pi i az/d}$, we find that $j(N\gamma z)$ has a Fourier expansion in powers of $q^{1/N}$ whose coefficients are in $\mathbb{Z}[e^{2\pi i/N}]$, and hence are algebraic integers. The same is then true of the symmetric polynomials in the $j(N\gamma_i z)$. We know that these symmetric polynomials lie in $\mathbb{C}[j(z)]$, and I claim that in fact they are polynomials in j with coefficients that are algebraic integers.

Consider a polynomial $P = \sum c_n j^n \in \mathbb{C}[j]$ whose coefficients are not all algebraic integers. If c_m is the coefficient having the largest subscript among those that are not algebraic integers, then the coefficient of q^{-m} in the q-expansion of P is not an algebraic integer, and so P can not be equal to a symmetric polynomial in the $j(N\gamma_i z)$.

Thus $F(X, Y) = \sum c_{m,n} X^m Y^n$ with the $c_{m,n}$ algebraic integers (and $c_{0,\mu} = 1$).

When we substitute (*) into the equation

$$F(j(z), j(Nz)) = 0,$$

and equate coefficients of powers of q, we obtain a set of linear equations for the $c_{m,n}$ with rational coefficients. When we adjoin the equation

$$c_{0,\mu} = 1,$$

then the equations determine the $c_{m,n}$ uniquely (because there is only one monic minimum equation for $j(Nz)$ over $\mathbb{C}(j)$). Because the system of linear equations has a solution in $\mathbb{C}$, it also has a solution in $\mathbb{Q}$; because the solution is unique, the solution in $\mathbb{C}$ must in fact lie in $\mathbb{Q}$. Thus the $c_{m,n} \in \mathbb{Q}$, but we know that they are algebraic integers, and so they lie in $\mathbb{Z}$.

Now assume $N > 1$. On replacing z with $-1/Nz$ in the equation $F(j(z), j(Nz)) = 0$, we obtain

$$F(j(-1/Nz), j(-1/z)) = 0,$$

which, because of the invariance of j, is just the equation

$$F(j(Nz), j(z)) = 0.$$

This shows that $F(Y, X)$ is a multiple of $F(X, Y)$ (recall that $F(X, Y)$ is irreducible in $\mathbb{C}(X)[Y]$, and hence in $\mathbb{C}[X, Y]$), say, $F(Y, X) = cF(X, Y)$. On equating coefficients, one sees that $c^2 = 1$, and so $c = \pm 1$. But $c = -1$ would imply that $F(X, X) = 0$, and so $X - Y$ would be a factor of $F(X, Y)$, which contradicts the irreducibility. Hence $c = 1$, and $F(X, Y)$ is symmetric.

Finally, suppose $N = p$, a prime. The argument following (*) shows in this case that the functions $j(p\gamma_i z)$ for $i \neq 1$ are exactly the functions:

$$j\left(\frac{z+m}{p}\right), \quad m = 0, 1, 2, \ldots, p-1.$$

Let $\zeta_p = e^{2\pi i/p}$, and let $\mathfrak{p}$ denote the prime ideal $(1 - \zeta_p)$ in $\mathbb{Z}[\zeta_p]$. Then $\mathfrak{p}^{p-1} = (p)$. When we regard the functions $j(\frac{z+m}{p})$ as power series in q, then we see that they are all congruent modulo $\mathfrak{p}$ (meaning that their coefficients are congruent modulo $\mathfrak{p}$), and so

$$\begin{aligned} f(j(z), Y) =_{df} (Y - j(pz)) \prod_{m=0}^{p-1} (Y - j(\frac{z+m}{p})) &\equiv (Y - j(pz))(Y - j(z/p))^p \\ &\equiv (Y - j(z)^p)(Y^p - j(z)) \pmod{\mathfrak{p}}. \end{aligned}$$

This implies the last equation in the theorem. □

EXAMPLE 6.2. For $N = 2$, the equation is

$$\begin{aligned} X^3 + Y^3 - X^2Y^2 + 1488XY(X+Y) - 162000(X^2+Y^2) + 40773375XY \\ + 8748000000(X+Y) - 157464000000000 = 0. \end{aligned}$$

Rather a lot of effort (for over a century) has been put into computing $F(X, Y)$ for small values of N. For a discussion of how to do it (complete with dirty tricks), see Birch's article in Modular Functions of One Variable, Vol I, SLN 320 (Ed. W. Kuyk).

The modular equation $F_N(X, Y) = 0$ was introduced by Kronecker, and used by Kronecker and Weber in the theory of complex multiplication. For $N = 3$, it was computed by Smith in 1878; for $N = 5$ it was computed by Berwick in 1916; for $N = 7$ it was computed by Herrmann in 1974; for $N = 11$ it was computed by MACSYMA in 1984. This last computation took 20 hours on a VAX-780; the result is a polynomial of degree 21 with coefficients up 10^{60} which takes 5 pages to write out. See Kaltofen and Yui, On the modular equation of order 11, Proc. of the Third MACSYMA's user's Conference, 1984, pp472-485.

Clearly one gets nowhere with brute force methods in this subject.

7. The Canonical Model of $X_0(N)$ over $\mathbb{Q}$

After reviewing some algebraic geometry, we define the canonical model of $X_0(N)$ over $\mathbb{Q}$.

Review of some algebraic geometry. (This is essentially Section 13 of my notes on Algebraic Geometry (Math 631).) Theorem 6.1 will allow us to define a model of $X_0(N)$ over $\mathbb{Q}$, but before explaining this I need to review some of the terminology from algebraic geometry.

First we need a slightly more abstract notion of sheaf than that on p11.

DEFINITION 7.1. A *presheaf* F on a topological space X is a map assigning to each open subset U of X a set $F(U)$ and to each inclusion $U \supset U'$ a "restriction" map

$$a \mapsto a|U' : F(U) \to F(U').$$

The restriction map corresponding to $U \supset U$ is required to be the identity map, and if $U \supset U' \supset U''$, then the restriction map $F(U) \to F(U'')$ is required to be the composite of the restriction maps $F(U) \to F(U')$ and $F(U') \to F(U'')$.

If the sets $F(U)$ are abelian groups and the restriction maps are homomorphisms, then F is called a *presheaf of abelian groups* (similarly for a sheaf of rings, modules, etc.). A presheaf F is a *sheaf* if for every open covering $\{U_i\}$ of $U \subset X$ and family of elements $a_i \in F(U_i)$ agreeing on overlaps (that is, such that $a_i|U_i \cap U_j = a_j|U_i \cap U_j$ for all i, j), there is a unique element $a \in F(U)$ such that $a_i = a|U_i$ for all i. A *ringed space* is a pair $(X, \mathcal{O}_X)$ where X is a topological space and $\mathcal{O}_X$ is a sheaf of rings on X. With the obvious notion of morphism, the ringed spaces form a category.

Let k_0 be a field, and let k be an algebraic closure of k_0. An *affine k_0-algebra* A is a finitely generated k_0-algebra A such that $A \otimes_{k_0} k$ is an integral domain.

This is stronger than saying that A itself is an integral domain—in fact, A can be an integral domain without $A \otimes_{k_0} k$ being reduced. Consider for example the algebra

$$A = k_0[X, Y]/(X^p + Y^p + a)$$

where $p = \mathrm{char}(k_0)$ and $a \notin k_0^p$; then A is an integral domain because $X^p + Y^p + a$ is irreducible, but obviously

$$A \otimes_{k_0} k = k[X, Y]/(X^p + Y^p + a) = k[X, Y]/((X + Y + \alpha)^p), \quad \alpha^p = a,$$

is not reduced. This problem arises only because of inseparability: if k_0 is perfect, then $A \otimes_{k_0} k$ is reduced whenever A is finitely generated k_0-algebra that is an integral domain. However, even then $A \otimes_{k_0} k$ need not be an integral domain—consider for example $A = k[X]/(f(X))$. We have the following criterion: a finitely generated algebra A over a perfect field k_0 is an affine k-algebra if and only if A is an integral domain and k_0 is algebraically closed in A (i.e., an element of A algebraic over k_0 is already in k_0).

EXAMPLE 7.2. An algebra $k_0[X, Y]/(f(X, Y))$ is an affine k_0-algebra if and only if $f(X, Y)$ is absolutely irreducible, i.e., it is irreducible in $k[X, Y]$.

Let A be a finitely generated k_0-algebra. We can write

$$A = k_0[x_1, ..., x_n] = k_0[X_1, ..., X_n]/(f_1, ..., f_m),$$

and then

$$A \otimes_{k_0} k = k[X_1, ..., X_n]/(f_1, ..., f_m).$$

Thus A is an affine algebra if and only if the elements $f_1, ..., f_m$ of $k_0[X_1, ..., X_n]$ generate a prime ideal when regarded as elements of $k[X_1, ..., X_n]$.

Let A be an affine k_0-algebra. Define $\operatorname{specm}(A)$ to be the set of maximal ideals in A, and endow it with the topology having as basis the sets $D(f)$, $f \in A$, where $D(f) = \{\mathfrak{m} \mid f \notin \mathfrak{m}\}$. There is a unique sheaf of k_0-algebras $\mathcal{O}$ on $\operatorname{specm}(A)$ such that $\mathcal{O}(D(f)) = A_f \overset{\text{df}}{=} A[f^{-1}]$ for all f. Here $\mathcal{O}$ is a sheaf in the abstract sense—the elements of $\mathcal{O}(U)$ are not functions on U with values in k_0, although we may wish to think of them as if they were. If $f \in A$ and $\mathfrak{m}_v \in \operatorname{specm} A$, then we define $f(v)$ to be the image of f in the $\kappa(v) \overset{\text{df}}{=} A/\mathfrak{m}_v$. Then $v \mapsto f(v)$ is not a function on $\operatorname{specm}(A)$ in the conventional sense because (unless $k_0 = k$) the fields $\kappa(v)$ are varying with v, but it does make sense to speak of the set $V(f)$ of zeros of f in X, and this zero set is the complement of $D(f)$.

The ringed space

$$\operatorname{Specm}(A) \overset{\text{df}}{=} (\operatorname{specm}(A), \mathcal{O}),$$

as well as any ringed space isomorphic to such a space, is called an *affine variety* over k_0. A ringed space $(X, \mathcal{O}_X)$ is a *prevariety* over k_0 if there is a finite covering (U_i) of X by open subsets such that $(U_i, \mathcal{O}_X|U_i)$ is an affine variety over k_0 for all i. A *morphism of prevarieties* over k_0 is a morphism of ringed spaces; in more detail, a morphism of prevarieties $(X, \mathcal{O}_X) \to (Y, \mathcal{O}_Y)$ is a continuous map $\varphi : X \to Y$ and, for every open subset U of Y, a map $\psi : \mathcal{O}_Y(U) \to \mathcal{O}_X(\varphi^{-1}(U))$ satisfying certain natural conditions. A prevariety X over k is *separated* if for all pairs of morphisms $\varphi, \psi : Z \to X$, the set where φ and ψ agree is closed in Z. A *variety* is a separated prevariety.

When $V = \operatorname{Specm} B$ and $W = \operatorname{Specm} A$, there is a one-to-one correspondence between the set of morphisms of varieties $W \to V$ and the set of homomorphisms of k_0-algebras $A \to B$. If $A = k_0[X_1, ..., X_m]/\mathfrak{a}$ and $B = k_0[Y_1, ..., Y_n]/\mathfrak{b}$, a homomorphism $A \to B$ is determined by a family of polynomials, $P_i(Y_1, ..., Y_n)$, $i = 1, ..., m$; the morphism $W \to V$ sends $(y_1, \ldots, y_n)$ to $(\ldots, P_i(y_1, ..., y_n), \ldots)$; in order to define a homomorphism, the P_i must be such that $F \in \mathfrak{a} \Rightarrow F(P_1, ..., P_n) \in \mathfrak{b}$; two families $P_1, ..., P_m$ and $Q_1, ..., Q_m$ determine the same map if and only if $P_i \equiv Q_i \bmod \mathfrak{b}$ for all i.

There is a canonical way of associating a variety X over k with a variety X_0 over k; for example, if $X_0 = \operatorname{Specm}(A)$, then $X = \operatorname{Specm}(A \otimes_{k_0} k)$. We then call X_0 a *model* for X over k_0. When $X \subset \mathbb{A}^n$, to give a model for X over k_0 is the same as to give an ideal $\mathfrak{a}_0 \subset k_0[X_1, ..., X_n]$ such that $\mathfrak{a}_0$ generates the ideal of X,

$$I(X) \overset{\text{df}}{=} \{f \in k[X_1, \ldots, X_n] \mid f = 0 \text{ on } X\}.$$

Of course, X need not have a model over k_0—for example, an elliptic curve E over k will have a model over $k_0 \subset k$ if and only if its j-invariant $j(E)$ lies in k_0. Moreover, when X has a model over k_0, it will usually have a large number of them, no two of which are isomorphic over k_0. For example, let X be a nondegenerate quadric surface in $\mathbb{P}^3$ over $\mathbb{Q}^{\mathrm{al}}$ (the algebraic closure of $\mathbb{Q}$); thus X is isomorphic to the surface

$$X^2 + Y^2 + Z^2 + W^2 = 0.$$

The models over X over $\mathbb{Q}$ are defined by equations

$$aX^2 + bY^2 + cZ^2 + dW^2 = 0, \quad a, b, c, d \in \mathbb{Q}.$$

Thus classifying the models of X over $\mathbb{Q}$ is equivalent to classifying quadratic forms over $\mathbb{Q}$ in 4 variables; this has been done, but it is quite complicated—there are an infinite number.

Let X be a variety over k_0. A *point* of X *with coordinates* in k_0, or a *point* of X *rational* over k_0, is a morphism $\operatorname{Specm} k_0 \to X$. For example, if X is affine, say $X = \operatorname{Specm} A$, then a point of X with coordinates in k_0 is a k_0-homomorphism $A \to k_0$. If $A = k[X_1, ..., X_n]/(f_1, ..., f_m)$, then to give a k_0-homomorphism $A \to k_0$ is the same as to give an n-tuple $(a_1, ..., a_n)$ such that

$$f_i(a_1, ..., a_n) = 0 \quad i = 1, ..., m;$$

thus a point of X with coordinates in k_0 is exactly what you expect it to be. Similar remarks apply to projective varieties. We write $X(k_0)$ for the points of X with coordinates in k_0.

It is possible to define the notion of a point of X with coordinates in *any* k_0-algebra R, and we write $X(R)$ for the set of such points. For example, when $X = \operatorname{Specm} A$,

$$X(R) = \operatorname{Hom}_{k-alg}(A, R).$$

When $k = k_0$, $X(k_0) = X$. What is the relation of the sets $X(k_0)$ and X when $k \neq k_0$? Let $v \in X$. Then v corresponds to a maximal ideal $\mathfrak{m}_v$ (actually, it *is* a maximal ideal), and we write $\kappa(v)$ for the residue field $\mathcal{O}_v/\mathfrak{m}_v$. It is a finite extension of k_0, and we call the degree of $\kappa(v)$ over k_0 the *degree* of v. Then $X(k_0)$ can be identified with the points v of X of degree 1. (Suppose for example that X is affine, say $X = \operatorname{Specm} A$. Then a point of X is a maximal ideal $\mathfrak{m}_v$ in A. Obviously, $\mathfrak{m}_v$ is the kernel of a k_0-homomorphism $A \to k_0$ if and only if $\kappa(v) \overset{\mathrm{df}}{=} A/\mathfrak{m}_v = k_0$, in which case it is the kernel of exactly one such homomorphism.)

The set $X(k)$ can be identified with the set of points on X_k, where X_k is the variety over k associated with X. When k_0 is perfect, there is an action of $\operatorname{Gal}(k/k_0)$ on $X(k)$, and one can show that there is a natural one-to-one correspondence between the orbits of the action and the points of X. (Again suppose $X = \operatorname{Specm} A$, and let $v \in X$; associate with v the set of k_0-homomorphisms $A \to k$ with kernel $\mathfrak{m}_v$.)

Assume k_0 is perfect. As we just noted, if X_0 is a variety over k_0, then there is an action of $\operatorname{Gal}(k/k_0)$ on $X_0(k)$. The variety $X \overset{\mathrm{df}}{=} X_{0,k}$ and the action of $\operatorname{Gal}(k/k_0)$ on $X(k)$ then determines X_0: for example, if $X = \operatorname{Specm} A$, then the action of $\operatorname{Gal}(k/k_0)$ on $X(k)$ determines an action of $\operatorname{Gal}(k/k_0)$ on A and $X_0 = \operatorname{Specm} A^{\operatorname{Gal}(k/k_0)}$.

All of the usual theory of algebraic varieties over algebraically closed fields carries over *mutatis mutandis* to varieties over a nonalgebraically closed field.

Curves and Riemann surfaces. Fix a field k_0, and let X be an algebraic variety over k_0. The function field $k_0(X)$ of X is the field of fractions of $\mathcal{O}_X(U)$ for any open affine subset U of X; for example, if $X = \operatorname{Specm} A$, then $k_0(X)$ is the field of fractions of A. The *dimension* of X is defined to be the transcendence degree of $k_0(X)$ over k_0. An *algebraic curve* is an algebraic variety of dimension 1.

To each point v of X there is attached a local ring $\mathcal{O}_v$. For example, if $X = \operatorname{Specm} A$, then a point v of X is a maximal ideal $\mathfrak{m}$ in A, and the local ring attached to v is $A_\mathfrak{m}$. An algebraic variety is said to be *regular* if all the local rings $A_\mathfrak{m}$ are regular ("regular" is a weaker condition than "nonsingular"; nonsingular implies regular, and the two are equivalent when the ground field k_0 is algebraically closed).

Consider an algebraic curve X. Then X is regular if and only if the local rings attached to it are discrete valuation rings. For example, $\operatorname{Specm} A$ is a regular curve if and only if A is a Dedekind domain. A regular curve X defines a set of discrete valuation rings in $k_0(X)$, each of which contains k_0, and X is complete if and only if this set includes all the discrete valuation rings in $k_0(X)$ having $k_0(X)$ as field of fractions and containing k_0.

A field K containing k_0 is said to be a *function field in n variables* over k_0 if it is finitely generated and has transcendence degree n over k_0. The *field of constants* of K is the algebraic closure of k_0 in K. Thus the function field of an algebraic variety over k_0 of dimension n is a function field in n variables over k_0 having k_0 as its field of constants (whence the terminology).

THEOREM 7.3. *The map $X \mapsto k_0(X)$ defines an equivalence from the category of complete regular algebraic curves over k_0 to the category of function fields in one variable over k_0 having k_0 as field of constants.*

PROOF. The curve corresponding to the field K can be constructed as follows: take X to be the set of discrete valuation rings in K containing k_0 and having K as their field of fractions; define a subset U of X to be open if it omits only finitely many elements of X; for such a U, define $\mathcal{O}_X(U)$ to be the intersection of the discrete valuation rings in U. □

COROLLARY 7.4. *A regular curve U can be embedded into a complete regular curve $\bar{U}$; the map $U \hookrightarrow \bar{U}$ is universal among maps from U into complete regular curves.*

PROOF. Take $\bar{U}$ to be the complete regular algebraic curve attached to $k_0(U)$. There is an obvious identification of U with an open subset of $\bar{U}$. □

EXAMPLE 7.5. Let $F(X,Y)$ be an absolutely irreducible polynomial in $k_0[X,Y]$, and let $A = k_0[X,Y]/(F(X,Y))$. Thus A is an affine k_0-algebra, and $C \overset{\text{df}}{=} \operatorname{Specm} A$ is the curve: $F(X,Y) = 0$. Let C^{ns} be the complement in C of the set of maximal ideals of A containing the ideal $(\partial F/\partial X, \partial F/\partial X) \mod F(X,Y)$. Then C^{ns} is a nonsingular curve, and hence can be embedded into a complete regular curve $\bar{C}$.

There is a geometric way of constructing $\bar{C}$, at least in the case that $k_0 = k$ is algebraically closed. First consider the plane projective curve C' defined by the homogeneous equation

$$Z^{\deg(F)} F(X/Z, Y/Z) = 0.$$

This is a projective (hence complete) algebraic curve which, in general, will have singular points. It is possible to resolve these singularities geometrically, and so obtain a nonsingular projective curve (see W. Fulton 1969, p179).

THEOREM 7.6. *Every compact Riemann surface X has a unique structure of a complete nonsingular algebraic curve.*

PROOF. We explain only how to construct the associated algebraic curve. The underlying set is the same; the topology is that for which the open sets are those with finite complements; the regular functions on an open set U are the holomorphic functions on U which are meromorphic on the whole of X. $\square$

REMARK 7.7. Theorems 7.3 and 7.6 depend crucially on the hypothesis that the variety has dimension 1.

In general, many different complete nonsingular algebraic varieties can have the same function field. A nonsingular variety U over a field of characteristic zero can be embedded in a complete nonsingular variety $\bar{U}$, but this is a very difficult theorem (proved by Hironaka in 1970), and $\bar{U}$ is very definately not unique. For a variety of dimension > 3 over a field of characteristic $p > 0$, even the existence of $\bar{U}$ is not known.

For a curve, "complete" is equivalent to "projective"; for smooth surfaces they are also equivalent, but in higher dimensions there are many complete nonprojective varieties (although Chow's lemma says that a complete variety is not too far away from a projective variety).

Many compact complex manifolds of dimension > 1 have no algebraic structure.

The curve $X_0(N)$ over $\mathbb{Q}$. According to Theorem 7.6, there is a unique structure of an complete nonsingular curve on $X_0(N)$ compatible with its structure as a Riemann surface. We write $X_0(N)_{\mathbb{C}}$ for $X_0(N)$ regarded as an algebraic curve over $\mathbb{C}$. Note that $X_0(N)_{\mathbb{C}}$ is the unique complete nonsingular curve over $\mathbb{C}$ having the field $\mathbb{C}(j(z), j(Nz)$ of modular functions for $\Gamma_0(N)$ as its field of rational functions.

Now write $F_N(X, Y)$ for the polynomial constructed in Theorem 6.1, and let C be the curve over $\mathbb{Q}$ defined by the equation:

$$F_N(X, Y) = 0.$$

As is explained above, we can remove the singular points of C to obtain a nonsingular curve C^{ns} over $\mathbb{Q}$, and then we can embed C^{ns} into a complete regular curve $\bar{C}$. The coordinate functions x and y are rational functions on $\bar{C}$, they generate the field of rational functions on $\bar{C}$, and they satisfy the relation $F_N(x, y) = 0$; these statements characterize $\bar{C}$ and the pair of functions x, y on it.

Let $\bar{C}_{\mathbb{C}}$ be the curve defined by $\bar{C}$ over $\mathbb{C}$. It can also be obtained in the same way as $\bar{C}$ starting from the curve $F_N(X, Y) = 0$, now thought of as a curve over $\mathbb{C}$. There is a unique isomorphism $\bar{C}_{\mathbb{C}} \to X_0(N)_{\mathbb{C}}$ making the rational functions x and y on $\bar{C}_{\mathbb{C}}$ correspond to the functions $j(z)$ and $j(Nz)$ on $X_0(N)$. We can use this isomorphism to identify the two curves, and so we can regard $\bar{C}$ as being a model of $X_0(N)_{\mathbb{C}}$ over $\mathbb{Q}$. We write it $X_0(N)_{\mathbb{Q}}$. (In fact, we often omit the subscripts from $X_0(N)_{\mathbb{C}}$ and $X_0(N)_{\mathbb{Q}}$.)

We can be a little more explicit: on an open subset, the isomorphism $X_0(N) \to \bar{C}_{\mathbb{C}}$ is simply the map $[z] \mapsto (j(z), j(Nz))$ (regarding this pair as a point on the affine curve $F_N(X, Y) = 0$).

The action of $\mathrm{Aut}(\mathbb{C})$ on $X_0(N)$ corresponding to the model $X_0(N)_{\mathbb{Q}}$ has the following description: for $\tau \in \mathrm{Aut}(\mathbb{C})$, $\tau[z] = [z']$ if $\tau j(z) = j(z')$ and $\tau j(Nz) = j(Nz')$.

The curve $X_0(N)_{\mathbb{Q}}$ is called the *canonical model* of $X_0(N)$ over $\mathbb{Q}$. The canonical model $X(1)_{\mathbb{Q}}$ of $X(1)$ is just the projective line $\mathbb{P}^1$ over $\mathbb{Q}$. If the field of rational functions on $\mathbb{P}^1$ is $\mathbb{Q}(T)$, then the identification of $\mathbb{P}^1$ with X(1) is made in such a way that T corresponds to j.

The quotient map $X_0(N) \to X(1)$ corresponds to the map of algebraic curves $X_0(N)_{\mathbb{Q}} \to X(1)_{\mathbb{Q}}$ defined by the inclusion of function fields $\mathbb{Q}(T) \to \mathbb{Q}(x, y)$, $T \mapsto x$. On an open subset of $X_0(N)_{\mathbb{Q}}$, it is the projection map $(a, b) \mapsto a$.

8. Modular Curves as Moduli Varieties

For over 100 years algebraic geometers have worked with "moduli varietes" that classify isomorphism classes of certain objects but, as far as I know, a precise definition of a moduli variety was not given before Mumford's work in the 1960's. In this section I explain the general notion of a moduli variety, and then I explain how to realize the modular curves as moduli varieties for elliptic curves with additional structure.

The general notion of a moduli variety. Fix a field k which initially we assume to be algebraically closed. A *moduli problem* over k is a contravariant functor $\mathcal{F}$ from the category of algebraic varieties over k to the category of sets. Thus for each variety V over k we are given a set $\mathcal{F}(V)$, and for each regular map $\varphi : W \to V$, we are given map $\varphi^* : \mathcal{F}(V) \to \mathcal{F}(W)$. Typically, $\mathcal{F}(V)$ will be the set of isomorphism classes of certain objects over V.

A solution to the moduli problem is a variety V over k together with an identification $V(k) = \mathcal{F}(k)$ and certain additional data sufficient to determine V uniquely. More precisely:

Definition 8.1. A pair (V, α) consisting of a variety V over k together with a bijection $\alpha : \mathcal{F}(k) \to V(k)$ is called a *solution to the moduli problem* $\mathcal{F}$ if it satisfies the following conditions:

(a) Let T be a variety over k and let $f \in \mathcal{F}(T)$; a point $t \in T(k)$ can be regarded as a map $\operatorname{Specm} k \to V$, and so (by the functoriality of $\mathcal{F}$) f defines an element f_t of $T(k)$; we therefore have a map $t \mapsto \alpha(f_t) : T(k) \to V(k)$, and this map is required to be regular (i.e., defined by a morphism of algebraic varieties);

(b) *(Universality)* Let Z be a variety over k and let $\beta : \mathcal{F}(k) \to Z(k)$ be a map such that, for any pair (T, f) as in (a), the map $t \mapsto \beta(f_t) : T(k) \to Z(k)$ is regular; then the map $\beta \circ \alpha^{-1} : V(k) \to Z(k)$ is regular.

A variety V that occurs as the solution of a moduli problem is called a *moduli variety.*

Proposition 8.2. *Up to a unique isomorphism, there exists at most one solution to a moduli problem.*

Proof. Suppose there are two solutions (V, α) and (V', α'). Then because of the universality of (V, α), $\alpha' \circ \alpha^{-1} : V \to V'$ is a regular map, and because of the universality of (V', α'), its inverse is also a regular map. □

Of course, in general there may exist no solution to a moduli problem, and when there does exist a solution, it may be very difficult to prove it. Mumford was given the Fields medal mainly because of his construction of the moduli varieties of curves and abelian varieties.

Remark 8.3. It is possible to modify the above definition for the case that the ground field k_0 is not algebraically closed. For simplicity, we assume k_0 to be perfect, and we let k be an algebraic closure of k_0. Now V is a variety over k_0 and α is a family of maps $\alpha(k') : \mathcal{F}(k') \to V(k')$ (one for each algebraic extension k' of k_0) compatible with inclusions of fields, and $(V_k, \alpha(k))$ is required to be a solution to the moduli problem over k. If (V, α) and (V', α') are two solutions to the same moduli problem, then $\alpha' \circ \alpha^{-1} : V(k) \to V'(k)$ and its inverse are both regular maps commuting with the action of $\mathrm{Gal}(k/k_0)$; they are both therefore defined over k_0. Consequently, up to a unique isomorphism, there again can be at most one solution to a moduli problem.

Note that we don't require $\alpha(k')$ to be a bijection when k' is not algebraically closed. In particular, V need not represent the functor $\mathcal{F}$. When V does represent the functor, V is called a *fine* moduli variety; otherwise it is a *coarse* moduli variety.

The moduli variety for elliptic curves. We show that $\mathbb{A}^1$ is the moduli variety for elliptic curves over a perfect field k_0.

An *elliptic curve* E over a field k' is a curve given by an equation of the form,

$$Y^2Z + a_1XYZ + a_3YZ^2 = X^3 + a_2X^2Z + a_4XZ^2 + a_6Z^3 \qquad (*)$$

for which the discriminant $\Delta(a_1, a_2, a_3, a_4, a_6) \neq 0$. It has a distinguished point $(0:1:0)$, and an isomorphism of elliptic curves over k' is an isomorphism of varieties carrying the distinguished point on one curve to the distinguished point on the second. (There is a unique group law on E having the distinguished element as zero, and a morphism of elliptic curves is automatically a homomorphism of groups.)

Let V be a variety over a field k'. An *elliptic curve* (better, *family of elliptic curves*) over V is a map of algebraic varieties $E \to V$ where E is the subvariety of $V \times \mathbb{P}^1$ defined by an equation of the form $(*)$ with the a_i regular functions on V; $\Delta(a_1, a_2, a_3, a_4, a_6)$ is now a regular function on V which is required to have no zeros.

For any variety V, let $\mathcal{E}(V)$ be the set of isomorphism classes of elliptic curves over V. Then $\mathcal{E}$ is a contravariant functor, and so can be regarded as a moduli problem over k_0.

For any field k' containing k_0, the j-invariant defines a map

$$E \mapsto j(E) : \mathcal{E}(k') \to \mathbb{A}^1(k') = k',$$

and the theory of elliptic curves (Math 679) shows that this map is an isomorphism if k' is algebraically closed (but not in general otherwise).

THEOREM 8.4. *The pair $(\mathbb{A}^1, j)$ is a solution to the moduli problem $\mathcal{E}$.*

PROOF. For any k_0-homomorphism $\sigma : k' \to k''$, $j(\sigma E) = \sigma j(E)$, and so it remains to show that $(\mathbb{A}^1, j)$ satisfies the conditions (a) and (b) over k.

Let $E \to T$ be a family of elliptic curves over T, where T is a variety over k. The map $t \mapsto j(E_t) : T(k) \to \mathbb{A}^1(k)$ is regular because $j(E_t) = c_4^3/\Delta$ where c_4 is a polynomial in the a_i's and Δ is a nowhere zero polynomial in the a_i's.

Now let (Z, β) be a pair as in (b). We have to show that $j \mapsto \beta(E_j) : \mathbb{A}^1(k) \to Z(k)$, where E_j is an elliptic curve over k with j-invariant j, is regular. Let U be the open subset of $\mathbb{A}^1$ obtained by removing the points 0 and 1728. Then

$$E : Y^2Z + XYZ = X^3 - \frac{36}{u-1728}XZ^2 - \frac{1}{u-1728}Z^3, \quad u \in U,$$

is an elliptic curve over U with the property that $j(E_u) = u$ (Silverman 1986, p52). Because of the property possessed by (Z, β), E/U defines a regular map $u \mapsto \beta(E_u) : U \to Z$. But this is just the restriction of the map $j \mapsto \beta(E_j)$ to $U(k)$, which is therefore regular, and it follows that j itself is regular. $\square$

The curve $Y_0(N)_{\mathbb{Q}}$ as a moduli variety. Let k be a perfect field, and let N be a positive integer not divisible by the characteristic of k (so there is no restriction on N when k has characteristic zero). Let E be an elliptic curve over k. When k is an algebraically closed field, a *cyclic subgroup of E of order N* is simply a cyclic subgroup of $E(k)$ of order N in the sense of abstract groups. When k is not algebraically closed, a *cyclic subgroup* of E is a Zariski-closed subset S such that $S(k^{\mathrm{al}})$ is cyclic subgroup of $S(k^{\mathrm{al}})$ of order N. Thus $S(k^{\mathrm{al}})$ is a cyclic subgroup of order N of $E(k^{\mathrm{al}})$ that is stable (as a set—not elementwise) under the action of $\mathrm{Gal}(k^{\mathrm{al}}/k)$, and every such group arises from a (unique) S.

An *isomorphism* from one pair (E,S) to a second (E',S') is an isomorphism $E\to E'$ mapping S onto S'.

These definitions can be extended in a natural way to families of elliptic curves over varieties.

For any variety V over k, define $\mathcal{E}_{0,N}(V)$ to be the set of isomorphism classes of pairs (E,S) where E is an elliptic curve over V, and S is a cyclic subgroup of E of order N. Then $\mathcal{E}_{0,N}$ is a contravariant functor, and hence is a moduli problem.

Recall that $\Lambda(\omega_1,\omega_2)$ is the lattice generated by a pair (ω_1,ω_2) with $\Im(\omega_1/\omega_2)>0$. Note that $\Lambda(\omega_1,N^{-1}\omega_2)/\Lambda(\omega_1,\omega_2)$ is a cyclic subgroup of order N of the elliptic curve $\mathbb{C}/\Lambda(\omega_1,\omega_2)$.

LEMMA 8.5. *The map*

$$\mathbb{H}\to\mathcal{E}_{0,N}(\mathbb{C}),\quad z\mapsto(\mathbb{C}/\Lambda(z,1),\Lambda(z,N^{-1})/\Lambda(z,1))$$

induces a bijection $\Gamma_0(N)\backslash\mathbb{H}\to\mathcal{E}_{0,N}(\mathbb{C})$.

PROOF. Easy—see Math 679, p124. □

Let $\mathcal{E}'_{0,N}(\mathrm{k})$ denote the set of isomorphism classes of homomorphisms of elliptic curves $\alpha: E\to E'$ over k whose kernel is a cyclic subgroup of E of order N. The map

$$\alpha\mapsto(E,\mathrm{Ker}(\alpha)):\mathcal{E}'_{0,N}(k)\to\mathcal{E}_{0,N}(k)$$

is a bijection; its inverse is $(E,S)\mapsto(E\to E/S)$. For example, the element $(\mathbb{C}/\Lambda(z,1),\Lambda(z,N^{-1})/\Lambda(z,1))$ of $\mathcal{E}_{0,N}(\mathbb{C})$ corresponds to the element $(\mathbb{C}/\Lambda(z,1)\xrightarrow{N}\mathbb{C}/\Lambda(Nz,1))$ of $\mathcal{E}'_{0,N}(\mathbb{C})$.

Let $F_N(X,Y)$ be the polynomial defined in Theorem 6.1 and let C be the (singular) curve $F_N(X,Y)=0$ over $\mathbb{Q}$. For any field $k\supset\mathbb{Q}$, consider the map

$$\mathcal{E}'_{0,N}(k)\to\mathbb{A}^2(k),\quad(E,E')\mapsto(j(E),j(E')).$$

When $k=\mathbb{C}$, the above discussion shows that the image of this map is contained in $C(\mathbb{C})$, and this implies that the same is true for any k.

Recall that $Y_0(N)=\Gamma_0(N)\backslash\mathbb{H}$. There is an affine curve $Y_0(N)_{\mathbb{Q}}\subset X_0(N)_{\mathbb{Q}}$ which is a model of $Y_0(N)\subset X_0(N)$. (This just says that the set of cusps on $X_0(N)$ is defined over $\mathbb{Q}$.)

THEOREM 8.6. *Let k be a field, and let N be an integer not divisible by the characteristic of k. The moduli problem $\mathcal{E}_{0,N}$ has a solution (M,α) over k. When*

$k = \mathbb{Q}$, M is canonically isomorphic to $Y_0(N)_{\mathbb{Q}}$, and the map

$$\mathcal{E}_{0,N}(k) \xrightarrow{\alpha} M(k) \xrightarrow{\approx} Y_0(N)_{\mathbb{Q}}(k) \xrightarrow{(j,j_N)} C(k)$$

is $(E, S) \mapsto (j(E), j(E/S))$.

PROOF. When $k = \mathbb{Q}$, it is possible to prove that $Y_0(N)_{\mathbb{Q}}$ is a solution to the moduli problem in much the same way as for $\mathbb{A}^1$ above. If $p \nmid N$, then it is possible to show that $Y_0(N)_{\mathbb{Q}}$ has good reduction at p, and the curve $Y_0(N)_{\mathbb{F}_p}$ over $\mathbb{F}_p$ it reduces to is a solution to the moduli problem over $\mathbb{F}_p$. □

The curve $Y(N)$ as a moduli variety. Let N be a positive integer, and let $\zeta \in \mathbb{C}$ be a primitive Nth root of 1. A *level-N structure* on an elliptic curve E is a pair of points $t = (t_1, t_2)$ in $E(k)$ such that the map

$$(m, m') \mapsto (mt, mt') : \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z} \to E(k)$$

is injective. This means that $E(k)_N$ has order N^2, and t_1 and t_2 form a basis for $E(k)_N$ as a $\mathbb{Z}/n\mathbb{Z}$-module. For any variety V over a field $k \supset \mathbb{Q}[\zeta]$, define $\mathcal{E}_N(V)$ to be the set of isomorphism classes of pairs (E, t) where E is an elliptic curve over V and $t = (t_1, t_2)$ is a level-N structure on E such that $e_N(t_1, t_2) = \zeta$ (here e_N is the Weil pairing—see Silverman III.8). Then $\mathcal{E}_N$ is a contravariant functor, and hence is a moduli problem.

LEMMA 8.7. *The map*

$$\mathbb{H} \to \mathcal{E}_N(\mathbb{C})$$

$z \mapsto (\mathbb{C}//\Lambda(z,1), (z,1) \mod \Lambda(z,1))$ *induces a bijection $\Gamma(N)\backslash\mathbb{H} \to \mathcal{E}_N(\mathbb{C})$.*

PROOF. Easy. □

THEOREM 8.8. *Let k be a field containing $\mathbb{Q}[\zeta]$, where ζ is a primitive Nth root of 1. The moduli problem $\mathcal{E}_N$ has a solution (M, α) over k. When $k = \mathbb{C}$, M is canonically isomorphic to $Y(N)_{\mathbb{C}}$ ($= X(N)_{\mathbb{C}}$ with the cusps removed). Let M be the solution to the moduli problem $\mathcal{E}_N$ over $\mathbb{Q}[\zeta]$; then M has good reduction at the prime ideals not dividing N.*

PROOF. Omit. □

EXAMPLE 8.9. For $N = 2$, the solution to the moduli problem is $\mathbb{A}^1$. In this case, there is a universal elliptic curve with level-2 structure over $\mathbb{A}^1$, namely, the curve

$$E : Y^2Z = X(X - Z)(X - \lambda Z).$$

Here λ is the coordinate on $\mathbb{A}^1$, and the map $E \to \mathbb{A}^1$ is $(x : y : z, \lambda) \mapsto \lambda$. The level-2 structure is the pair of points $(0 : 0 : 1)$, $(1 : 0 : 1)$. The curve E is universal in the following sense: for any family of elliptic curves $E' \to V$ with level-2 structure over a variety V (with the same base field k), there is a unique morphism $V \to \mathbb{A}^1$ such that E' is the pull-back of E. In this case the map $\mathcal{E}(k) \to \mathbb{A}^1(k)$ is an isomorphism for all fields $k \supset \mathbb{Q}$, and $\mathbb{A}^1$ is a *fine* moduli variety.

9. Modular Forms, Dirichlet Series, and Functional Equations

The most famous Dirichlet series, $\zeta(s) \overset{\text{df}}{=} \sum_{n=1}^{\infty} n^{-s}$, was shown by Riemann (in 1859) to have an analytic continuation to the whole complex plane except for a simple pole at $s = 1$, and to satisfy a functional equation

$$Z(s) = Z(1-s)$$

where $Z(s) = \pi^{-s/2}\Gamma(s/2)\zeta(s)$. One now believes (Hasse-Weil conjecture) that all Dirichlet series arising as the zeta functions of algebraic varieties over number fields should have meromorphic continuations to the whole complex plane and satisfy functional equations. In this section we investigate the relation between Dirichlet series with functional equations and modular forms.

We saw in (2.12) that the modular group $\Gamma(1)$ is generated by the matrices $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$. Therefore a modular function $f(z)$ of weight $2k$ satisfies the following two conditions:

$$f(z+1) = f(z), \quad f(-1/z) = (-z)^{2k} f(z).$$

The first condition implies that $f(z)$ has a Fourier expansion $f(z) = \sum a_n q^n$, and so defines a Dirichlet series $\varphi(s) = \sum a_n n^{-s}$. Hecke showed that the second condition implies that the Dirichlet series satisfies a functional equation, and conversely every Dirichlet series satisfying a functional equation of the correct form (and certain holomorphicity conditions) arises from a modular form. Weil extended this result to the subgroup $\Gamma_0(N)$ of $\Gamma(1)$, which needs more than two generators (and so we need more than one functional equation for the Dirichlet series). In this section we explain Hecke's and Weil's results, and in later sections we explain the implications of Weil's results for elliptic curves over $\mathbb{Q}$.

The Mellin transform. Let $a_1, a_2, \ldots$ be a sequence of complex numbers such that $a_n = O(n^M)$ for some M. This can be regarded as the sequence of coefficients of either the power series $f(q) = \sum_1^{\infty} a_n q^n$, which is absolutely convergent for $|q| < 1$ at least, or for the Dirichlet series $\varphi(s) = \sum_1^{\infty} a_n n^{-s}$, which is absolutely convergent for $\Re(s) > M + 1$ at least. In this subsection, we give explicit formulae that realize the formal correspondence between $f(y)$ and $\varphi(s)$.

Recall that the gamma function $\Gamma(s)$ is defined by the formula,

$$\Gamma(s) = \int_0^{\infty} e^{-x} x^{s-1} dx, \quad \Re(s) > 0.$$

It has the following properties: $\Gamma(s+1) = s\Gamma(s)$, $\Gamma(1) = 1$, and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$; $\Gamma(s)$ extends to a function that is holomorphic on the whole complex plane, except for simple poles at $s = -n$, where it has a residue $\frac{(-1)^n}{n!}$, $n = 0, 1, 2, \ldots$.

Proposition 9.1 (Mellin Inversion Formula). *For every real $c > 0$,*

$$e^{-x} = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \Gamma(s) x^{-s} ds, \quad x > 0.$$

(The integral is taken upwards on a vertical line.)

PROOF. Regard the integral as taking place on a vertical circumference on the Riemann sphere. The calculus of residues shows that the integral is equal to

$$2\pi i\sum_{n=0}^{\infty}\mathrm{res}_{s=-n}x^{-s}\Gamma(s)=2\pi i\sum_{n=0}^{\infty}\frac{(-x)^n}{n!}=2\pi i\cdot e^{-x}.$$

□

THEOREM 9.2. *Let $a_1, a_2, \ldots$ be a sequence of complex numbers such that $a_n = O(n^M)$ for some M. Write $f(x)=\sum_1^\infty a_n e^{-nx}$ and $\phi(s)=\sum_1^\infty a_n n^{-s}$. Then*

$$\Gamma(s)\phi(s)=\int_0^\infty f(x)x^{s-1}dx \quad \text{for } \Re(s)>\max(0,M+1), \qquad (*)$$

$$f(x)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\phi(s)\Gamma(s)x^{-s}ds \text{ for } c>\max(0,M+1) \text{ and } \Re(x)>0. \qquad (**)$$

PROOF. First consider $(*)$. Formally we have

$$\begin{aligned}\int_0^\infty f(x)x^{s-1}dx &= \int_0^\infty\sum_1^\infty a_n e^{-nx}x^{s-1}dx\\ &= \sum_1^\infty\int_0^\infty a_n e^{-nx}x^{s-1}dx\\ &= \sum_1^\infty a_n\Gamma(s)n^{-s}\\ &= \Gamma(s)\phi(s)\end{aligned}$$

on writing x for nx in the last integral and using the definition of $\Gamma(s)$. The only problem is in justifying the interchange of the integral with the summation sign.

The equation $(**)$ follows from Proposition 9.1. □

The functions $f(x)$ and $\phi(s)$ are called the *Mellin transforms* of each other.

The equation $(*)$ provides a means of analytically continuing $\phi(s)$ provided $f(x)$ tends to zero sufficiently rapidly at $x=0$. In particular, if $f(x)=O(x^A)$ for every $A>0$ as $x\to 0$ through real positive values, then $\Gamma(s)\phi(s)$ can be extended to a holomorphic function over the entire complex plane. Of course, this condition on $f(x)$ implies that $x=0$ is an essential singularity.

We say that a function $\varphi(s)$ on the complex plane is *bounded on vertical strips,* if for all real numbers $a<b$, $\varphi(s)$ is bounded on the strip $a\le\Re(s)\le b$ as $\Im(s)\to\pm\infty$.

THEOREM 9.3 (Hecke 1936). *Let $a_0, a_1, a_2, \ldots$ be a sequence of complex numbers such that $a_n = O(n^M)$ for some M. Given $\lambda>0$, $k>0$, $C=\pm1$, write*

$\varphi(s)=\sum a_n n^{-s}$; $\quad$ *($\varphi(s)$ converges for $\Re(s)>M+1$)*
$\Phi(s)=\left(\frac{2\pi}{\lambda}\right)^{-s}\Gamma(s)\varphi(s)$;
$f(z)=\sum_{n\ge0}a_n e^{2\pi i nz/\lambda}$; $\quad$ (converges for $\Im(z)>0$).

Then the following conditions are equivalent:

(i) *The function* $\Phi(s) + \frac{a_0}{s} + \frac{Ca_0}{k-s}$ *can be analytically continued to a holomorphic function on the entire complex plane which is bounded on vertical strips, and it satisfies the functional equation*

$$\Phi(k-s) = C\Phi(s).$$

(ii) *In the upper half plane, f satisfies the functional equation*

$$f(-1/z) = C(z/i)^k f(z).$$

PROOF. Given (ii), apply ($*$) to obtain (i); given (i), apply ($**$) to obtain (ii). □

REMARK 9.4. Let $\Gamma'(\lambda)$ be the subgroup of $\Gamma(1)$ generated by the maps $z \mapsto z+\lambda$ and $z \mapsto -1/z$. A *modular form of weight k and multiplier C* for $\Gamma'(\lambda)$ is a holomorphic function $f(z)$ on $\mathbb{H}$ such that

$$f(z+\lambda) = f(z), \quad f(-1/z) = C(z/i)^k f(z),$$

and f is holomorphic at $i\infty$. This is a slightly more general notion than in Section 4—if k is an even integer and $C = 1$ then it agrees with it.

The theorem says that there is a one-to-one correspondence between modular forms of weight k and multiplier C for $\Gamma'(\lambda)$ whose Fourier coefficients satisfy $a_n = O(n^M)$ for some M, and Dirichlet series satisfying (i). Note that $\Phi(s)$ is holomorphic if f is a cusp form.

For example $\zeta(s)$ corresponds to a modular form of weight 1/2 and multiplier 1 for $\Gamma'(2)$.

Weil's theorem. Given a sequence of complex numbers $a_1, a_2, \ldots$ such that $a_n = O(n^M)$ for some M, write

$$(1) \qquad L(s) = \sum_{n=1}^{\infty} a_n n^{-s}, \quad \Lambda(s) = (2\pi)^{-s}\Gamma(s)L(s), \quad f(z) = \sum_{n=1}^{\infty} a_n e^{2\pi i n z}.$$

More generally, let $m > 0$ be an integer, and let χ be a primitive character on $(\mathbb{Z}/m\mathbb{Z})^\times$ (primitive means that it is not a character on $(\mathbb{Z}/d\mathbb{Z})^\times$ for any proper divisor d of n). As usual, we extend $\chi(s)$ to the whole of $\mathbb{Z}/m\mathbb{Z}$ by setting $\chi(n) = 0$ if n is not relatively prime to m. We write

$$L_\chi(s) = \sum_{n=1}^{\infty} a_n \chi(n) n^{-s}, \quad \Lambda_\chi(s) = (\frac{m}{2\pi})^{-s}\Gamma(s)L_\chi(s), \quad f_\chi(z) = \sum_{n=1}^{\infty} a_n \chi(n) e^{2\pi i n z}.$$

Note that L_χ and f_χ are the Mellin transforms of each other.

For any χ, the associated Gauss sum is

$$g(\chi) = \sum_{n=1}^{m} \chi(n) e^{-2\pi i n/m}.$$

Obviously $\bar{\chi}(a)g(\chi) = \sum \chi(n)e^{-2\pi i a n/m}$, and hence

$$\chi(n) = m^{-1} g(\chi) \sum \bar{\chi}(a) e^{2\pi i a n/m}.$$

It follows from this last equation that

$$f_\chi = m^{-1} g(\chi) \sum_{1}^{m} \bar{\chi}(a) f|_k \begin{pmatrix} m & a \\ 0 & m \end{pmatrix}.$$

THEOREM 9.5. *Let $f(z)$ be a modular form of weight $2k$ for $\Gamma_0(N)$, and suppose that $f|_k\begin{pmatrix} 0 & -1 \\ N & 0 \end{pmatrix} = C(-1)^k f$ for some $C = \pm 1$. Define*

$$C_\chi = Cg(\chi)\chi(-N)/g(\bar{\chi}).$$

Then $\Lambda_\chi(s)$ satisfies the functional equation:

$$\Lambda_\chi(s) = C_\chi N^{k-s}\Lambda_{\bar{\chi}}(2k-s) \text{ whenever } \gcd(m, N) = 1.$$

PROOF. Apply Theorem 9.3. □

The most interesting result is the converse to this theorem.

THEOREM 9.6 (Weil 1967). *Fix a $C = \pm 1$, and suppose that for all but finitely many primes p not dividing N the following condition holds: for every primitive character χ of $(\mathbb{Z}/p\mathbb{Z})^\times$, $\Lambda(s)$ and $\Lambda_\chi(s)$ can be analytically continued to holomorphic functions in the entire complex plane and that each of them is bounded on vertical strips; suppose also that they satisfy the functional equations:*

$$\Lambda(s) = CN^{k-s}\Lambda(2k-s)$$

$$\Lambda_\chi(s) = C_\chi N^{k-s}\Lambda_{\bar{\chi}}(2k-s)$$

where C_χ is defined above; suppose further that the Dirichlet series $L(s)$ is absolutely convergent for $s = k - \epsilon$ for some $\epsilon > 0$. Then $f(z)$ is a cusp form of weight $2k$ for $\Gamma_0(N)$.

PROOF. Several pages of manipulation of 2×2 matrices. □

Let E be an elliptic curve over $\mathbb{Q}$, and let $L(E, s)$ be the associated L-series. As we shall see shortly, it is generally conjectured that $L(s)$ satisfies the hypotheses of the theorem, and hence is attached to a modular form $f(z)$ of weight 2 for $\Gamma_0(N)$. Granted this, one can show that there is nonconstant map $\alpha : X_0(N) \to E$ (defined over $\mathbb{Q}$) such that the pull-back of the canonical differential on E is the differential on $X_0(N)$ attached to $f(z)$.

REMARK 9.7. Complete proofs of the statements in this section can be found in Ogg 1969, especially Chapter V. They are not particularly difficult—it would only add about 5 pages to the notes to include them.

10. Correspondences on Curves; the Theorem of Eichler-Shimura

In this section we sketch a proof of the key theorem of Eichler and Shimura relating the Hecke correspondence T_p to the Frobenius map. In the next section we explain how this enables us to realize certain zeta functions as the Mellin transforms of modular forms.

The ring of correspondences of a curve. Let X and X' be projective nonsingular curves over a field k which, for simplicity, we take to be algebraically closed.

A *correspondence* T between X and X' is a pair of finite surjective morphisms

$$X \xleftarrow{\alpha} Y \xrightarrow{\beta} X'.$$

It can be thought of as a many-valued map $X \to X'$ sending a point $P \in X(k)$ to the set $\{\beta(Q_i)\}$ where the Q_i run through the elements of $\alpha^{-1}(P)$ (the Q_i need not be distinct). Better, define $\mathrm{Div}(X)$ to be the free abelian group on the set of points of X; thus an element of $\mathrm{Div}(X)$ is a finite formal sum

$$D = \sum n_P P, \quad n_P \in \mathbb{Z}, \quad P \in C.$$

A correspondence T then defines a map

$$\mathrm{Div}(X) \to \mathrm{Div}(X'), \quad P \mapsto \sum \beta(Q_i),$$

(notations as above). This map multiplies the degree of a divisor by $\deg(\alpha)$. It therefore sends the divisors of degree zero on X into the divisors of degree zero on X', and one can show that it sends principal divisors to principal divisors. It therefore defines a map $T : J(X) \to J(X')$ where

$$J(X) \overset{\mathrm{df}}{=} \mathrm{Div}^0(X)/\{\text{principal divisors}\}.$$

We define the *ring of correspondences* $\mathcal{A}(X)$ on X to be the subring of $\mathrm{End}(J(X))$ generated by the maps defined by correspondences.

If T is the correspondence

$$X \xleftarrow{\beta} Y \xrightarrow{\alpha} X'.$$

then the transpose T' of T is the correspondence

$$X \xleftarrow{\alpha} Y \xrightarrow{\beta} X'.$$

A morphism $\alpha : X \to X'$ can be thought of as a correspondence

$$X \leftarrow \Gamma \to X'$$

where $\Gamma \subset X \times X'$ is the graph of α and the maps are the projections.

Aside 10.1. Attached to any complete nonsingular curve X there is an abelian variety $\mathrm{Jac}(X)$ whose set of points is $J(X)$. The ring of correspondences is the endomorphism ring of $\mathrm{Jac}(X)$—see the next section.

The Hecke correspondence. Let Γ be a subgroup of $\Gamma(1)$ of finite index, and let α be a matrix with integer coefficients and determinant > 0. Write $\Gamma\alpha\Gamma = \cup\Gamma\alpha_i$ (disjoint union). Then we get a map

$$T(\alpha) : J(X(\Gamma)) \to J(X(\Gamma)), \quad [z] \mapsto \sum [\alpha_i z].$$

As was explained on in Section 5, this is the map defined by the correspondence:

$$X(\Gamma) \leftarrow X(\Gamma_\alpha) \xrightarrow{\alpha} X(\Gamma)$$

where $\Gamma_\alpha = \Gamma \cap \alpha^{-1}\Gamma\alpha$. In this way, we get a homomorphism $\mathcal{H} \to \mathcal{A}$ from the ring of Hecke operators into the ring of correspondences.

Consider the case $\Gamma = \Gamma_0(N)$ and $T = T(p)$ the Hecke correspondence defined by the double coset $\Gamma_0(N)\begin{pmatrix} 1 & 0 \\ 0 & p \end{pmatrix}\Gamma_0(N)$. Assume that $p \nmid N$. We give two further descriptions of $T(p)$.

First, identify a point of $Y_0(N)$ (over $\mathbb{C}$) with an isomorphism class of homomorphisms $E \to E'$ of elliptic curves with kernel a cyclic group of order N. The subgroup E_p of E of points of order dividing p is isomorphic to $(\mathbb{Z}/p\mathbb{Z}) \times (\mathbb{Z}/p\mathbb{Z})$. Hence there are $p+1$ cyclic subgroups of E_p of order p, say $S_0, S_1, \ldots, S_p$ (they correspond to the lines through the origin in $\mathbb{F}_p^2$). Then (as a many-valued map), $T(p)$ sends $\alpha : E \to E'$ to $\{E_i \to E_i' \mid i = 0, 1, \ldots, p\}$ where $E_i = E/S_i$ and $E_i' = E'/\alpha(S_i)$.

Second, regard $Y_0(N)$ as the curve C defined by the polynomial $F_N(X, Y)$ constructed in Theorem 6.1 (of course, this isn't quite correct—there is a map $Y_0(N) \to C, \quad [z] \mapsto (j(z), j(Nz))$, which is an isomorphism over the nonsingular part of C). Let (j, j') be a point on C; then there are elliptic curves E and E' (well-defined up to isomorphism) such that $j = j(E)$ and $j' = j(E')$. The condition $F_N(j, j') = 0$ implies that there is a homomorphism $\alpha : E \to E'$ with kernel a cyclic subgroup of order N. Then $T(p)$ maps (j, j') to $\{(j_i, j_i') \mid i = 0, \ldots, p\}$ where $j_i = j(E/S_i)$ and $j_i' = j(E'/\alpha S_i)$.

These last two descriptions of the action of $T(p)$ are valid over any field of characteristic 0.

The Frobenius map. Let C be a curve defined over a field k of characteristic $p \neq 0$. Assume (for simplicity) that k is algebraically closed. If C is defined by equations $\sum c_{i_0 i_1 \ldots} X_0^{i_0} X_1^{i_1} \cdots = 0$ and q is a power of p, then $C^{(q)}$ is the curve defined by the equations $\sum c_{i_0 i_1 \ldots}^q X_0^{i_0} X_1^{i_1} \cdots = 0$, and the *Frobenius map* $\Pi_q : C \to C^{(q)}$ sends the point $(a_0 : a_1 : \cdots)$ to $(a_0^q : a_1^q : \cdots)$. Note that if C is defined over $\mathbb{F}_q$, so that the equations can be chosen to have coefficients $c_{i_0 i_1 \ldots}$ in $\mathbb{F}_q$, then $C = C^{(q)}$ and the Frobenius map is a map from C to itself.

Recall that a nonconstant morphism $\alpha : C \to C'$ of curves defines an inclusion $\alpha^* : k(C') \hookrightarrow k(C)$ of function fields, and that the degree of α is defined to be $[k(C) : \alpha^* k(C')]$. The map α is said to be *separable* or *purely inseparable* according as $k(C)$ is a separable of purely inseparable extension of $\alpha^* k(C')$. If the separable degree of $k(C)$ over $\alpha^* k(C')$ is m, then the map $C(k) \to C'(k)$ is $m : 1$ except on a finite set (assuming k to be algebraically closed).

PROPOSITION 10.2. *The Frobenius map $\Pi_q : C \to C^{(q)}$ is purely inseparable of degree q, and any purely inseparable map $\varphi : C \to C'$ of degree q (of complete* *nonsingular curves) factors as*

$$C \xrightarrow{\Pi_q} C^{(q)} \xrightarrow{\approx} C'$$

PROOF. See Silverman 1986, II.2.12. [First check that

$$\Pi_q^* k(C) = k(C^{(q)}) = k(C)^q \stackrel{\text{df}}{=} \{a^q \mid a \in k(C)\}$$

Then show that $k(C)$ is purely inseparable of degree q over $k(C)^q$, and that this statement uniquely determines $k(C)^q$. The last sentence is obvious when $k(C) = k(T)$ (field of rational functions in T), and the general case follows because $k(C)$ is a separable extension of such a field $k(T)$]. □

Brief review of the points of order p on elliptic curves. Let E be an elliptic curve over an algebraically closed field k. The map $p : E \to E$ (multiplication by p) is of degree p^2. If k has characteristic zero, then the map is separable, which implies that its kernel has order p^2. If k has characteristic p, the map is never separable: either it is purely inseparable (and so E has no points of order p) or its separable and inseparable degrees are p (and so E has p points of order dividing p). In the first case, (10.2) tells us that multiplication by p factors as

$$E \to E^{(p^2)} \xrightarrow{\approx} E.$$

Hence this case occurs only when $E \approx E^{(p^2)}$, i.e., when $j(E) = j(E^{(p^2)}) = j(E)^{p^2}$. Thus if E has no points of order p, then $j(E) \in \mathbb{F}_{p^2}$.

The Eichler-Shimura theorem. The curve $X_0(N)$ is defined over $\mathbb{Q}$ and the Hecke correspondence $T(p)$ is defined over some number field K. For almost all primes $p \nmid N$, $X_0(N)$ will reduce to a nonsingular curve $\tilde{X}_0(N)$.[4] For such a prime p, the correspondence $T(p)$ defines a correspondence $\tilde{T}(p)$ on $\tilde{X}_0(N)$.

THEOREM 10.3. *For a prime p where $X_0(N)$ has good reduction,*

$$\bar{T}_p = \Pi_p + \Pi_p'$$

(equality in the ring $\mathcal{A}(\tilde{X}_0(N))$ of correspondences on $\tilde{X}_0(N)$ over the algebraic closure $\mathbb{F}$ of $\mathbb{F}_p$; here Π_p' is the transpose of Π_p).

PROOF. We show that they agree as many-valued maps on an open subset of $\tilde{X}_0(N)$.

Over $\mathbb{Q}_p^{\text{al}}$ we have the following description of T_p (see above): a homomorphism of elliptic curves $\alpha : E \to E'$ with cyclic kernel of order N defines a point $(j(E), j(E'))$ on $X_0(N)$; let $S_0, \ldots, S_p$ be the subgroups of order p in E; then

$$T_p(j(E), j(E')) = \{(j(E_i), j(E_i'))\}$$

where $E_i = E/S_i$ and $E_i' = E'/\alpha(S_i)$.

Consider a point $\tilde{P}$ on $\tilde{X}_0(N)$ with coordinates in $\mathbb{F}$. Ignoring a finite number of points of $\tilde{X}_0(N)$, we can suppose $\tilde{P} \in \tilde{Y}_0(N)$ and hence is of the form $(j(\tilde{E}), j(\tilde{E}'))$ for some map $\tilde{\alpha} : \tilde{E} \to \tilde{E}'$. Moreover, we can suppose that $\tilde{E}$ has p points of order dividing p.

[4]In fact, it is known that $X_0(N)$ has good reduction for all primes $p \nmid N$, but this is hard to prove. It is easy to see that $X_0(N)$ does not have good reduction at primes dividing N.

Let $\alpha : E \to E'$ be a lifting of $\tilde{\alpha}$ to $\mathbb{Q}_p^{\mathrm{al}}$. The reduction map $E_p(\mathbb{Q}_p^{\mathrm{al}}) \to \tilde{E}_p(\mathbb{F}_p^{\mathrm{al}})$ has a kernel of order p. Number the subgroups of order p in E so that S_0 is the kernel of this map. Then each S_i, $i \neq 0$, maps to a subgroup of order p in $\tilde{E}$.

The map $p : \tilde{E} \to \tilde{E}$ factors as

$$\tilde{E} \xrightarrow{\varphi} \tilde{E}/S_i \xrightarrow{\psi} \tilde{E}.$$

When $i = 0$, φ is a purely inseparable map of degree p (it is the reduction of the map $E \to E/S_0$—it therefore has degree p and has zero (visible) kernel), and so ψ must be separable of degree p (we are assuming $\tilde{E}$ has p points of order dividing p). Proposition 10.2 shows that there is an isomorphism $\tilde{E}^{(p)} \to \tilde{E}/S_0$. Similarly $\tilde{E}'^{(p)} \approx \tilde{E}'/S_0$. Therefore

$$(j(\tilde{E}_0), j(\tilde{E}_0')) = (j(\tilde{E}^{(p)}), j(\tilde{E}'^{(p)})) = (j(\tilde{E})^p, j(\tilde{E}')^p) = \Pi_p(j(\tilde{E}), j(\tilde{E}')).$$

When $i \neq 0$, φ is separable (its kernel is the reduction of S_i), and so ψ is purely inseparable. Therefore $\tilde{E} \approx \tilde{E}_i^{(p)}$, and similarly $\tilde{E}' \approx \tilde{E}_i'^{(p)}$. Therefore

$$(j(\tilde{E}_i)^{(p)}, j(\tilde{E}_i')^{(p)}) = (j(\tilde{E}), j(\tilde{E}')).$$

Hence

$$\{(j(\tilde{E}_i), j(\tilde{E}_i')) \mid i = 1, 2, \ldots, p\}$$

is the inverse image of Π_p, i.e., it is $\Pi_p'(j(\tilde{E}), j(\tilde{E}'))$. This completes the proof of the theorem. □

11. Curves and their Zeta Functions

We begin by reviewing the theory of the zeta functions of curves over $\mathbb{Q}$; then we explain the relation between the various representations of the ring of correspondences; finally we explain the implications of the Eichler-Shimura theorem for the zeta functions of the curves $X_0(N)$ and elliptic curves; in particular, we state the conjecture of Taniyama-Weil, and briefly indicate how it implies Fermat's last theorem.

Two elementary results. We begin with two results from linear algebra that will be needed later.

PROPOSITION 11.1. *Let Λ be a free $\mathbb{Z}$-module of finite rank, and let $\alpha : \Lambda \to \Lambda$ be a $\mathbb{Z}$-linear map with nonzero determinant. Then the kernel of the map*

$$\tilde{\alpha} : (\Lambda \otimes \mathbb{Q})/\Lambda \to (\Lambda \otimes \mathbb{Q})/\Lambda$$

defined by α has order $|\det(\alpha)|$.

PROOF. Consider the commutative diagram:

$$\begin{array}{ccccccccc}
0 & \longrightarrow & \Lambda & \longrightarrow & \Lambda \otimes \mathbb{Q} & \longrightarrow & (\Lambda \otimes \mathbb{Q})/\Lambda & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle \alpha \otimes 1} & & \downarrow{\scriptstyle \tilde{\alpha}} & & \\
0 & \longrightarrow & \Lambda & \longrightarrow & \Lambda \otimes \mathbb{Q} & \longrightarrow & (\Lambda \otimes \mathbb{Q})/\Lambda & \longrightarrow & 0.
\end{array}$$

Because $\det(\alpha) \neq 0$, the middle vertical map is an isomorphism. Therefore the snake lemma gives an isomorphism

$$\operatorname{Ker}(\tilde{\alpha}) \to \operatorname{Coker}(\alpha),$$

and it is easy to see that $\operatorname{Coker}(\alpha)$ is finite with order equal to $\det(\alpha)$ (especially if the map is given by a diagonal matrix). $\square$

Let V be a real vector space. To give the structure of a complex vector space on V (compatible with its real structure), it suffices to give an $\mathbb{R}$-linear map $J : V \to V$ such that $J^2 = -1$.

The map J extends by linearity to $V \otimes_{\mathbb{R}} \mathbb{C}$, and $V \otimes_{\mathbb{R}} \mathbb{C}$ splits as a direct sum

$$V \otimes_{\mathbb{R}} \mathbb{C} = V^+ \oplus V^-,$$

$V^{\pm}$ the ± 1 eigenspaces of J.

PROPOSITION 11.2. (a) *The map*

$$V \xrightarrow{v \mapsto v \otimes 1} V \otimes_{\mathbb{R}} \mathbb{C} \xrightarrow{project} V^+$$

is an isomorphism of complex *vector spaces.*

(b) *Denote by $w \mapsto \bar{w}$ the map $v \otimes z \mapsto v \otimes \bar{z} : V \otimes_{\mathbb{R}} \mathbb{C} \to V \otimes_{\mathbb{R}} \mathbb{C}$; this is an $\mathbb{R}$-linear involution of $V \otimes_{\mathbb{R}} \mathbb{C}$ interchanging V^+ and V^-.*

PROOF. Easy exercise. $\square$

COROLLARY 11.3. *Let α be an endomorphism of V that is $\mathbb{C}$-linear. Write A for the matrix of α regarded as an endomorphism of V, and A_1 for the matrix of α as a $\mathbb{C}$-linear endomorphism of V. Then*

$$A \sim A_1 \oplus \bar{A}_1.$$

(By this I mean that the matrix A is similar to the matrix $\begin{pmatrix} A_1 & 0 \\ 0 & \bar{A}_1 \end{pmatrix}$

PROOF. Follows immediately from the above Proposition. [In the case that V has dimension 2, we can identify V (as a real or complex vector space) with $\mathbb{C}$; for the map "multiplication by $\alpha = a + ib$" the statement becomes,

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix} \sim \begin{pmatrix} a+ib & 0 \\ 0 & a-ib \end{pmatrix},$$

which is obviously true because the two matrices are semisimple and have the same trace and determinant.] □

The zeta function of a curve over a finite field. The next theorem summarizes what is known.

THEOREM 11.4. *Let C be a complete nonsingular curve of genus g over $\mathbb{F}_q$. Let N_n be the number of points of C with coordinates in $\mathbb{F}_{q^n}$. Then there exist algebraic integers $\alpha_1, \alpha_2, \ldots, \alpha_{2g}$ (independent of n) such that*

$$N_n = 1 + q^n - \sum_{i=1}^{2g} \alpha_i^n; \qquad (*)$$

moreover, the numbers q/α_i^{-1} are a permutation of the α_i, and for each i, $|\alpha_i| = q^{1/2}$.

All but the last of these assertions follow in a straightforward way from the Riemann-Roch theorem (see M. Eichler, Introduction to the Theory of Algebraic Numbers and Function, Academic Press, 1966, V.5.1). The last is the famous "Riemann hypothesis" for curves, proved in this case by Weil in the 1940's.

Define $Z(C,t)$ to be the power series with rational coefficients such that

$$\log Z(C,t) = \sum_{n=1}^{\infty} N_n t^n/n.$$

Then (*) is equivalent to the formula

$$Z(C,t) = \frac{(1-\alpha_1 t)\cdots(1-\alpha_n t)}{(1-t)(1-qt)}$$

(because $-\log(1-at) = \sum a^n t^n/n$).

Define $\zeta(C,s) = Z(C,q^{-s})$. Then the "Riemann hypothesis" is equivalent to $\zeta(C,s)$ having all its zeros on the line $\Re(s) = 1/2$, whence its name. One can show that $\zeta(C,s) = \prod_{x\in C} \frac{1}{(1-\mathbb{N}x^{-s})}$, where $\mathbb{N}x$ is the number of elements in the residue field at x, and so the definition of $\zeta(C,s)$ is quite similar to that of $\zeta(\mathbb{Q},s)$.

The zeta function of a curve over $\mathbb{Q}$. Let C be a complete nonsingular curve over $\mathbb{Q}$. For all but finitely many primes p, the reduction $C(p)$ of C modulo p will be a complete nonsingular curve over $\mathbb{F}_p$. We call the primes for which this is true the "good primes" for C and the remainder the "bad primes". We set

$$\zeta(C,s)=\prod_p \zeta_p(C,s)$$

where $\zeta_p(C,s)$ is the zeta function of $C(p)$ when p is a good prime and is as defined in (Serre, Seminaire DPP 1969/70; Oeuvres, Vol II, pp 581–592) when p is a bad prime.

On comparing the expansion of $\zeta(C,s)$ as a Dirichlet series with $\sum n^{-s}$ and using the Riemann hypothesis, one finds that $\zeta(C,s)$ converges for $\Re(s)>3/2$. It is conjectured that it can be analytically continued to the entire complex plane except for simple poles at the negative integers, and that it satisfies a functional equation relating $\zeta(s)$ to $\zeta(2-s)$. Note that we can write

$$\zeta(C,s)=\frac{\zeta(s)\zeta(s-1)}{L(C,s)}$$

where

$$L(C,s)=\prod_p \frac{1}{(1-\alpha_1(p)p^{-s})\cdots(1-\alpha_{2g}(p)p^{-s})}.$$

For an elliptic curve E over $\mathbb{Q}$, there is a pleasant geometric definition of the factors of $L(E,s)$ at the bad primes. Choose a Weierstrass minimal model for E, and reduce it mod p. If $E(p)$ has a node at which each of the two tangents are rational over $\mathbb{F}_p$, then the factor is $(1-p^{-s})^{-1}$; if $E(p)$ has a node at which the tangents are not separately rational over $\mathbb{F}_p$ (this means that the tangent cone is a homogeneous polynomial of degree two variables with coefficients in $\mathbb{F}_p$ that does not factor over $\mathbb{F}_p$), then the factor is $(1+p^{-s})^{-1}$; if $E(p)$ has a cusp, then the factor is 1.

The *geometric conductor* of E is defined to be

$$N=\prod_p p^{f_p}$$

where $f_p=0$ if E has good reduction at p, $f_p=1$ if $E(p)$ has a node as its only singularity, and $f_p\geq 2$ if $E(p)$ has a cusp (with equality unless $p=2,3$). Write

$$\Lambda(s)=(2\pi)^{-s}\Gamma(s)L(E,s).$$

Then it is conjectured that $Z(s)$ can be analytically continued to the entire complex plane as a holomorphic function, and satisfies the functional equation:

$$\Lambda(s)=\pm N^{1-s}\Lambda(2-s).$$

More generally, let $m>0$ be a prime not dividing N and let χ be primitive character of $(\mathbb{Z}/m\mathbb{Z})^{\times}$. If

$$L(E,s)=\sum c_n n^{-s},$$

we define

$$L_\chi(E,s)=\sum c_n\chi(n)n^{-s},$$

and

$$\Lambda_\chi(E,s) = (m/2\pi)^s\Gamma(s)L_\chi(E,s).$$

It is conjectured that $\Lambda_\chi(E,s)$ can be analytically continued to the whole complex plane as a holomorphic function, and that it satisfies the functional equation

$$\Lambda_\chi(E,s) = \pm(g(\chi)\chi(-N)/g(\bar{\chi}))N^{1-s}\Lambda_{\bar{\chi}}(E,2-s)$$

where

$$g(\chi) = \sum_{n=1}^{m} \chi(n)e^{2\pi i n/m}.$$

Review of elliptic curves. (See also Math 679 or, for complete details, Silverman 1986.) Let E be an elliptic curve over an algebraically closed field k, and let $\mathcal{A} = \mathrm{End}(E)$. Then $\mathcal{A} \otimes_{\mathbb{Z}} \mathbb{Q}$ is $\mathbb{Q}$, an imaginary quadratic field, or a quaternion algebra over $\mathbb{Q}$ (the last case only occurs when k has characteristic $p \neq 0$, and then only for supersingular elliptic curves).

Because E has genus 1, the map $\sum n_i[P_i] \mapsto \sum n_i P_i : \mathrm{Div}^0(E) \to E(k)$ defines an isomorphism $J(k) \to E(k)$.

Here $\mathcal{A}$ is the full ring of correspondences of E. Certainly, any element of $\mathcal{A}$ can be regarded as a correspondence on E. Conversely a correspondence

$$E \leftarrow Y \rightarrow E$$

defines a map $E(k) \to E(k)$, and it is easy to see that this is regular.

There are three natural representations of $\mathcal{A}$.

First, let $W = \mathrm{Tgt}_0(E)$. This is a one-dimensional vector space over k. Since every element α of $\mathcal{A}$ fixes 0, α defines an endomorphism $d\alpha$ of W. We therefore obtain a homomorphism $\rho : \mathcal{A} \to \mathrm{End}(W)$.

Next, for any prime $\ell \neq \mathrm{char}(k)$, the Tate module $T_\ell E$ of E is a free $\mathbb{Z}_\ell$-module of rank 2. We obtain a homomorphism $\rho_\ell : \mathcal{A} \to \mathrm{End}(T_\ell E)$.

Finally, when $k = \mathbb{C}$, $H_1(E,\mathbb{Z})$ is a free $\mathbb{Z}$-module of rank 2, and we obtain a homomorphism $\rho_B : \mathcal{A} \to \mathrm{End}(H_1(E,\mathbb{Z}))$.

PROPOSITION 11.5. *When $k = \mathbb{C}$,*

$$\rho_B \otimes \mathbb{Z}_\ell \sim \rho_\ell, \qquad \rho_B \otimes \mathbb{C} \sim \rho \oplus \bar{\rho}.$$

(By this I mean that they are isomorphic as representations; from a more down-to-earth point of view, this means that if we choose bases for the various modules, then the matrix $(\rho_B(\alpha))$ is similar to $(\rho_\ell(\alpha))$ and to $\begin{pmatrix} \rho(\alpha) & 0 \\ 0 & \bar{\rho}(\alpha) \end{pmatrix}$ for all $\alpha \in \mathcal{A}$.)

PROOF. Write $E = \mathbb{C}/\Lambda$. Then $\mathbb{C}$ is the universal covering space of E and Λ is the group of covering transformations. Therefore $\Lambda = \pi_1(E,0)$. From algebraic topology, we know that H_1 is the maximal abelian quotient of π_1, and so (in this case), $H_1(E,\mathbb{Z}) \cong \pi_1(E,0) \cong \Lambda$ (canonical isomorphisms).

The map $\mathbb{C} \to E$ defines an isomorphism $\mathbb{C} \to \mathrm{Tgt}_0(E)$. But Λ is a lattice in $\mathbb{C}$ (regarded as a real vector space), which means that the canonical map $\Lambda \otimes_{\mathbb{Z}} \mathbb{R} \to \mathbb{C}$ is an isomorphism. Now the relation $\rho_B \sim \rho \oplus \bar{\rho}$ follows from (11.3).

Next, note that the group of points of order ℓ^N on E, E_{ℓ^N}, is equal to $\ell^{-N}\Lambda/\Lambda$. There are canonical isomorphisms

$$\Lambda \otimes_{\mathbb{Z}} (\mathbb{Z}/\ell^N\mathbb{Z}) = \Lambda/\ell^N\Lambda \xrightarrow{\ell^{-N}} \ell^{-N}\Lambda/\Lambda = E_{\ell^N}.$$

When we pass to the inverse limit, these isomorphisms give an isomorphism $\Lambda \otimes \mathbb{Z}_\ell \cong T_\ell E$. □

REMARK 11.6. There is yet another representation of $\mathcal{A}$. Let $\Omega^1(E)$ be the space of holomorphic differentials on E. It is a one-dimensional space over k. Moreover, there is a canonical pairing

$$\Omega^1(E) \times \mathrm{Tgt}_0(E) \to k.$$

This is nondegenerate. Therefore the representation of $\mathcal{A}$ on $\Omega^1(E)$ is the transpose of the representation on $\mathrm{Tgt}_0(E)$. Since both representations are one-dimensional, this means that they are equal.

PROPOSITION 11.7. *For any nonzero endomorphism α of E, the degree of α is equal to* $\det(\rho_\ell\alpha)$.

PROOF. Suppose first that $k = \mathbb{C}$, so that we can identify $E(\mathbb{C})$ with $\mathbb{C}/\Lambda$. Then $E(\mathbb{C})_{\mathrm{tors}} = (\Lambda \otimes \mathbb{Q})/\Lambda$, and (11.1) shows that the kernel of the map $E(\mathbb{C})_{\mathrm{tors}} \to E(\mathbb{C})_{\mathrm{tors}}$ defined by α is finite and has order equal to $\det(\rho_B(\alpha))$. But the order of the kernel is $\deg(\alpha)$ and (11.5) shows that $\det(\rho_B(\alpha)) = \det(\rho_\ell(\alpha))$.

For the case of a general k, see Silverman 1986, V, Proposition 2.3. □

COROLLARY 11.8. *Let E be an elliptic curve over $\mathbb{F}_p$; then the numbers α_1 and α_2 occurring in (11.4) are the eigenvalues of Π_p acting on $T_\ell E$ for any $\ell \neq p$.*

PROOF. The elements of $E(\mathbb{F}_q)$ are exactly the elements of $E(\bar{\mathbb{F}}_p)$ that are fixed by $\Pi_q \overset{\mathrm{df}}{=} \Pi_p^n$, i.e., $E(\mathbb{F}_q)$ is the kernel of the endomorphism $\Pi_p^n - 1$. This endomorphism is separable (Π_p obviously acts as zero on the tangent space), and so

$$N_n = \deg(\Pi_p^n - 1) = \det(\rho_\ell(\Pi_p)) = (\alpha_1^n - 1)(\alpha_2^n - 1) = q - \alpha_1^n - \alpha_2^n + 1.$$

□

We need one last fact.

PROPOSITION 11.9. *Let α' be the transpose of the endomorphism α of E; then $\rho_\ell(\alpha')$ is the transpose of $\rho_\ell(\alpha)$.*

The zeta function of $X_0(N)$: case of genus 1. When N is one of the integers 11, 14, 15, 17, 19, 20, 21, 24, 17, 32, 36, or, 49, the curve $X_0(N)$ has genus 1. Recall (discussion before Theorem 6.1) that the number of cusps[5] of $\Gamma_0(N)$ is $\sum_{d|N} \varphi(d, N/d)$. If N is prime, then there are two cusps, 0 and $i\infty$, and they are both rational over $\mathbb{Q}$. If N is one of the above values, and we take $i\infty$ to be the zero element of $X_0(N)$, then it becomes an elliptic curve over $\mathbb{Q}$.

LEMMA 11.10. *There is a natural one-to-one correspondence between the cusp forms of weight 2 for $\Gamma_0(N)$ and the holomorphic differential forms $X_0(N)$ (over $\mathbb{C}$).*

[5]For a description of the cusps on $X_0(N)$ and their fields of rationality, see Ogg, Rational points on certain elliptic modular curves, Proc. Symp. P. Math, 24, AMS, 1973, 221-231.

PROOF. We know that $f \mapsto f dz$ gives a one-to-one correspondence between the meromorphic modular forms of weight 2 for $\Gamma_0(N)$ and the meromorphic differentials on $X_0(N)$, but Lemma 4.11 shows that the cusp forms correspond to the holomorphic differential forms. $\square$

Assume $X_0(N)$ has genus one. Let ω be a holomorphic differential on $X_0(N)$; when we pull it back to $\mathbb{H}$ and write it $f(z)dz$, we obtain a cusp form $f(z)$ for $\Gamma_0(N)$ of weight 2. It is automatically an eigenform for $T(p)$ all $p \nmid N$, and we assume that it is normalized so that $f(z) = \sum a_n q^n$ with $a_1 = 1$. Then $T(p) \cdot f = a_p f$. One can show that a_p is real.

Now consider $\tilde{X}_0(N)$, the reduction of $X_0(N)$ modulo p. Here we have endomorphisms Π_p and Π_p', and $\Pi_p \circ \Pi_p' = \deg(\Pi_p) = p$. Therefore

$$(I_2 - \rho_\ell(\Pi_p)T)(I_2 - \rho_\ell(\Pi_p')T) = I_2 - (\rho_\ell(\Pi_p + \Pi_p'))T + pT^2.$$

According to the Eichler-Shimura theorem, we can replace $\Pi_p + \Pi_p'$ by $\tilde{T}(p)$, and since the ℓ-adic representation doesn't change when we reduce modulo p, we can replace $\tilde{T}(p)$ by $T(p)$. The right hand side becomes

$$I_2 - \begin{pmatrix} a_p & 0 \\ 0 & a_p \end{pmatrix} T + pT^2.$$

Now take determinants, noting that Π_p and Π_p', being transposes, have the same characteristic polynomial. We get that

$$(1 - a_p T + pT^2)^2 = \det(1 - \Pi_p T)^2.$$

On taking square roots, we conclude that

$$(1 - a_p T + pT^2) = \det(1 - \Pi_p T) = (1 - \alpha_p T)(1 - \bar{\alpha}_p T).$$

On replacing T with p^{-s} in this equation, we obtain the equality of the p-factors of the Euler products for the Mellin transform of $f(z)$ and of $L(X_0(N), s)$. We have therefore proved the following theorem.

THEOREM 11.11. *The zeta function of $X_0(N)$ (as a curve over $\mathbb{Q}$) is, up to a finite number of factors, the Mellin transform of $f(z)$.*

COROLLARY 11.12. *The strong Hasse-Weil conjecture (see below) is true for $X_0(N)$.*

PROOF. Apply Theorem 9.5. $\square$

Review of the theory of curves. We repeat the above discussion with E replaced by a general (projective nonsingular) curve C. Proofs can be found (at least when the ground field is $\mathbb{C}$) in Griffiths 1989. Let C be a complete nonsingular curve over an algebraically closed field k. Attached to C there is an abelian variety J, called the *Jacobian variety* of C such that

$$J(k) = \mathrm{Div}^0(C)/\{\text{principal divisors}\}.$$

In the case that C is an elliptic curve, $J = C$, i.e., an elliptic curve is its own Jacobian.

When $k = \mathbb{C}$ it is easy to define J, at least as a complex torus. As we have already mentioned, the Riemann-Roch theorem shows that the holomorphic differentials $\Omega^1(C)$ on C form a vector space over k of dimension $g =$ genus of C.

Now assume $k = \mathbb{C}$. The map

$$H_1(C, \mathbb{Z}) \to \Omega^1(C)^\vee, \quad \gamma \mapsto (\omega \mapsto \int_\gamma \omega),$$

identifies $H_1(C, \mathbb{Z})$ with a lattice in $\Omega^1(C)^\vee$ (linear dual to the vector space $\Omega^1(C)$). Therefore we have a g-dimensional complex torus $\Omega^1(C)^\vee / H_1(C, \mathbb{Z})$. One proves that there is a unique abelian variety J over $\mathbb{C}$ such that $J(\mathbb{C}) = \Omega^1(C)^\vee / H_1(C, \mathbb{Z})$. (Recall that *not* every compact complex manifold of dimension > 1 arises from an algebraic variety.)

We next recall two very famous theorems. Fix a point $P \in C$.

Abel's Theorem: Let $P_1, \ldots, P_r$ and $Q_1, \ldots, Q_r$ be elements of $C(\mathbb{C})$; then there is a meromorphic function on $C(\mathbb{C})$ with its poles at the P_i and its zeros at the Q_i if and only if, for any path γ_i from P to P_i and path γ_i' from P to Q_i, there exists a γ in $H_1(C(\mathbb{C}), \mathbb{Z})$ such that

$$\sum_{i=1}^{r} \int_{\gamma_i} \omega - \sum_{i=1}^{r} \int_{\gamma_i'} \omega = \int_\gamma \omega \quad \text{all } \omega.$$

Jacobi Inversion Formula: For any linear mapping $l : \Omega^1(C) \to \mathbb{C}$, there exist g points $P_1, \ldots, P_g$ in $C(\mathbb{C})$ and paths $\gamma_1, \ldots, \gamma_g$ from P to P_i such that $l(\omega) = \sum \int_{\gamma_i} \omega$ for all $\omega \in \Omega^1(C)$.

These two statements combine to show that there is an isomorphism:

$$\sum n_i P_i \mapsto \left(\omega \mapsto \sum n_i \int_{\gamma_i} \omega\right) : \mathrm{Div}^0(C)/\{\text{principal divisors}\} \to J(\mathbb{C}).$$

(The γ_i are paths from P to P_i.) The construction of J is much more difficult over a general field k. (See my second article in: *Arithmetic Geometry,* eds. G. Cornell and Silverman, Springer, 1986.)

The ring of correspondences $\mathcal{A}$ of C can be identified with the endomorphism ring of J, i.e., with the ring of regular maps $\alpha : J \to J$ such that $\alpha(0) = 0$.

Again, there are three representations of $\mathcal{A}$.

First, we have a representation ρ of $\mathcal{A}$ on $\mathrm{Tgt}_0(J) = \Omega^1(C)^\vee$. This is a vector space of dimension g over the ground field k.

Second, for any $\ell \neq \mathrm{char}(k)$, we have a representation on the Tate module $T_\ell(J) = \varprojlim J_{\ell^n}(k)$. This is a free $\mathbb{Z}_\ell$-module of rank $2g$.

Third, when $k = \mathbb{C}$, we have a representation on $H_1(C, \mathbb{Z})$. This is a free $\mathbb{Z}$-module of rank 2.

PROPOSITION 11.13. *When $k = \mathbb{C}$,*

$$\rho_B \otimes \mathbb{Z}_\ell \sim \rho_\ell, \quad \rho_B \otimes \mathbb{C} \sim \rho \oplus \bar{\rho}.$$

PROOF. This can be proved exactly as in the case of an elliptic curve. □

The rest of the results for elliptic curves extend in an obvious way to a curve C of genus g and its Jacobian variety $J(C)$.

The zeta function of $X_0(N)$: general case. Exactly as in the case of genus 1, the Eichler-Shimura theorem implies the following result.

THEOREM 11.14. *Let $f_1, f_2, ..., f_g$ be a basis for the cusp forms of degree 2 for $\Gamma_0(N)$, chosen to be normalized eigenforms for the Hecke operators $T(p)$ for p prime to N. Then, apart from the factors corresponding to a finite number of primes, the zeta function of $X_0(N)$ is equal to the product of the Mellin transforms of the f_i.*

THEOREM 11.15. *Let f be a cusp form of weight 2, which is a normalized eigenform for the Hecke operators, and write $f = \sum a_n q^n$. Then for all primes $p \nmid N$, $|a_p| \leq 2p^{1/2}$.*

PROOF. In the course of the proof of the theorem, one finds that $a_p = \alpha + \bar{\alpha}$ where α occurs in the zeta function of the reduction of $X_0(N)$ at p. Thus this follows from the Riemann hypothesis. $\square$

REMARK 11.16. As discussed in Section 4, Deligne has proved the analogue of Theorem 11.15 for all weights: let f be a cusp form of weight $2k$ for $\Gamma_0(N)$ and assume f is an eigenform for all the $T(p)$ with p a prime not dividing N and that f is "new" (see below); write $f = \sum_1^\infty a_n q^n$ with $a_1 = 1$; then

$$|a_p| \leq 2p^{2k-1/2},$$

for all p not dividing N. The proof identifies the eigenvalues of the Hecke operator with sums of eigenvalues of Frobenius endomorphisms acting on the étale cohomology of a power of the universal elliptic curve; thus the conjecture follows from the Riemann hypothesis for such varieties. See Deligne, Sém. Bourbaki, Fév. 1969. In fact, Deligne's paper Weil II simplifies the proof (for a few hints concerning this, see E. Freitag and R. Kiehl, Etale Cohomology and the Weil Conjecture, p278).

The Conjecture of Taniyama and Weil. Let E be an elliptic curve over $\mathbb{Q}$. Let N be its geometric conductor. It has an L-series

$$L(E, s) = \sum_{n=1}^{\infty} a_n q^n.$$

For any prime m not dividing N, and primitive character $\chi : (\mathbb{Z}/m\mathbb{Z})^\times \to \mathbb{C}^\times$, let

$$\Lambda_\chi(E, s) = N^{s/2} \left(\frac{m}{2\pi}\right)^s \Gamma(s) \sum_{n=1}^{\infty} a_n \chi(n) q^n.$$

CONJECTURE 11.17 (Strong Hasse-Weil conjecture). *For all m prime to N, and all primitive Dirichlet characters χ, $\Lambda_\chi(E, s)$ has an analytic continuation of $\mathbb{C}$, bounded in vertical strips, satisfying the functional equation*

$$\Lambda_\chi(E, s) = \pm(g(\chi)\chi(-N)/g(\bar{\chi}))N^{1-s}\Lambda_{\bar{\chi}}(E, 2-s)$$

where

$$g(\chi) = \sum_{n=1}^{m} \chi(n) e^{2\pi i n/m}.$$

An elliptic curve E over $\mathbb{Q}$ is said to be *modular* if there is a nonconstant map $X_0(N) \to E$ (defined over $\mathbb{Q}$).

REMARK 11.18. Let C be a complete nonsingular curve, and fix a rational point P on C (assumed to exist). Then there is a canonical map $\varphi_P : C \to J(C)$ sending P to 0, and the map is universal: for any abelian variety A and regular map $\varphi : C \to A$ sending P to 0, there is a unique map $\psi : J(C) \to A$ such that $\psi \circ \varphi_P = \varphi$. Thus to say that E is modular means that there is a surjective homomorphism $J_0(N) \to E$.

THEOREM 11.19. *An elliptic curve E over $\mathbb{Q}$ is modular if and only if it satisfies the strong Hasse-Weil conjecture (and in fact, there is a map $X_0(N) \to E$ with N equal to the geometric conductor of E).*

PROOF. Suppose E is modular, and let ω be the Néron differential on E. The pull-back of ω to $X_0(N)$ can be written $f(z)dz$ with $f(z)$ a cusp form of weight 2 for $\Gamma_0(N)$, and the Eichler-Shimura theorem shows that $\Lambda(E, s)$ is the Mellin transform of f. (Actually, it is not quite this simple...)

Conversely, suppose E satisfies the strong Hasse-Weil conjecture. Then according to Weil's theorem, $\Lambda(E, s)$ is the Mellin transform of a cusp form f. The cusp form has rational Fourier coefficients, and the next proposition shows that there is a quotient E' of $J_0(N)$ whose L-series is the Mellin transform of f; thus we have found a modular elliptic curve having the same zeta function as E, and a theorem of Faltings then shows that there is an isogeny $E' \to E$. □

THEOREM 11.20 (Faltings 1983). *Let E and E' be elliptic curves over $\mathbb{Q}$. If $\zeta(E, s) = \zeta(E', s)$ then E is isogenous to E'.*

PROOF. See his paper proving Mordell's conjecture (Invent. Math. 1983). □

Suppose $M|N$; then we have a map $X_0(N) \to X_0(M)$, and hence a map $J_0(N) \to J_0(M)$. The intersection of the kernels is the "new" part of $J_0(N)$, $J_0^{\text{new}}(N)$.

Similarly, it is possible to define a subspace $\mathcal{S}_0^{\text{new}}(N)$ of new cusp forms of weight 2.

PROPOSITION 11.21. *There is a one-to-one correspondence between the elliptic curves over $\mathbb{Q}$ that are images of $X_0(N)$ but of no $X_0(M)$ with $M < N$ and newforms for $\Gamma_0(N)$ that are eigenforms with rational eigenvalues.*

PROOF. Given a "new" form $f(z) = \sum a_n q^n$ as in the Proposition, we define an elliptic curve E equal to the intersection of the kernels of the endomorphisms $T(p) - a_p$ acting on $J(X_0(N))$. Some quotient of E by a finite subgroup will be the modular elliptic curve sought. □

CONJECTURE 11.22 (Taniyama). *Let E be an elliptic curve over $\mathbb{Q}$ with geometric conductor N. Then there is a nonconstant map $X_0(N) \to E$; in particular, every elliptic curve over $\mathbb{Q}$ is a modular elliptic curve.*

We have proved the following.

THEOREM 11.23. *The strong Hasse-Weil conjecture for elliptic curves over $\mathbb{Q}$ is equivalent to the Taniyama-Weil conjecture.*

Conjecture 6.22 was suggested (a little vaguely) by Taniyama [6] in 1955, and promoted by Shimura. Weil gave rather compelling evidence for it.

Notes. There is a vast literature on the above questions. The best introduction to it is: Elliptic curves and modular functions, H.P.F. Swinnerton-Dyer and B.J. Birch, in Modular Functions of One Variable IV, (eds. Birch and Kuyk), SLN 476, QA343.M72 v.4, pp 2–32. See also: Manin, Parabolic points and zeta-functions of modular curves, Math. USSR 6 (1972), 19–64.

Fermat's last theorem.

THEOREM 11.24. *The Taniyama conjecture implies Fermat's last theorem.*

Idea: It is clear that the Taniyama conjecture restricts the number of elliptic curves over $\mathbb{Q}$ that there can be with small conductor. For example, $X_0(N)$ has genus zero for $N = 1, 2, 3, ..., 10, 12, 13, 16, 18, 25$ and so for these values, the Taniyama-Weil conjecture implies that there can be no elliptic curve with this conductor. (Tate showed a long time ago that there is no elliptic curve over $\mathbb{Q}$ with conductor 1, that is, with good reduction at every prime.)

More precisely, the one proves the following:

THEOREM 11.25. *Let p be a prime > 2, and suppose that*

$$a^p - b^p = c^p$$

with a, b, c all nonzero integers and $\gcd(a, b, c) = 1$. *Then the elliptic curve*

$$E: \quad Y^2 = X(X - a^p)(X + b^p)$$

is not a modular elliptic curve.

PROOF. We can assume that $p > 163$; moreover that $2|b$ and $a \equiv 3 \mod 3$. An easy calculation shows that the curve has bad reduction exactly at the primes p dividing abc, and at each such prime the reduced curve has a node. Thus the geometric conductor is a product of the primes dividing abc.

Suppose that E is a Weil curve. There is a weight 2 cusp form for $\Gamma_0(N)$ with integral q-expansion, and Ribet proves that there is a cusp form of weight 2 for $\Gamma_0(2)$ such that $f \equiv f'$ modulo ℓ. But $X_0(2)$ has genus zero, and so there are no cusp forms of weight 2. □

REMARK 11.26. Ribet's proof is very intricate; it involves a delicate interplay between three primes ℓ, p, and q, which is one more than most of us can keep track of (Ribet, On modular representations of $\mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$ arising from modular forms, Invent. Math 100 (1990), 431–476). As far as I know, the idea of using the elliptic curve in (11.25) to attempt to prove Fermat's last theorem is due to G. Frey. He has published many talks about it, see for example, Frey, Links between solutions of $A - B = C$ and elliptic curves, in Number Theory, Ulm 1987, (ed. H. Schlickewei and Wirsing), SLN 1380.

[6]Taniyama was a very brilliant Japanese mathematician who was the main founder of the theory of complex multiplication of abelian varietes of dimension > 1. He killed himself in late 1958, shortly after his 31st birthday.

Application to the conjecture of Birch and Swinnerton-Dyer. Recall (Silverman 1986) that, for an elliptic curve E over $\mathbb{Q}$, the conjecture of Birch and Swinnerton-Dyer predicts that

$$\lim_{s\to 1}(s-1)^{-r}L(E,s)=\frac{\Omega\prod_p c_p[\mathrm{TS}(E/\mathbb{Q})]R(E/\mathbb{Q})}{[E(\mathbb{Q})_{\mathrm{tors}}]^2}$$

where $r=\mathrm{rank}(E(\mathbb{Q}))$, $\Omega=\int_{E(\mathbb{R})}|\omega|$ where ω is the Néron differential on E, the product of the c_p is over the bad primes, TS is the Tate-Shafarevich group of E, and $R(E/\mathbb{Q})$ is the discriminant of the height pairing.

Now suppose E is a modular elliptic curve. Put the equation for E in Weierstrass minimal form, and let $\omega = dx/(2y+a_1x+a_3)$ be the Néron differential. Assume $\alpha^*\omega = f dz$, for $f(z)$ a newform for $\Gamma_0(N)$. Then $L(E,s)$ is the Mellin transform of $f(z)$. Write $f(z)=c(q+a_2q^2+...)q^{-1}dq$, where c is a positive rational number. Conjecturally $c=1$, and so I drop it.

Assume that $i\infty$ maps to $0\in E$. Then q is real for z on the imaginary axis between 0 and $i\infty$. Therefore $j(z)$ and $j(Nz)$ are real, and, as we explained (end of Section 8) this means that the image of the imaginary axis in $X_0(N)(\mathbb{C})$ is in $X_0(N)(\mathbb{R})$, i.e., the points in the image of the imaginary axis have real coordinates.

The Mellin transform formula (cf. 9.2) implies that

$$L(E,1)=\Gamma(1)L(E,1)=\int_0^{i\infty}f(z)dz.$$

Define M by the equation

$$\int_0^{i\infty}f(z)dz=M\cdot\int_{E(\mathbb{R})}\omega.$$

Intuitively at least, M is the winding number of the map from the imaginary axis from 0 to $i\infty$ onto $E(\mathbb{R})$. The image of the point 0 in $X_0(N)$ is known to be a point of finite order, and this implies that the winding number is a rational number. Thus, for a modular curve (suitably normalized), the conjecture of Birch and Swinnerton-Dyer can be restated as follows.

CONJECTURE 11.27 (Birch and Swinnerton-Dyer). (a) *$E(\mathbb{Q})$ is infinite if and only if $M=0$.*

(b) *If $M\neq 0$, then $M[E(\mathbb{Q})]^2=[\mathrm{TS}(E/\mathbb{Q})]\prod_p c_p$.*

REMARK 11.28. Some remarkable results have been obtained in this context by Kolyvagin and others. (See: Rubin, The work of Kolyvagin on the arithmetic of elliptic curves, SLN 1399, MR 90h:14001), and the papers of Kolyvagin.)

More details can be found in the article of Birch and Swinnerton-Dyer mentioned above. Winding numbers and the mysterious c are discussed in Mazur and Swinnerton-Dyer, Inventiones math., 25, 1-61, 1974. See also the article of Manin mentioned above and my Math 679 notes.

12. Complex Multiplication for Elliptic Curves

The theory of complex multiplication is not only the most beautiful part of mathematics but also of the whole of science.
D. Hilbert.

It was known to Gauss that $\mathbb{Q}[\zeta_n]$ is an abelian extension of $\mathbb{Q}$. Towards the end of the 1840's Kronecker had the idea that cyclotomic fields, and their subfields, exhaust the abelian extensions of $\mathbb{Q}$, and furthermore, that every abelian extension of a quadratic imaginary number field $E = \mathbb{Q}[\sqrt{-d}]$ is contained in the extension given by adjoining to E roots of 1 and certain special values of the modular function j. Many years later, he was to refer to this idea as the most cherished dream[7] of his youth (mein liebster Jugendtraum) (Kronecker, Werke, V, p435).

Abelian extensions of $\mathbb{Q}$. Let $\mathbb{Q}^{\text{cyc}} = \cup\mathbb{Q}[\zeta_n]$; it is a subfield of the maximal abelian extension $\mathbb{Q}^{\text{ab}}$ of $\mathbb{Q}$.

THEOREM 12.1 (Kronecker-Weber). *The field* $\mathbb{Q}^{\text{cyc}} = \mathbb{Q}^{ab}$.

The proof has two steps.

Elementary part. Note that there is a homomorphism

$$\chi : \mathrm{Gal}(\mathbb{Q}[\zeta_n]/\mathbb{Q}) \to (\mathbb{Z}/n\mathbb{Z})^\times, \quad \sigma\zeta = \zeta^{\chi(m)},$$

which is obviously injective. Proving that it is surjective is equivalent to proving that the cyclotomic polynomial

$$\Phi_n(X) \stackrel{\text{df}}{=} \prod_{(m.n)=1} (X - \zeta^m)$$

is irreducible in $\mathbb{Q}[X]$, or that $\mathrm{Gal}(\mathbb{Q}[\zeta_n]/\mathbb{Q})$ acts transitively on the primitive n^{th} roots of 1. One way of doing this is to look modulo p, and exploit the Frobenius map (see Math 594f, 5.9).

Application of class field theory. For any abelian extension F of $\mathbb{Q}$, class field theory provides us with a surjective homomorphism (the Artin map)

$$\phi : \mathbb{I} \to \mathrm{Gal}(F/\mathbb{Q})$$

where $\mathbb{I}$ is the group of idèles of $\mathbb{Q}$ (see Math 776). When we pass to the inverse limit over all F's, then we obtain an exact sequence

$$1 \to (\mathbb{Q}^\times \cdot \mathbb{R}^+)^- \to \mathbb{I} \to \mathrm{Gal}(\mathbb{Q}^{\text{ab}}/\mathbb{Q}) \to 1$$

where $\mathbb{R}^+ = \{r \in \mathbb{R} \mid r > 0\}$, and the bar denotes the closure.

Consider the homomorphisms

$$\mathbb{I} \to \mathrm{Gal}(\mathbb{Q}^{\text{ab}}/\mathbb{Q}) \to \mathrm{Gal}(\mathbb{Q}^{\text{cyc}}/\mathbb{Q}) \xrightarrow{\chi} \varprojlim(\mathbb{Z}/m\mathbb{Z})^\times = \hat{\mathbb{Z}}^\times.$$

All maps are surjective. In order to show that the middle map is an isomorphism, we have to prove that the kernel of $\mathbb{I} \to \hat{\mathbb{Z}}^\times$ is $(\mathbb{Q}^\times \cdot \mathbb{R}^+)^-$; it clearly contains $(\mathbb{Q}^\times \cdot \mathbb{R}^+)^-$.

Note that $\hat{\mathbb{Z}} = \prod \mathbb{Z}_\ell$, and that $\hat{\mathbb{Z}}^\times = \prod \mathbb{Z}_\ell^\times$. There is therefore a canonical embedding $i : \hat{\mathbb{Z}} \hookrightarrow \mathbb{I}$, and to complete the proof of the theorem, it suffices to show:

[7]See a series of articles in preparation by N. Schappacher for a complete and very careful treatment of Kronecker's Jugendtraum and the work that grew out of it.

(i) the composite $\hat{\mathbb{Z}}^\times \xrightarrow{i} \mathbb{I} \to \hat{\mathbb{Z}}^\times$ is the identity map;
(ii) $(\mathbb{Q}^\times \cdot \mathbb{R}^+)^- \cdot i(\hat{\mathbb{Z}}^\times) = \mathbb{I}$.

Assume these statements, and let $\alpha \in \mathbb{I}$. Then (ii) says that $\alpha = a \cdot i(z)$ with $a \in (\mathbb{Q}^\times \cdot \mathbb{R}^+)^-$ and $z \in \mathbb{Z}^\times$, and (i) shows that $\varphi(a \cdot z) = z$. Thus, if $\alpha \in \mathrm{Ker}(\varphi)$, then $z = 1$, and $\alpha \in (\mathbb{Q}^\times \cdot \mathbb{R}^+)^-$.

The proofs of (i) and (ii) are left as an exercise (see Math 776, V.5.9).

Alternative: Find the kernel of $\phi : (\mathbb{A}^\times/\mathbb{Q}^\times) \to \mathrm{Gal}(\mathbb{Q}[\zeta_n]/\mathbb{Q})$, and show that every open subgroup of finite index contains such a subgroup.

Alternative: For a proof using only local (i.e., not global) class field theory, see Math 776, I.4.16.

Orders in K. Let K be a quadratic imaginary number field. An *order* of K is a subring R containing $\mathbb{Z}$ and free of rank 2 over $\mathbb{Z}$. Clearly every element of R is integral over $\mathbb{Z}$, and so $R \subset \mathcal{O}_K$ (ring of integers in K). Thus $\mathcal{O}_K$ is the unique maximal order.

PROPOSITION 12.2. *Let R be an order in K. Then there is a unique integer $f > 0$ such that $R = \mathbb{Z} + f \cdot \mathcal{O}_K$. Conversely, for any integer $f > 0$, $\mathbb{Z} + f \cdot \mathcal{O}_K$ is an order in K.*

PROOF. Let $\{1, \alpha\}$ be a $\mathbb{Z}$-basis for $\mathcal{O}_K$, so that $\mathcal{O}_K = \mathbb{Z} + \mathbb{Z}\alpha$. Then $R \cap \mathbb{Z}\alpha$ is a subgroup of $\mathbb{Z}\alpha$, and hence equals $\mathbb{Z}\alpha f$ for some positive integer f. Now $\mathbb{Z} + f\mathcal{O}_K \subset \mathbb{Z} + \mathbb{Z}\alpha f \subset R$. Conversely, if $m + n\alpha \in R$, $m, n \in \mathbb{Z}$, then $n\alpha \in R$, and so $n \in f\mathbb{Z}$. Thus, $m + n\alpha \in \mathbb{Z} + f\alpha\mathbb{Z} \subset \mathbb{Z} + f\mathcal{O}_K$. $\square$

The number f is called the *conductor* of R. We often write R_f for $\mathbb{Z} + f \cdot \mathcal{O}_K$.

PROPOSITION 12.3. *Let R be an order in K. The following conditions on an R-submodule $\mathfrak{a}$ of K are equivalent:*

(a) *$\mathfrak{a}$ is a projective R-module;*
(b) *$R = \{a \in K \mid a \cdot \mathfrak{a} \subset \mathfrak{a}\}$;*
(c) *$\mathfrak{a} = x \cdot \mathcal{O}_K$ for some $x \in \mathbb{I}$ (this means that for all primes v of $\mathcal{O}_K$, $\mathfrak{a} \cdot \mathcal{O}_v = x_v \cdot \mathcal{O}_v$).*

PROOF. For (b) $\implies$ (c), see Shimura 1971, (5.4.2), p 122. $\square$

Such an R-submodule of K is called a *proper* R-ideal. A proper R-ideal of the form αR, $\alpha \in K^\times$, is said to be *principal.* If $\mathfrak{a}$ and $\mathfrak{b}$ are two proper R-ideals, then

$$\mathfrak{a} \cdot \mathfrak{b} \overset{\text{df}}{=} \{\sum a_i b_i \mid a_i \in \mathfrak{a}, \quad b_i \in \mathfrak{b}\}$$

is again a proper R-ideal.

PROPOSITION 12.4. *For any order R in K, the proper R-ideals form a group with respect to multiplication, with R as the identity element.*

PROOF. Shimura 1971, Proposition 4.11, p105. $\square$

The *class group* $Cl(R)$ is defined to be the quotient of the group of proper R-ideals by the subgroup of principal ideals. When R is the full ring of integers in E, then $Cl(R)$ is the usual class group.

REMARK 12.5. The class number of R is

$$h(R) = h \cdot f \cdot (\mathcal{O}_K^\times : R^\times)^{-1} \cdot \prod_{p|f} \left(1 - \left(\frac{K}{p}\right) p^{-1}\right)$$

where h is the class number of $\mathcal{O}_K$, and $(\frac{K}{p}) = 1, -1, 0$ according as p splits in K, stays prime, or ramifies. (If we write $\{\pm 1\}$ for the Galois group of K over $\mathbb{Q}$, then $p \mapsto (\frac{K}{p})$ is the reciprocity map.) See Shimura 1971, Exercise 4.12.

Elliptic curves over $\mathbb{C}$. For any lattice Λ in $\mathbb{C}$, the Weierstrass $\wp$ and $\wp'$ functions realize $\mathbb{C}/\Lambda$ as an elliptic curve $E(\Lambda)$, and every elliptic curve over $\mathbb{C}$ arises in this way. If Λ and Λ' are two lattices, and α is an element of $\mathbb{C}$ such that $\alpha\Lambda \subset \Lambda'$, then $[z] \mapsto [\alpha z]$ is a homomorphism $E(\Lambda) \to E(\Lambda')$, and every homomorphism is of this form; thus

$$\mathrm{Hom}(E(\Lambda), E(\Lambda')) = \{\alpha \in \mathbb{C} \mid \alpha\Lambda \subset \Lambda'\}.$$

In particular, $E(\Lambda) \approx E(\Lambda')$ if and only if $\Lambda' = \alpha\Lambda$ for some $\alpha \in \mathbb{C}^\times$.

These statements reduce much of the theory of elliptic curves over $\mathbb{C}$ to linear algebra. For example, $\mathrm{End}(E)$ is either $\mathbb{Z}$ or an order R in a quadratic imaginary field K. Consider $E = E(\Lambda)$; if $\mathrm{End}(E) \neq \mathbb{Z}$, then there is an $\alpha \in \mathbb{C}$, $\alpha \notin \mathbb{Z}$, such that $\alpha\Lambda \subset \Lambda$, and

$$\mathrm{End}(E) = \{\alpha \in \mathbb{C} \mid \alpha\Lambda \subset \Lambda\},$$

which is an order in $\mathbb{Q}[\alpha]$ having Λ as a proper ideal.

When $\mathrm{End}(E) = R \neq \mathbb{Z}$, we say E has *complex multiplication* by R.

Write $E = E(\Lambda)$, so that $E(\mathbb{C}) = \mathbb{C}/\Lambda$. Clearly $E_n(\mathbb{C})$, the set of points of order dividing n on E, is equal to $n^{-1}\Lambda/\Lambda$, and so it is a free $\mathbb{Z}/n\mathbb{Z}$-module of rank 2. The inverse limit, $T_\ell E \overset{\text{df}}{=} \varprojlim E_{\ell^m} = \varprojlim \ell^{-m}\Lambda/\Lambda = \Lambda \otimes \mathbb{Z}_\ell$, and so $V_\ell E = \Lambda \otimes \mathbb{Q}_\ell$.

Algebraicity of j. When R is an order in a quadratic imaginary field $K \subset \mathbb{C}$, we write $Ell(R)$ for the set of isomorphism classes of elliptic curves over $\mathbb{C}$ with complex multiplication by R.

PROPOSITION 12.6. *For each proper R-ideal $\mathfrak{a}$, $E(\mathfrak{a}) \overset{\text{df}}{=} \mathbb{C}/\mathfrak{a}$ is an elliptic curve with complex multiplication by R, and the map $\mathfrak{a} \mapsto \mathbb{C}/\mathfrak{a}$ induces a bijection*

$$Cl(R) \to Ell(R).$$

PROOF. If $\mathfrak{a}$ is a proper R-ideal, then

$$\begin{aligned} \mathrm{End}(E(\mathfrak{a})) &= \{\alpha \in \mathbb{C} \mid \alpha\mathfrak{a} \subset \mathfrak{a}\} \text{ (see above)} \\ &= \{\alpha \in K \mid \alpha\mathfrak{a} \subset \mathfrak{a}\} \text{ (easy)} \\ &= R \text{ (definition of proper } R\text{-ideal).} \end{aligned}$$

Since $E(\alpha \cdot \mathfrak{a}) \approx E(\mathfrak{b})$ we get a well-defined map $Cl(R) \to Ell(R)$. Similar arguments show that it is bijective. □

COROLLARY 12.7. *Up to isomorphism, there are only finitely many elliptic curves over $\mathbb{C}$ with complex multiplication by R; in fact there are exactly $h(R)$.*

With an elliptic curve E over $\mathbb{C}$, we can associate its j-invariant $j(E) \in \mathbb{C}$, and $E \approx E'$ if and only if $j(E) = j(E')$. For an automorphism σ of $\mathbb{C}$, we define σE to be the curve obtained by applying σ to the coefficients of the equation defining E. Clearly $j(\sigma E) = \sigma j(E)$.

THEOREM 12.8. *If E has complex multiplication then $j(E)$ is algebraic.*

PROOF. Let $z \in \mathbb{C}$. If z is algebraic (meaning algebraic over $\mathbb{Q}$), then z has only finitely many conjugates, i.e., as σ ranges over the automorphisms of $\mathbb{C}$, σz ranges over a finite set. The converse of this is also true: if z is transcendental, then σz takes on uncountably many different values (if z' is any other transcendental number, there is an isomorphism $\mathbb{Q}[z] \to \mathbb{Q}[z']$ which can be extended to an automorphism of $\mathbb{C}$).

Now consider $j(E)$. As σ ranges over $\mathbb{C}$, σE ranges over finitely many isomorphism classes, and so $\sigma j(E)$ ranges over a finite set. This shows that $j(E)$ is algebraic. □

COROLLARY 12.9. *Let j be the (usual) modular function for $\Gamma(1)$, and let $z \in \mathbb{H}$ be such that $\mathbb{Q}[z]$ is a quadratic imaginary number field. Then $j(z)$ is a algebraic.*

PROOF. The function j is defined so that $j(z) = j(E(\Lambda))$, where $\Lambda = \mathbb{Z} + \mathbb{Z}z$. Suppose $\mathbb{Q}[z]$ is a quadratic imaginary number field. Then

$$\{\alpha \in \mathbb{C} \mid \alpha(\mathbb{Z} + \mathbb{Z}z) \subset \mathbb{Z} + \mathbb{Z}z\}$$

is an order R in $\mathbb{Q}[z]$, and $E(\Lambda)$ has complex multiplication by R, from which the statement follows. □

The integrality of j. Let E be an elliptic curve over a field k and let R be an order in a quadratic imaginary number field K. When we are given a isomorphism $i : R \to \operatorname{End}(E)$, we say that E has complex multiplication by R (defined over k). Then R and $\mathbb{Z}_\ell$ act on $T_\ell E$, and therefore $R \otimes_{\mathbb{Z}} \mathbb{Z}_\ell$ acts on $T_\ell E$; moreover, $K \otimes_{\mathbb{Q}} \mathbb{Q}_\ell$ acts on $V_\ell E \overset{\mathrm{df}}{=} T_\ell E \otimes \mathbb{Q}$. These actions commute with the actions of $\operatorname{Gal}(k^{\mathrm{al}}/k)$ on the modules.

Let α be an endomorphism of an elliptic curve E over a field k. Define,

$$\operatorname{Tr}(\alpha) = 1 + \deg(\alpha) - \deg(1-\alpha) \in \mathbb{Z},$$

and define the characteristic polynomial of α to be

$$f_\alpha(X) = X^2 - \operatorname{Tr}(\alpha)X + \deg(\alpha) \in \mathbb{Z}[X].$$

PROPOSITION 12.10. (a) *The endomorphism $f_\alpha(\alpha)$ of E is zero.*
(b) *For all $\ell \neq \operatorname{char}(k)$, $f_\alpha(X)$ is the characteristic polynomial of α acting on $V_\ell E$.*

PROOF. Part (b) is proved in Silverman 1986, 2.3, p134. Part (a) follows from (b), the Cayley-Hamilton theorem, and the fact that the $\operatorname{End}(E)$ acts faithfully on $V_\ell E$ (Silverman 1986, 7.4, p92). □

COROLLARY 12.11. *If E has complex multiplication by $R \subset K$, then $V_\ell E$ is a free $K \otimes \mathbb{Q}_\ell$-module.*

PROOF. When the ground field $k = \mathbb{C}$, this is obvious because $V_\ell E = \Lambda \otimes_{\mathbb{Z}} \mathbb{Q}_\ell$, and $\Lambda \otimes_{\mathbb{Z}} \mathbb{Q}_\ell = (\Lambda \otimes_{\mathbb{Z}} \mathbb{Q}) \otimes_{\mathbb{Q}} \mathbb{Q}_\ell = K \otimes_{\mathbb{Q}} \mathbb{Q}_\ell$. When $K \otimes_{\mathbb{Q}} \mathbb{Q}_\ell$ is a field, it is again obvious (every module over a field is free). Otherwise $K \otimes_{\mathbb{Q}} \mathbb{Q}_\ell = K_v \oplus K_w$ where v and w are the primes of K lying over p, and we have to see that $V_\ell E$ is isomorphic to the $K \otimes \mathbb{Q}_\ell$-module $K_v \oplus K_w$ (rather than $K_v \oplus K_v$ for example). But for $\alpha \in K$, $\alpha \notin \mathbb{Q}$, the proposition shows that characteristic polynomial of α acting on $V_\ell E$ is the minimum polynomial of α over K, and this implies what we want. □

REMARK 12.12. In fact $T_\ell E$ is a free $R \otimes \mathbb{Z}_\ell$-module (see J-P. Serre and J. Tate, Good reduction of abelian varieties, Ann. of Math. 88, 1968, pp 492-517, p502).

PROPOSITION 12.13. *The action of* $\mathrm{Gal}(k^{al}/k)$ *on* $V_\ell E$ *factors through* $K \otimes \mathbb{Q}_\ell$*, i.e., there is a homomorphism* $\rho_\ell : \mathrm{Gal}(k^{al}/k) \to (K \otimes \mathbb{Q}_\ell)^\times$ *such that*

$$\rho_\ell(\sigma) \cdot x = \sigma x, \quad \text{all } \sigma \in \mathrm{Gal}(k^{al}/k), \quad x \in V_\ell A.$$

PROOF. The action of $\mathrm{Gal}(k^{\mathrm{al}}/k)$ on $V_\ell E$ commutes with the action R (because we are assuming that the action of R is defined over k). Therefore the image of $\mathrm{Gal}(k^{\mathrm{al}}/k)$ lies in $\mathrm{End}_{K\otimes\mathbb{Q}_\ell}(V_\ell E)$, which equals $K \otimes \mathbb{Q}_\ell$, because $V_\ell E$ is free $K \otimes \mathbb{Q}_\ell$-module of rank 1. □

In particular, we see that the image of ρ_ℓ is abelian, and so the action of $\mathrm{Gal}(k^{\mathrm{al}}/k)$ factors through $\mathrm{Gal}(k^{\mathrm{ab}}/k)$—all the ℓ^m-torsion points of E are rational over k^{ab} for all m. As $\mathrm{Gal}(k^{\mathrm{al}}/k)$ is compact, $Im(\rho_\ell) \subset \mathcal{O}_\ell^\times$, where $\mathcal{O}_\ell$ is the ring of integers in $K \otimes_{\mathbb{Q}} \mathbb{Q}_\ell$ ($\mathcal{O}_\ell$ is either a complete discrete valuation ring or the product of two such rings).

THEOREM 12.14. *Let E be an elliptic curve over a number field k having complex multiplication by R over k. Then E has potential good reduction at every prime v of k.*

PROOF. Let ℓ be a prime number not divisible by v. According to Silverman 1986, VII.7.3, p186, we have to show that the action of the inertia group I_v at v on $T_\ell A$ factors through a finite quotient. But we know that it factors through the inertia subgroup J_v of $\mathrm{Gal}(k^{\mathrm{ab}}/k)$, and class field theory tells us that there is a surjective map

$$\mathcal{O}_v^\times \to J_v$$

where $\mathcal{O}_v$ is the ring of integers in k_v. Thus we obtain a homomorphism

$$\mathcal{O}_v^\times \to J_v \to \mathcal{O}_\ell^\times \subset \mathrm{Aut}(T_\ell E),$$

where $\mathcal{O}_\ell$ is the ring of integers in $K \otimes \mathbb{Q}_\ell$. I claim that any homomorphism $\mathcal{O}_v^\times \to \mathcal{O}_\ell^\times$ automatically factors through a finite quotient. In fact algebraic number theory shows that $\mathcal{O}_v^\times$ has a subgroup U^1 of finite index which is a pro-p-group, where p is the prime lying under v. Similary, $\mathcal{O}_\ell^\times$ has a subgroup of finite index V which is a pro-ℓ-group. Any map from a pro-p group to a pro-ℓ-group is zero, and so $\mathrm{Ker}(U^1 \to \mathcal{O}^\times) = \mathrm{Ker}(U^1 \to \mathcal{O}^\times/V)$, which shows that the homomorphism is zero on a subgroup of finite index of U^1. □

COROLLARY 12.15. *If E is an elliptic curve over a number field k with complex multiplication, then $j(E) \in \mathcal{O}_K$.*

PROOF. An elementary argument shows that, if E has good reduction at v, then $j(E) \in \mathcal{O}_v$ (cf. Silverman 1986, VII.5.5, p181). □

COROLLARY 12.16. *Let j be the (usual) modular function for $\Gamma(1)$, and let $z \in \mathbb{H}$ be such that $\mathbb{Q}[z]$ is a quadratic imaginary number field. Then $j(z)$ is an algebraic integer.*

REMARK 12.17. There are analytic proofs of the integrality of $j(E)$, but they are less illuminating.

Statement of the main theorem (first form). Let K be a quadratic imaginary number field, with ring of integers $\mathcal{O}_K$, and let $Ell(\mathcal{O}_K)$ be the set of isomorphism classes of elliptic curves over $\mathbb{C}$ with complex multiplication by $\mathcal{O}_K$. For any fractional $\mathcal{O}_K$-ideal Λ in K, we write $j(\Lambda)$ for $j(\mathbb{C}/\Lambda)$. (Thus if $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$ where $z = \omega_1/\omega_2$ lies in the upper half plane, then $j(\Lambda) = j(z)$, where $j(z)$ is the standard function occurring in the theory of elliptic modular functions.)

THEOREM 12.18. (a) *For any elliptic curve E over $\mathbb{C}$ with complex multiplication by $\mathcal{O}_K$, $K[j(E)]$ is the Hilbert class field K^{hcf} of K.*
(b) *The group $\mathrm{Gal}(K^{\mathrm{hcf}}/K)$ permutes the set $\{j(E) \mid E \in Ell(\mathcal{O}_K)\}$ transitively.*
(c) *For each prime ideal $\mathfrak{p}$ of K, $\mathrm{Frob}(\mathfrak{p})(j(\Lambda)) = j(\Lambda \cdot \mathfrak{p}^{-1})$.*

The proof will occupy the next few subsections.

The theory of $\mathfrak{a}$-isogenies. Let R be an order in K, and let $\mathfrak{a}$ be a proper ideal in R. For an elliptic curve E over a field k with complex multiplication by R, we define

$$\mathrm{Ker}(\mathfrak{a}) = \cap_{a \in \mathfrak{a}} \mathrm{Ker}(a : E \to E).$$

Note that if $\mathfrak{a} = (a_1, \dots, a_n)$, then $\mathrm{Ker}(\mathfrak{a}) = \cap \mathrm{Ker}(a_i : E \to E)$. Let Λ be a proper R-ideal, and consider the elliptic curve $E(\Lambda)$ over $\mathbb{C}$. Then $\Lambda \cdot \mathfrak{a}^{-1}$ is also a proper ideal.

LEMMA 12.19. *There is a canonical map $E(\Lambda) \to E(\Lambda \cdot \mathfrak{a}^{-1})$ with kernel $\mathrm{Ker}(\mathfrak{a})$.*

PROOF. Since $\Lambda \subset \Lambda \cdot \mathfrak{a}^{-1}$, we can take the map to be $z + \Lambda \mapsto z + \Lambda \cdot \mathfrak{a}^{-1}$. □

PROPOSITION 12.20. *Let E be an elliptic curve over k with complex multiplication by R, and let $\mathfrak{a}$ be a proper ideal in R. Assume k has characteristic zero. Then there is an elliptic curve $\mathfrak{a} \cdot E$ and a homomorphism map $\varphi_{\mathfrak{a}} : E \to \mathfrak{a} \cdot E$ whose kernel is $\mathrm{Ker}(\mathfrak{a})$. The pair $(\mathfrak{a} \cdot E, \varphi_{\mathfrak{a}})$ has the following universal property: for any homomorphism $\varphi : E \to E'$ with $\mathrm{Ker}(\varphi) \supset \mathrm{Ker}(\mathfrak{a})$, there is a unique homomorphism $\psi : \mathfrak{a} \cdot E \to E'$ such that $\psi \circ \varphi_{\mathfrak{a}} = \varphi$.*

PROOF. When $k = \mathbb{C}$, we write $E = E(\Lambda)$ and take $\mathfrak{a} \cdot E = E(\Lambda \cdot \mathfrak{a}^{-1})$. If k is a field of characteristic zero, we define $\mathfrak{a} \cdot E = E(\Lambda \cdot \mathfrak{a}^{-1})$ (see Silvermann 1986, 4.12, 4.13.2, p78). □

We want to extend the definition of $\mathfrak{a} \cdot E$ to the case where k need not have characteristic zero. For this, we define $\mathfrak{a} \cdot E$ to be the image of the map

$$x \mapsto (a_1 x, ..., a_n x) : E \to E^n, \quad \mathfrak{a} = (a_1, \dots, a_n),$$

and $\varphi_{\mathfrak{a}}$ to be this map. We call the isogeny $\varphi_{\mathfrak{a}} : E \to \mathfrak{a} \cdot E$ (or any isogeny that differs from it by an isomorphism) an $\mathfrak{a}$-*isogeny*. The degree of an $\mathfrak{a}$-isogeny is $\mathbb{N}(\mathfrak{a}) \overset{\mathrm{df}}{=} (\mathcal{O}_K : \mathfrak{a})$.

We obtain an action, $(\mathfrak{a}, E) \mapsto \mathfrak{a} \cdot E$, of $Cl(R)$ on $Ell_k(R)$, the set of isomorphism classes of elliptic curves over k having complex multiplication by R.

PROPOSITION 12.21. *The action of $Cl(R)$ on $Ell_k(R)$ makes $Ell_k(R)$ into a principal homogeneous space for $Cl(R)$, i.e., for any $x_0 \in Ell(R)$, the map $\mathfrak{a} \mapsto \mathfrak{a} \cdot x_0 : Cl(R) \to Ell(R)$ is a bijection.*

PROOF. When $k = \mathbb{C}$, this is a restatement of an earlier result (before we implicitly took x_0 to be the isomorphism of class of $\mathbb{C}/R$, and considered the map $Cl(R) \to Ell(R)$, $\mathfrak{a} \to \mathfrak{a}^{-1} \cdot x_0$). We omit the proof of the general case, although this is a key point. □

Reduction of elliptic curves. Let E be an elliptic curve over a number field k with good reduction at a prime ideal v of k. For simplicity, assume that $\mathfrak{p}$ does not divide 2 or 3. Then E has an equation

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

with coefficients in $\mathcal{O}_v$ whose discriminant Δ is not divisible by $\mathfrak{p}_\mathfrak{v}$.

Reduction of the tangent space. Recall that for a curve C defined by an equation $F(X, Y) = 0$, the tangent space at (a, b) on the curve is defined by the equation:

$$\left.\frac{\partial F}{\partial X}\right|_{(a,b)} (X - a) + \left.\frac{\partial F}{\partial X}\right|_{(a,b)} (X - b) = 0.$$

For example, for

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

we find that the tangent space to E at $(0, 0)$ is given by the equation

$$Z = 0.$$

Now take a Weierstrass minimal equation for E over $\mathcal{O}_v$—we can think of the equation as defining a curve $\mathcal{E}$ over $\mathcal{O}_v$, and use the same procedure to define the tangent space $Tgt_0(\mathcal{E})$ at 0 on $\mathcal{E}$—it is an $\mathcal{O}_v$-module.

PROPOSITION 12.22. *The tangent space $Tgt_0(\mathcal{E})$ at 0 to $\mathcal{E}$ is a free $\mathcal{O}_v$-module of rank one such that*

$$Tgt_0(\mathcal{E}) \otimes_{\mathcal{O}_v} K_v = Tgt_0(E/K_v), \quad Tgt_0(\mathcal{E}) \otimes_{\mathcal{O}_v} \kappa(v) = Tgt_0(E(v))$$

where $\kappa(v) = \mathcal{O}_v/\mathfrak{p}_\mathfrak{v}$ and $E(v)$ is the reduced curve.

PROOF. Obvious. □

Thus we can identify $Tgt_0(\mathcal{E})$ (in a natural way) with a submodule of $Tgt_0(E)$, and $Tgt_0(E(v)) = Tgt_0(\mathcal{E})/\mathfrak{m}_v \cdot Tgt_0(\mathcal{E})$, where $\mathfrak{m}_v$ is the maximal ideal of $\mathcal{O}_v$.

Reduction of endomorphisms. Let $\alpha : E \to E'$ be a homomorphism of elliptic curves over k, and assume that both E and E' have good reduction at a prime v of k. Then α defines a homomorphism $\alpha(v) : E(v) \to E'(v)$ of the reduced curves. Moreover, α acts as expected on the tangent spaces and the points of finite order. In more detail:

(a) the map $Tgt_0(\alpha) : Tgt_0(E) \to Tgt_0(E')$ maps $Tgt_0(\mathcal{E})$ into $Tgt_0(\mathcal{E}')$, and induces the map $Tgt_0(\alpha(v))$ on the quotient modules;

(b) recall (Silverman 1986) that for $\ell \neq char(\kappa(v))$ the reduction map defines an isomorphism $T_\ell(E) \to T_\ell(E_0)$; there is a commutative diagram:

$$\begin{array}{ccc} T_\ell E & \xrightarrow{\alpha} & T_\ell E' \\ \downarrow & & \downarrow \\ T_\ell E(v) & \xrightarrow{\alpha_0} & T_\ell E'(v). \end{array}$$

It follows from (b) and Proposition 12.10 that α and α_0 have the same characteristic polynomial (hence the same degree).

Also, we shall need to know that the reduction of an $\mathfrak{a}$-isogeny is an $\mathfrak{a}$-isogeny (this is almost obvious from the definition of an $\mathfrak{a}$-isogeny).

Finally, consider an $\mathfrak{a}$-isogeny $\varphi : E \to E'$; it gives rise to a homomorphism

$$Tgt_0(E) \to Tgt_0(E')$$

whose kernel is $\cap Tgt_0(a)$, a running through the elements of $\mathfrak{a}$ (this again is almost obvious from the definition of $\mathfrak{a}$-isogeny).

The Frobenius map. Let E be an elliptic curve over the finite field $k \supset \mathbb{F}_p$. If E is defined by

$$Y^2 = X^3 + aX + b,$$

then write $E^{(q)}$ for the elliptic curve

$$Y^2 = X^3 + a^q X + b^q.$$

Then the Frobenius map Frob_q is defined to be

$$(x, y) \mapsto (x^q, y^q) : E \to E^{(q)},$$

PROPOSITION 12.23. *The Frobenius map $Frob_p$ is a purely inseparable isogeny of degree p; if $\varphi : E \to E'$ is a second purely inseparable isogeny of degree p, then there is an isomorphism $\alpha : E^{(p)} \to E'$ such that $\alpha \circ Frob_p = \varphi$.*

PROOF. This is similar to Silverman 1986, 2.11, p30. We have $(Frob_p)^*(k(E^{(p)})) = k(E)^p$, which the unique subfield of $k(E)$ such that $k(E) \supset k(E)^p$ is a purely inseparable extension of degree p. □

REMARK 12.24. There is the following criterion: A homomorphism $\alpha : E \to E'$ is separable if and only the map it defines on the tangent spaces $Tgt_0(E) \to Tgt_0(E')$ is an isomorphism.

Proof of the main theorem. The group $G = \mathrm{Gal}(\mathbb{Q}^{\mathrm{al}}/K)$ acts on $Ell(R)$, and commutes with the action of $Cl(R)$. Fix an $x_0 \in Ell(R)$, and for $\sigma \in G$, define $\varphi(\sigma) \in Cl(R)$ by:

$$\sigma x_0 = x_0 \cdot \varphi(\sigma).$$

One checks directly that $\varphi(\sigma)$ is independent of the choice of x_0, and that φ is a homomorphism. Let L be a finite extension of K such that

(a) φ factors through $\mathrm{Gal}(L/\mathbb{Q})$;
(b) there is an elliptic curve E defined over L with j-invariant $j(\mathfrak{a})$, some proper R-ideal $\mathfrak{a}$.

LEMMA 12.25. *There is a set S of prime ideals of K of density one excluding those that ramify in L, such that*

$$\varphi(\varphi_{\mathfrak{p}}) = Cl(\mathfrak{p})$$

where $\varphi_{\mathfrak{p}} \in \mathrm{Gal}(L/K)$ is a Frobenius element.

PROOF. Let $\mathfrak{p}$ be a prime ideal of K such that

(i) $\mathfrak{p}$ is unramified in L;
(ii) E has good reduction at some prime ideal $\mathfrak{P}$ lying over $\mathfrak{p}$;
(iii) $\mathfrak{p}$ has degree 1, i.e., $\mathbb{N}(\mathfrak{p}) = p$, a prime number.

The set of such $\mathfrak{p}$ has density one (conditions (i) and (ii) exclude only finitely many primes, and it is a standard result (Math 776, VI.3.2) that the primes satisfying (iii) have density one).

To prove the equation, we have to show that

$$\varphi_{\mathfrak{p}}(E) \approx \mathfrak{p} \cdot E.$$

We can verify this after reducing mod $\mathfrak{P}$.

We have a $\mathfrak{p}$-isogeny $E \to \mathfrak{p} \cdot E$. When we reduce modulo $\mathfrak{p}$, this remains a $\mathfrak{p}$-isogeny. It is of degree $\mathbb{N}(\mathfrak{p}) = \mathfrak{p}$, and by looking at the tangent space, one sees that it is purely inseparable. Now $\varphi_{\mathfrak{p}}(E)$ reduces to $E^{(p)}$, and we can apply Proposition 12.23 to see that $E^{(p)}$ is isomorphic $\mathfrak{p} \cdot E$. □

We now prove the theorem. Since the Frobenius elements $\mathrm{Frob}_{\mathfrak{p}}$ generate $\mathrm{Gal}(L/K)$, we see that φ is surjective; whence (a) of the theorem. Part (b) is just what we proved.

The main theorem for orders. (Outline) Let R_f be an order in K. Just as for the maximal order $\mathcal{O}_K$, the ideal class group $Cl(R)$ can be identified with a quotient of the idèle class group of K, and so class field theory shows that there is an abelian extension K_f of K such that the Artin reciprocity map defines an isomorphism

$$\phi : Cl(R_f) \to \mathrm{Gal}(K_f/K).$$

Of course, when $f = 1$, K_f is the Hilbert class field. The field K_f is called the *ring class field*. Note that in general $Cl(R_f)$ is much bigger than $Cl(\mathcal{O}_K)$.

The same argument as before shows that if E_f has complex multiplication by R_f, then $K[j(E_f)]$ is the ring class field for K. Kronecker predicted (I think)[8] that K^{ab} should equal $K^* \stackrel{\mathrm{df}}{=} \mathbb{Q}^{\mathrm{cyc}} \cdot K'$, where $K' = \cup K(j(E_f))$ (union over positive integers). Note that

$$K' = \cup K(j(\tau)) \quad (\text{union over } \tau \in K, \quad \tau \in \mathbb{H}),$$

and so K^* is obtained from K by adjoining the special values $j(z)$ of j and the special values $e^{2\pi i m/n}$ of e^z.

THEOREM 12.26. *The Galois group $\mathrm{Gal}(K^{ab}/K*)$ is a product of groups of order 2.*

PROOF. Examine the kernel of the map $\mathbb{I}_K \to \mathrm{Gal}(K^*/K)$. □

[8]Actually, it is not too clear exactly what Kronecker predicted—see the articles of Schappacher.)

Points of order m. (Outline) We strengthen the main theorem to take account of the points of finite order. Fix an m, and let E be an elliptic curve over $\mathbb{C}$ with complex multiplication by $\mathcal{O}_K$. For any $\sigma \in \mathrm{Aut}(\mathbb{C})$ fixing K, there is an isogeny $\alpha : E \to \sigma E$, which we may suppose to be of degree prime to m. Then α maps E_m into σE_m, and we can choose α so that $\alpha(x) \equiv \sigma x \mod m)$ for all $x \in T_f E(\stackrel{\mathrm{df}}{=} \prod T_\ell E)$. We know that α will be an $\mathfrak{a}$-isogeny for some $\mathfrak{a}$, and under our assumptions $\mathfrak{a}$ is relatively prime to m.

Write $Id(m)$ for the set of ideals in K relatively prime to m, and $Cl(m)$ for the corresponding ideal class group. The above construction gives a homomorphism

$$\mathrm{Aut}(\mathbb{C}/K) \to Cl(m).$$

Let K_m be the abelian extension of K (given by class field theory) with Galois group $Cl(m)$.

THEOREM 12.27. *The homomorphism factors through* $\mathrm{Gal}(K_m/K)$, *and is the reciprocal of the isomorphism given by the Artin reciprocity map.*

PROOF. For $m = 1$, this is the original form of the main theorem. A similar argument works in the more general case. □

Adelic version of the main theorem. Omitted.

Index