CHAPTER II

MODULAR SYMBOL ALGORITHMS

In this chapter we describe the modular symbol method in detail. First, in Sections 2.1 to 2.5, we describe the use of modular symbols and M-symbols to compute the homology space $H_1(X_0(N), \mathbb{Q})$ and the action of the Hecke algebra, for an arbitrary positive integer N. At this stage it is already possible to identify rational newforms f, and obtain some information about the modular elliptic curves E_f attached to them: these are introduced in Section 2.6. To obtain equations for the curves E_f we compute their period lattices: the methods used for this stage occupy most of the remaining sections of the chapter. The final section 2.15 shows how to compute the degree of the associated map $\varphi : X_0(N) \to E_f$.

To illustrate the methods, we also give some worked examples in an Appendix to the chapter.

2.1 Modular Symbols and Homology

2.1.1. The upper half-plane, the modular group and cusp forms.

Let $\mathcal{H}$ denote the upper half-plane

$$\mathcal{H} = \{z = x + iy \in \mathbb{C} \mid y > 0\},$$

and $\mathcal{H}^* = \mathcal{H} \cup \mathbb{Q} \cup \{\infty\}$ the extended upper half-plane, obtained by including the cusps $\mathbb{Q} \cup \{\infty\}$. The group $\mathrm{PSL}_2(\mathbb{R})$ acts on $\mathcal{H}^*$ via linear fractional transformations:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} : \quad z \mapsto \frac{az+b}{cz+d};$$

these are the isometries of the hyperbolic geometry on $\mathcal{H}$, for which geodesics are either half-lines perpendicular to the real axis $\mathbb{R}$, or semicircles perpendicular to $\mathbb{R}$.

The modular group $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$ is a discrete subgroup of $\mathrm{PSL}_2(\mathbb{R})$ (in the topology induced from $\mathrm{SL}(2, \mathbb{R}) \subset M_2(\mathbb{R}) \cong \mathbb{R}^4$), and acts discontinuously on $\mathcal{H}$, in the sense that for each $z \in \mathcal{H}$ the orbit $\Gamma.z$ is discrete. Note that the cusps $\mathbb{Q} \cup \{\infty\}$ form a complete Γ-orbit.

The elements $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} : \quad z \mapsto -1/z$ (of order 2) and $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} : \quad z \mapsto z + 1$ (of infinite order) generate Γ. This fact, and the related fact that a fundamental region for the action of Γ on $\mathcal{H}$ is given by the set $\mathcal{F}$ defined by

$$\mathcal{F} = \{z = x + iy \in \mathcal{H} \mid |x| \le \frac{1}{2}, |z| \ge 1\}, \tag{2.1.1}$$

are standard and will not be proved here. Both results depend essentially on the fact that $\mathbb{Z}$ is Euclidean.

Let G be a subgroup of Γ of finite index e. Then G also acts discretely on $\mathcal{H}$. A fundamental region for G on $\mathcal{H}$ is given by $\cup M_i.\mathcal{F}$, where the M_i (for $1 \le i \le e$) are right coset representatives for G in Γ.

Let $X_G = G\backslash\mathcal{H}^*$ denote the quotient space; this may be given the structure of a compact Riemann surface. Around most points the local parameter is just z, but more care is needed about the "parabolic points" or cusps, and the "elliptic points" which have non-trivial stabilizers in the Γ-action. These elliptic points for G (if any) are in the Γ-orbits of i (stabilized by S of order 2), and of $\rho = (1+\sqrt{-3})/2$ (stabilized by TS of order 3). See the books of Lang [32], Shimura [55] or Knapp [28] for details of the Riemann surface construction.

Let g denote the genus of the surface X_G; as a real manifold[1], X_G is a g-holed torus. We will be concerned with the explicit computation of the 1-homology $H_1(X_G, \mathbb{Z})$, which is a free $\mathbb{Z}$-module of rank $2g$. (See Subsection 2.1.2 below for a brief review of homology). This homology will be expressed in terms of "modular symbols", defined below. We must also explain the connection between homology, modular forms, and elliptic curves. First we review the definition of cusp forms.

The space of *cusp forms of weight* 2 for G will be denoted by $S_2(G)$. These cusp forms are holomorphic functions $f(z)$ for $z \in \mathcal{H}$ which satisfy

(1) $f\,|M = f$ for all $M \in G$, where

$$\left(f\left|\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right.\right)(z) = (cz+d)^{-2} f\left(\frac{az+b}{cz+d}\right).$$

Thus, since $(cz+d)^{-2} = (d/dz)(M(z))$ for $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, we have, for all $M \in G$,

$$f(M(z))d(M(z)) = f(z)dz.$$

(2) $f(z)$ behaves nicely at the cusps. The significance of this condition is that, by (1), a cusp form of weight 2 for G is the pull-back of a (holomorphic) differential on the Riemann surface $G\backslash\mathcal{H}$, of which X_G is the compactification after adding the (finitely many) G-inequivalent cusps, and we want this differential to be holomorphic on the whole of X_G. In future we will identify cusp forms of weight 2 for G with holomorphic differentials on X_G. From standard Riemann surface theory, we then know that $S_2(G)$ is a complex vector space of dimension g.

We can make explicit the condition that $f(z)dz$ is holomorphic at the cusp ∞ (for the other cusps, see one of the references on the theory of modular forms). The stabilizer of ∞ in Γ is the infinite cyclic subgroup generated by T; if h is the least positive integer such that $T^h \in G$, then clearly we have

$$\mathrm{Stab}(\infty) \cap G = \left\langle T^h \right\rangle,$$

and every $f \in S_2(G)$ has a Fourier expansion of the form

$$f(z) = \sum_{n=1}^{\infty} a_n e^{2\pi i n z/h} \tag{2.1.2}$$

with coefficients $a_n \in \mathbb{C}$. The integer h is called the *width* of the cusp ∞; for $G = \Gamma_0(N)$, the case we will be most interested in, we have $h = 1$, since $T \in \Gamma_0(N)$.

[1]Strictly speaking, X_G is not a manifold unless G has no elements of finite order, because of the branching over the elliptic points. However this will make no difference in practice and we may safely ignore it.

2.1.2. The duality between cusp forms and homology.

The basis for our method is the explicit computation of the homology (specifically, the rational 1-homology) of the Riemann surface X_G. This is useful for various reasons. On the one hand, this gives us a very explicit vector space on which Hecke operators act, which is isomorphic (or more strictly, dual) to the space of cusp forms. Thus by computing homology and the Hecke action on it, we are indirectly also able to obtain information about the space of cusp forms. The Fourier coefficients of the cusp forms are determined by their Hecke eigenvalues (see Section 2.6), so we obtain these indirectly as eigenvalues of Hecke operators acting on homology. Secondly, in order to actually compute the elliptic curves attached to these cusp forms, we need to know their periods, which are obtained by integrating the corresponding differentials around closed paths on the surface X_G; since two paths give the same period (for all forms) if and only if they are homologous (essentially by Cauchy's Theorem on X_G), it is clear that to determine the whole period lattice we will also require an explicit knowledge of the homology of X_G.

The integral homology $H_1(X_G, \mathbb{Z})$ is most easily defined geometrically: it is the abelian group obtained by taking as generators all closed paths on X_G, and factoring out by the relation that two closed paths are equivalent (or *homologous*) if one can be continuously deformed into the other. If the genus of the surface X_G is g, this gives a free abelian group of rank $2g$: roughly speaking, the surface is a g-holed torus, and there are two generating loops around each hole. To determine this homology group in practice, one triangulates the surface, so that every path is homologous to a path along the edges of the triangulation. Now the generators are the directed edges of the triangulation, modulo relations given by the sum of the edges around each triangle being homologous to zero. A typical element of $H_1(X_G, \mathbb{Z})$ will then be given as a $\mathbb{Z}$-linear combination of these directed edges. In Subsection 2.1.6 below, we will make this very explicit: there will be one edge of the triangulation for each coset of G in Γ, and the triangle relations will be expressed algebraically in terms of the coset action of Γ. This description will entirely algebraicize the situation, in a way which is then easy to implement on a computer.

For any other ring R, the homology with coefficients in R is obtained simply by tensoring with R:

$$H_1(X_G, R) = H_1(X_G, \mathbb{Z}) \otimes_{\mathbb{Z}} R.$$

Explicitly, one just takes R-linear combinations of the $2g$ generators of the $\mathbb{Z}$-module $H_1(X_G, \mathbb{Z})$ ("extension of scalars"); the result is then an R-module. In what follows we will only need to take $R = \mathbb{Q}$, $R = \mathbb{R}$, and $R = \mathbb{C}$.

Let $H_1(X_G, \mathbb{R}) = H_1(X_G, \mathbb{Z}) \otimes_{\mathbb{Z}} \mathbb{R}$, which is a real vector space of dimension $2g$. Abstractly, this space is obtained by formal extension of scalars from $H_1(X_G, \mathbb{Z})$; but we can be more concrete, if we introduce the notion of modular symbols.

First let α, $\beta \in \mathcal{H}^*$ be points equivalent under the action of G, so that $\beta = M(\alpha)$ for some $M \in G$. Any smooth path (for instance, a geodesic path) from α to β in $\mathcal{H}^*$ projects to a closed path in the quotient space X_G, and hence determines an integral homology class in $H_1(X_G, \mathbb{Z})$, which depends only on α and β and not on the path chosen, because $\mathcal{H}^*$ is simply connected. (In fact, the class depends only on M: see (5) in Proposition 2.1.1 below). We denote this homology class by the *modular symbol* $\{\alpha, \beta\}_G$, or simply $\{\alpha, \beta\}$ when the group G is clear from the context.

Conversely, every integral homology class $\gamma \in H_1(X_G, \mathbb{Z})$ can be represented by such a modular symbol $\{\alpha, \beta\}$. Also, if $f \in S_2(G)$ then the integral

$$\int_\gamma 2\pi i f(z) dz = 2\pi i \int_\alpha^\beta f(z) dz$$

is well-defined, since $f(z)$ is holomorphic, and will be denoted either as $\langle \gamma, f \rangle$ or as $I_f(\alpha, \beta)$. The (complex) value of such an integral is called a *period* of the cusp form f, or of the associated differential $2\pi i f(z)dz$.

Let $f_1, f_2, \ldots, f_g$ be a fixed basis for $S_2(G)$, so that the differentials $2\pi i f_j(z)dz$ are a basis for the holomorphic differentials on X_G. Also let $\gamma_1, \gamma_2, \ldots, \gamma_{2g}$ be a fixed $\mathbb{Z}$-basis for the integral homology $H_1(X_G, \mathbb{Z})$. Then we may form the $2g \times g$ complex *period matrix*

$$\Omega = (\omega_{jk}) = (\langle \gamma_j, f_k \rangle).$$

By standard Riemann surface theory, the $2g$ rows of Ω are linearly independent over $\mathbb{R}$, and so their $\mathbb{Z}$-span is a lattice (discrete subgroup) Λ of rank $2g$ in $\mathbb{C}^g$. The quotient $J(G) = \mathbb{C}^g/\Lambda$ is the Jacobian of X_G; it is an abelian variety of dimension g.

The symbols $\{\alpha, \beta\}$ give $\mathbb{C}$-linear functionals $S_2(G) \to \mathbb{C}$ via $f \mapsto I_f(\alpha, \beta)$. We may identify $H_1(X_G, \mathbb{R})$ with the space of all $\mathbb{C}$-linear functionals on $S_2(G)$ as follows: given an element $\gamma \in H_1(X_G, \mathbb{R})$, we can write γ uniquely in the form

$$\gamma = \sum_{j=1}^{2g} c_j \gamma_j$$

with coefficients $c_j \in \mathbb{R}$. Define $\langle \gamma, f \rangle = \sum_{j=1}^{2g} c_j \langle \gamma_j, f \rangle$. Then the corresponding functional is $f \mapsto \langle \gamma, f \rangle$. Conversely, given a functional $\omega\colon S_2(G) \to \mathbb{C}$, the vector $(\omega(f_1), \omega(f_2), \ldots, \omega(f_g)) \in \mathbb{C}^g$ may be expressed uniquely as an $\mathbb{R}$-linear combination of the rows of Ω, so there exist real scalars c_j $(1 \le j \le 2g)$ such that

$$\omega(f) = \sum_{j=1}^{2g} c_j \langle \gamma_j, f \rangle$$

for all $f \in S_2(G)$; then $\omega(f) = \langle \gamma, f \rangle$ where $\gamma = \sum_{j=1}^{2g} c_j \gamma_j \in H_1(X_G, \mathbb{R})$.

In particular, let α, $\beta \in \mathcal{H}^*$ be arbitrary (not necessarily in the same G-orbit); then the functional $f \mapsto I_f(\alpha, \beta)$ corresponds to a unique element $\gamma = \sum_{j=1}^{2g} c_j \gamma_j \in H_1(X_G, \mathbb{R})$, and we *define* the modular symbol $\{\alpha, \beta\}_G \in H_1(X_G, \mathbb{R})$ to be this element. Clearly this definition agrees with the earlier one in the special case where $\beta = M(\alpha)$ for some $M \in G$; indeed, this case holds if and only if all $c_j \in \mathbb{Z}$.

By the *field of definition* of an element $\gamma \in H_1(X_G, \mathbb{R})$ we mean the field generated over $\mathbb{Q}$ by its coefficients c_j (with respect to the $\mathbb{Z}$-basis for the integral homology, as above). For example, γ is rational (has field of definition $\mathbb{Q}$) if and only if $\gamma \in H_1(X_G, \mathbb{Q})$.

We now have an $\mathbb{R}$-bilinear pairing

$$S_2(G) \times H_1(X_G, \mathbb{R}) \longrightarrow \mathbb{C} \tag{2.1.3}$$

given by

$$(f, \gamma) \mapsto \langle \gamma, f \rangle = \int_\gamma 2\pi i f(z)dz$$

which gives an exact duality between the two spaces on the left if we view $S_2(G)$ as a real vector space of dimension $2g$ by restriction of scalars from $\mathbb{C}$ to $\mathbb{R}$.

To interpret this as a duality over $\mathbb{C}$, we can give $H_1(X_G, \mathbb{R})$ the structure of a vector space over $\mathbb{C}$ (of dimension g) as follows. Given $\gamma \in H_1(X_G, \mathbb{R})$ and $c \in \mathbb{C}$, we define $c\gamma$ to be that element of $H_1(X_G, \mathbb{R})$ which satisfies $\langle c\gamma, f \rangle = \langle \gamma, cf \rangle$ for all $f \in S_2(G)$; in other words, $c\gamma$ is the element corresponding to the functional $f \mapsto c \langle \gamma, f \rangle$. Now the map $(f, \gamma) \mapsto \langle \gamma, f \rangle$ is $\mathbb{C}$-bilinear, and the dual pairing (2.1.3) between homology and cusp forms is an exact duality over $\mathbb{C}$.

2.1.3. Real structure.

For suitable groups G we can restrict the duality described above to a duality between real vector spaces of dimension g. This has important implications for explicit computations, where a halving of the dimension (from $2g$ to g) gives a significant saving of effort.

Let $J = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$. We say that a subgroup G of Γ is *of real type* if J normalizes G. Explicitly, let $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in G$; then $J^{-1}MJ = \begin{pmatrix} a & -b \\ -c & d \end{pmatrix} = M^*$, say, and G is of real type when $M \in G \iff M^* \in G$. This will be true, in particular, for the congruence subgroups $\Gamma_0(N)$ and $\Gamma_1(N)$ of most interest to us.

For $z \in \mathcal{H}$ set $z^* = -\overline{z}$. Then a trivial calculation shows that $w = M(z) \iff w^* = M^*(z^*)$; it follows that, for G of real type, the map $z \mapsto z^*$ induces a well-defined map on the quotient X_G, and hence also on homology, via $\{\alpha, \beta\} \mapsto \{\alpha^*, \beta^*\}$. Clearly this is an $\mathbb{R}$-linear involution on $H_1(X_G, \mathbb{R})$. Hence we obtain a decomposition into $+1$ and -1 eigenspaces for $*$:

$$H_1(X_G, \mathbb{R}) = H_1^+(X_G, \mathbb{R}) \oplus H_1^-(X_G, \mathbb{R}).$$

Remark. The involution $*$ also acts on the integral homology $H_1(X_G, \mathbb{Z})$, and we may set $H_1^\pm(X_G, \mathbb{Z}) = H_1^\pm(X_G, \mathbb{R}) \cap H_1(X_G, \mathbb{Z})$. However the direct sum $H_1^+(X_G, \mathbb{Z}) \oplus H_1^-(X_G, \mathbb{Z})$ will in general have finite index in $H_1(X_G, \mathbb{Z})$.

We now define dually an involution, also denoted $*$, on the space $S_2(G)$ where G is of real type. For a holomorphic function f on $\mathcal{H}$, we set $f^*(z) = \overline{f(z^*)}$. Then f^* is also holomorphic on $\mathcal{H}$, and the following facts are easily verified:

(1) If f has Fourier expansion $f(z) = \sum a_n q^n$ (where $q = \exp(2\pi i z/h)$ as in (2.1.2) above), then $f^*(z) = \sum \overline{a_n} q^n$. In other words, the Fourier coefficients of f^* are the conjugates of those of f.

(2) For $M \in \Gamma$, we have $f^* \mid M = (f \mid M^*)^*$.

(3) $\langle \gamma^*, f^* \rangle = \overline{\langle \gamma, f \rangle}$ for all f, γ.

As a formal consequence of fact (2), we immediately see that, for G of real type, the map $f \mapsto f^*$ is an $\mathbb{R}$-linear map from $S_2(G)$ to itself, which is an involution. Denote by $S_2(G)_\mathbb{R}$ the $\mathbb{R}$-subspace of $S_2(G)$ fixed by this involution, which by fact (1) consists of those cusp forms with real Fourier coefficients. Then $\dim_\mathbb{R}(S_2(G)_\mathbb{R}) = g$, and $S_2(G)_\mathbb{R}$ spans $S_2(G)$ over $\mathbb{C}$.

For nonzero $f \in S_2(G)_\mathbb{R}$ we have (from fact (3)):

$$\langle \gamma, f \rangle \in \mathbb{R} \iff \gamma \in H_1^+(X_G, \mathbb{R}),$$

and also

$$\langle \gamma, f \rangle \in i\mathbb{R} \iff \gamma \in H_1^-(X_G, \mathbb{R}).$$

Moreover, multiplication by i on $H_1(X_G, \mathbb{R})$ interchanges the "real" and "pure imaginary" eigenspaces $H_1^\pm(X_G, \mathbb{R})$ since

$$\begin{aligned} \gamma \in H_1^+ &\iff \langle \gamma, f \rangle \in \mathbb{R} \qquad \forall f \in S_2(G)_\mathbb{R} \\ &\iff \langle i\gamma, f \rangle \in i\mathbb{R} \qquad \forall f \in S_2(G)_\mathbb{R} \\ &\iff i\gamma \in H_1^-. \end{aligned}$$

It follows that $\dim H_1^+(X_G, \mathbb{R}) = \dim H_1^-(X_G, \mathbb{R}) = g$.

Also, since the period $\langle \gamma, f \rangle$ is real for $\gamma \in H_1^+$ and $f \in S_2(G)_\mathbb{R}$, the duality over $\mathbb{C}$ we had earlier now restricts to a duality over $\mathbb{R}$:

$$S_2(G)_\mathbb{R} \times H_1^+(X_G, \mathbb{R}) \longrightarrow \mathbb{R}. \tag{2.1.4}$$

It follows that the real vector spaces $S_2(G)_{\mathbb{R}}$ and $H_1^+(X_G, \mathbb{R})$ of dimension g are dual to each other. We will exploit this duality (which also respects the action of Hecke and other operators, see below), as we will compute $H_1(X_G, \mathbb{R})$ explicitly in order to gain information about the cusp forms in $S_2(G)$. Also, this duality is crucial in the definition of modular elliptic curves.

2.1.4. Modular symbol formalism.

We will need the following simple properties of the modular symbols $\{\alpha, \beta\}$.

PROPOSITION 2.1.1. *Let $\alpha, \beta, \gamma \in \mathcal{H}^*$, and let M, M_1, $M_2 \in G$. Then*

(1) $\{\alpha, \alpha\} = 0$;
(2) $\{\alpha, \beta\} + \{\beta, \alpha\} = 0$;
(3) $\{\alpha, \beta\} + \{\beta, \gamma\} + \{\gamma, \alpha\} = 0$;
(4) $\{M\alpha, M\beta\}_G = \{\alpha, \beta\}_G$;
(5) $\{\alpha, M\alpha\}_G = \{\beta, M\beta\}_G$;
(6) $\{\alpha, M_1M_2\alpha\}_G = \{\alpha, M_1\alpha\}_G + \{\alpha, M_2\alpha\}_G$;
(7) $\{\alpha, M\alpha\}_G \in H_1(X_G, \mathbb{Z})$.

PROOF. Only (5) and (6) are not quite obvious. For (5), write $\{\alpha, M\alpha\} = \{\alpha, \beta\} + \{\beta, M\beta\} + \{M\beta, M\alpha\}$, using (2) and (3); now the first and third terms cancel by (4). For (6), we have $\{\alpha, M_1M_2\alpha\} = \{\alpha, M_1\alpha\} + \{M_1\alpha, M_1M_2\alpha\} = \{\alpha, M_1\alpha\} + \{\alpha, M_2\alpha\}$ using (4). □

COROLLARY 2.1.2. *The map $M \mapsto \{\alpha, M\alpha\}_G$ is a surjective group homomorphism $G \to H_1(X_G, \mathbb{Z})$, which is independent of $\alpha \in \mathcal{H}^*$.*

The kernel of this homomorphism contains all commutators and elliptic elements (since the latter have finite order, and the image is a torsion-free abelian group), and also all parabolic elements: for if $M \in G$ is parabolic, it is a conjugate of T and hence fixes some $\alpha \in \mathbb{Q} \cup \{\infty\}$, so $M \mapsto \{\alpha, M\alpha\} = 0$. In fact, the kernel is generated by these elements, but we will not prove that here.

2.1.5. Rational structure and the Manin-Drinfeld Theorem.

We have seen that every element γ of $H_1(X_G, \mathbb{Z})$ has the form $\{\alpha, M\alpha\}$ with $M \in G$ and $\alpha \in \mathcal{H}^*$ arbitrary; usually we take α to be a cusp, so that γ is a path between G-equivalent cusps. It is not clear in general what is the field of definition of a modular symbol $\{\alpha, \beta\}_G$ for which α and β are both cusps. However, when G is a congruence subgroup, this is answered by the Manin-Drinfeld Theorem.

A *congruence subgroup* of Γ is a subgroup G such that membership of G is determined by means of congruence conditions on the entries of a matrix in Γ. A moment's thought shows that this is equivalent to the condition that for some positive integer N, G contains the *principal congruence subgroup* $\Gamma(N)$, which is defined to be the subgroup of Γ consisting of matrices congruent to the identity modulo N. The least such N is called the *level* of G.

The most important congruence subgroups are $\Gamma(N)$ itself;

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma \mid c \equiv 0 \pmod{N} \right\};$$

and

$$\Gamma_1(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma \mid c \equiv 0, a \equiv 1 \pmod{N} \right\}.$$

We can now state the Manin-Drinfeld Theorem.

THEOREM 2.1.3. *(Manin, Drinfeld) Let G be a congruence subgroup of the modular group Γ. Then for all pairs of cusps α, $\beta \in \mathcal{H}^*$ we have*

$$\{\alpha, \beta\}_G \in H_1(X_G, \mathbb{Q}).$$

In particular, the modular symbol $\{0, \infty\}_G$ is rational; the denominator of this element is very important in many ways.

Thus the rational homology $H_1(X_G, \mathbb{Q})$ is generated by paths between cusps (since it is generated by the integral homology), and conversely every path between cusps is rational. We will see later how to use this fact to develop an algorithm for computing the rational homology.

The proof of the Manin-Drinfeld Theorem involves the use of Hecke operators: see the Remark in Section 2.9 for a sketch of this argument in the case of $\Gamma_0(N)$. Using Hecke operators, we can also prove that the space of cusp forms $S_2(G)$ has a $\mathbb{Q}$-structure, namely a basis consisting of forms with rational Fourier coefficients (when G is a congruence subgroup). This is related to the fact that the modular curve X_G, which as a Riemann surface is certainly an algebraic curve over $\mathbb{C}$, can in fact be given the structure of an algebraic curve over the field of algebraic numbers $\overline{\mathbb{Q}}$ (and even over the Nth cyclotomic field, if G has level N). This rational structure is crucial to the construction of modular elliptic curves, to ensure that we obtain elliptic curves defined over $\mathbb{Q}$. For further details, see the books of Lang [32] and Knapp [28].

The duality between cusp forms and homology does not descend entirely to $\mathbb{Q}$, however, because even if $f \in S_2(G)$ has rational Fourier coefficients and $\gamma \in H_1(X_G, \mathbb{Q})$, the period $\langle \gamma, f \rangle$ will not be rational. Rationality questions for periods of modular forms have been studied extensively, notably by Manin, but we will not go into this further here.

2.1.6. Triangulations and homology.

From now on we will assume that G is a congruence subgroup, so that the rational homology of X_G is precisely the homology generated by paths between cusps.

We will compute the homology of X_G by first triangulating the upper half-plane $\mathcal{H}^*$, using a tessellation of hyperbolic triangles, and then passing to the quotient. This will give us a very explicit triangulation of the surface X_G, using which we can write down explicit generators and relations for its 1-homology.

For $M \in \Gamma$ let (M) denote $\{M(0), M(\infty)\}$, viewed as a path in $\mathcal{H}^*$; this is the image under M of the imaginary axis $\{0, \infty\}$. These geodesic paths form the oriented edges of a triangulation of $\mathcal{H}^*$ whose vertices are the cusps $\mathbb{Q} \cup \{\infty\}$. Explicitly, there is an edge from ∞ to n for all $n \in \mathbb{Z}$, and an edge between rational numbers a/c and b/d such that $ad - bc = 1$. The triangles of the triangulation are images under $M \in \Gamma$ of the basic triangle $\mathcal{T}$ with vertices at 0, 1 and ∞ and edges (I), (TS), $((TS)^2)$. We denote by $\langle M \rangle$ the image of this triangle under M, which has vertices $M(0)$, $M(1)$ and $M(\infty)$ and edges (M), (MTS) and $(M(TS)^2)$. This representation of the triangles is unique except for the relation

$$\langle M \rangle = \langle MTS \rangle = \left\langle M(TS)^2 \right\rangle.$$

Also, triangles $\langle M \rangle$ and $\langle MS \rangle$ meet along the edge (M), since $(MS) = -(M)$ (the negative sign indicating reverse orientation).

We will use the symbol $(M)_G$ to denote the image of the path (M) in the quotient X_G, and also its image in the rational homology. The geometric observations of the previous paragraph now give us the following 2- and 3-term relations between these symbols:

$$\begin{aligned} (M)_G + (MTS)_G + (M(TS)^2)_G &= 0 \\ (M)_G + (MS)_G &= 0. \end{aligned} \tag{2.1.5}$$

We also clearly have the relations

$$(M'M)_G = (M)_G \tag{2.1.6}$$

for all $M' \in G$, so we may use as generators of the rational homology the finite set of symbols $(M_i)_G$, $(1 \le i \le e)$, where as before $M_1, \ldots, M_e$ are a set of right coset representatives for G in Γ.

Let $C(G)$ be the $\mathbb{Q}$-vector space with basis the formal symbols $(M)_G$ for each M in Γ, identified by the relations (2.1.6), so that $\dim(C(G)) = e = [\Gamma : G]$.

Let $B(G)$ be the subspace of $C(G)$ spanned by all elements of the form

$$(M)_G + (MS)_G,$$
$$(M)_G + (MTS)_G + (M(TS)^2)_G.$$

Let $C_0(G)$ be the $\mathbb{Q}$-vector space spanned by the G-cusps $[\alpha]_G$ for $\alpha \in \mathbb{Q} \cup \{\infty\}$ (so that $[\alpha]_G = [\beta]_G \iff \beta = M(\alpha)$ for some $M \in G$). Define the boundary map $\delta\colon C(G) \to C_0(G)$ by

$$\delta((M)_G) = [M(\infty)]_G - [M(0)]_G$$

and set $Z(G) = \ker(\delta)$. Note that $B(G) \subseteq Z(G)$, by a trivial calculation using the facts that S transposes 0 and ∞ while TS cycles 0, 1 and ∞.

Finally we define $H(G) = Z(G)/B(G)$. The crucial result, due in this form to Manin [37, Theorem 1.9], is that this formal construction does in fact give us the rational homology of X_G:

THEOREM 2.1.4. *$H(G)$ is isomorphic to $H_1(X_G, \mathbb{Q})$, the isomorphism being induced by*

$$(M)_G \mapsto \{M(0), M(\infty)\}_G.$$

We may thus use the symbol $(M)_G$ either as an abstract symbol obeying certain relations, or to denote an element of $H_1(X_G, \mathbb{Q})$, without confusion. In future, as the subgroup G will be fixed, we will omit the subscript on these symbols and blur the distinction between (M) as a path in the upper half-plane and $(M)_G$ as representing an element of the rational 1-homology of X_G.

Note that the form of the relations between the generating symbols (M) does not depend at all on the specific group G. In particular we do not have to consider explicitly the shape of a fundamental region for the action of G on $\mathcal{H}^*$, or how the edges of such a region are identified. This represents a major simplification compared with earlier approaches, such as that used by Tingley [67]. In order to develop this result into an explicit algorithm for computing homology, we need to have a specific set of right coset representatives for the subgroup G of Γ, and also to have a test for G-equivalence of cusps. These are purely algebraic problems which can easily be solved for arithmetically defined subgroups G such as congruence subgroups. Observe that from this point on, we do not have to do any geometry at all.

One final remark before we specialize to the case $G = \Gamma_0(N)$: every path between cusps may be expressed as a finite sum of paths of the form (M) with $M \in \Gamma$. Writing $\{\alpha, \beta\} = \{0, \beta\} - \{0, \alpha\}$, it suffices to do this for modular symbols of the form $\{0, \alpha\}$. Let

$$\frac{p_{-2}}{q_{-2}} = \frac{0}{1}, \frac{p_{-1}}{q_{-1}} = \frac{1}{0}, \frac{p_0}{1} = \frac{p_0}{q_0}, \frac{p_1}{q_1}, \frac{p_2}{q_2}, \ldots, \frac{p_k}{q_k} = \alpha \tag{2.1.7}$$

denote the continued fraction convergents of the rational number α. Then, as is well-known,

$$p_j q_{j-1} - p_{j-1} q_j = (-1)^{j-1} \qquad \text{for } -1 \le j \le k.$$

Hence

$$\{0,\alpha\} = \sum_{j=-1}^{k} \left\{ \frac{p_{j-1}}{q_{j-1}}, \frac{p_j}{q_j} \right\} = \sum_{j=-1}^{k} \{M_j(0), M_j(\infty)\} = \sum_{j=-1}^{k} (M_j) \tag{2.1.8}$$

where $M_j = \begin{pmatrix} (-1)^{j-1}p_j & p_{j-1} \\ (-1)^{j-1}q_j & q_{j-1} \end{pmatrix}$.

2.2 M-symbols and $\Gamma_0(N)$

We now specialize to the case $G = \Gamma_0(N)$:

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma \mid c \equiv 0 \pmod{N} \right\}.$$

The index of $\Gamma_0(N)$ in Γ is given (see [55, Proposition 1.43]) by

$$[\Gamma : \Gamma_0(N)] = N \prod_{p \mid N} \left(1 + p^{-1}\right).$$

Define $H(N) = H(\Gamma_0(N))$ and $X_0(N) = X_{\Gamma_0(N)}$. After Theorem 2.1.4, we will identify $H(N)$ with $H_1(X_0(N), \mathbb{Q})$ by identifying (M) with $\{M(0), M(\infty)\}$.

The next lemma is used to determine right coset representatives for $\Gamma_0(N)$ in Γ.

PROPOSITION 2.2.1. *For $j = 1, 2$ let $M_j = \begin{pmatrix} a_j & b_j \\ c_j & d_j \end{pmatrix} \in \Gamma$. The following are equivalent.*

(1) *The right cosets $\Gamma_0(N)M_1$ and $\Gamma_0(N)M_2$ are equal;*
(2) *$c_1 d_2 \equiv c_2 d_1 \pmod{N}$;*
(3) *There exists u with $\gcd(u, N) = 1$ such that $c_1 \equiv u c_2$ and $d_1 \equiv u d_2 \pmod{N}$.*

PROOF. We have

$$M_1 M_2^{-1} = \begin{pmatrix} a_1 d_2 - b_1 c_2 & * \\ c_1 d_2 - d_1 c_2 & a_2 d_1 - b_2 c_1 \end{pmatrix},$$

which is in $\Gamma_0(N)$ if and only if $c_1 d_2 - d_1 c_2 \equiv 0 \pmod{N}$. Thus (1) and (2) are equivalent. Also, if (1) holds, then from $\det(M_1 M_2^{-1}) = 1$, we deduce also that $\gcd(u, N) = 1$, where $u = a_2 d_1 - b_2 c_1$. Now

$$\begin{aligned} u c_2 &= a_2 d_1 c_2 - b_2 c_1 c_2 \\ &\equiv a_2 d_2 c_1 - b_2 c_2 c_1 && \text{since } d_1 c_2 \equiv d_2 c_1 \pmod{N} \\ &= c_1 && \text{since } a_2 d_2 - b_2 c_2 = 1 \end{aligned}$$

and $u d_2 \equiv d_1$ similarly. Conversely, if $c_1 \equiv u c_2$ and $d_1 \equiv u d_2 \pmod{N}$ with $\gcd(u, N) = 1$, then the congruence in (2) follows easily. □

On the set of ordered pairs $(c, d) \in \mathbb{Z}^2$ such that $\gcd(c, d, N) = 1$ we now define the relation $\sim$, where

$$(c_1, d_1) \sim (c_2, d_2) \iff c_1 d_2 \equiv c_2 d_1 \pmod{N}. \tag{2.2.1}$$

By Proposition 2.2.1, this is an equivalence relation. The equivalence class of (c, d) will be denoted $(c : d)$, and such symbols will be called M-symbols (after Manin, who introduced them in [37]). The set of these M-symbols modulo N is $P^1(N) = P^1(\mathbb{Z}/N\mathbb{Z})$, the projective line over the ring of integers modulo N.

Notice that in an M-symbol $(c : d)$, the integers c and d are only determined modulo N, and that we can always choose them such that $\gcd(c, d) = 1$.

Proposition 2.2.1 now implies the following.

PROPOSITION 2.2.2. *There exist bijections*

$$P^1(N) \longleftrightarrow [\Gamma : \Gamma_0(N)] \longleftrightarrow \{(M) : M \in [\Gamma : \Gamma_0(N)]\}$$

given by

$$(c : d) \leftrightarrow M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \leftrightarrow (M) = \{b/d, a/c\} \tag{2.2.2}$$

where $a, b \in \mathbb{Z}$ are chosen so that $ad - bc = 1$. $\square$

Note that a different choice of a, b in (2.2.2) has the effect of multiplying M on the left by a power of T which does not change the right coset of M, or the symbol (M), since $T \in \Gamma_0(N)$ for all N.

The right coset action of Γ on $[\Gamma : \Gamma_0(N)]$ induces an action on $P^1(N)$:

$$(c : d) \begin{pmatrix} p & q \\ r & s \end{pmatrix} = (cp + dr : cq + ds). \tag{2.2.3}$$

In particular, we have

$$(c : d)S = (d : -c) \qquad \text{and} \qquad (c : d)T = (c : c + d). \tag{2.2.4}$$

The boundary map δ now takes the form

$$\delta\colon \quad (c : d) \mapsto [a/c] - [b/d]. \tag{2.2.5}$$

In order to compute $\ker(\delta)$, we must be able to determine when two cusps are $\Gamma_0(N)$-equivalent. This is achieved by the following result.

PROPOSITION 2.2.3. *For $j = 1, 2$ let $\alpha_j = p_j/q_j$ be cusps written in lowest terms. The following are equivalent:*

(1) *$\alpha_2 = M(\alpha_1)$ for some $M \in \Gamma_0(N)$;*
(2) *$q_2 \equiv uq_1 \pmod{N}$ and $up_2 \equiv p_1 \pmod{\gcd(q_1, N)}$, with $\gcd(u, N) = 1$.*
(3) *$s_1 q_2 \equiv s_2 q_1 \pmod{\gcd(q_1 q_2, N)}$, where s_j satisfies $p_j s_j \equiv 1 \pmod{q_j}$.*

PROOF. (1) $\implies$ (2): Let $M = \begin{pmatrix} a & b \\ Nc & d \end{pmatrix} \in \Gamma_0(N)$; Then $p_2/q_2 = (ap_1 + bq_1)/(Ncp_1 + dq_1)$, with both fractions in lowest terms. Equating numerators and denominators (up to sign) gives (2), with $u = \pm d$, since $ad \equiv 1 \pmod{N}$.

(2) $\implies$ (1): Here we use Proposition 2.2.1. Assume (2), and write $p_1 s_1' - q_1 r_1' = p_2 s_2 - q_2 r_2 = 1$ with $s_1', r_1', s_2, r_2 \in \mathbb{Z}$. Then $p_1 s_1' \equiv 1 \pmod{q_1}$ and $p_2 s_2 \equiv 1 \pmod{q_2}$. Also $\gcd(q_1, N) = \gcd(q_2, N) = N_0$, say, since $q_2 \equiv uq_1 \pmod{N}$. Now $up_2 \equiv p_1 \pmod{N_0}$ implies $us_1' \equiv s_2 \pmod{N_0}$, so we may find $x \in \mathbb{Z}$ such that $uxq_1 \equiv us_1' - s_2 \pmod{N}$. Set $s_1 = s_1' - xq_1$ and $r_1 = r_1' - xp_1$. Then $p_1 s_1 - q_1 r_1 = 1$ and now $us_1 \equiv s_2 \pmod{N}$. By Proposition 2.2.1, there exists $M \in \Gamma_0(N)$ such that $\begin{pmatrix} p_2 & r_2 \\ q_2 & s_2 \end{pmatrix} = M \begin{pmatrix} p_1 & r_1 \\ q_1 & s_1 \end{pmatrix}$, and so $M(p_1/q_1) = p_2/q_2$ as required.

(1) $\iff$ (3): As before, solve the equations $p_j s_j - q_j r_j = 1$ for $j = 1, 2$. Set $M_j = \begin{pmatrix} p_j & r_j \\ q_j & s_j \end{pmatrix}$, so that $M_j(\infty) = \alpha_j$, and $M_2 M_1^{-1}(\alpha_1) = \alpha_2$. This matrix is in $\Gamma_0(N)$ if and only if $q_2 s_1 - q_1 s_2 \equiv 0 \pmod{N}$. The most general such matrix is obtained by replacing s_1 by

$s_1' = s_1 + xq_1$, and it follows that α_1 and α_2 are equivalent if and only if we can solve for $x \in \mathbb{Z}$ the congruence

$$0 \equiv q_2 s_1' - q_1 s_2 \equiv q_2 s_1 - q_1 s_2 + x q_1 q_2 \pmod{N},$$

which is if and only if the congruence in (3) holds. □

Henceforth, we can therefore assume that $H(N)$ is given explicitly in terms of M-symbols. Certain symbols will be generators, and each M-symbol $(c : d)$ will be expressed as a $\mathbb{Q}$-linear combination of these generating symbols, by means of the 2-term relations

$$(c : d) + (-d : c) = 0 \tag{2.2.6}$$

and 3-term relations

$$(c : d) + (c + d : -c) + (d : -c - d) = 0. \tag{2.2.7}$$

These are just the relations (2.1.5) expressed in terms of M-symbols, using (2.2.4).

Implementation. We make a list of inequivalent M-symbols as follows: first, list the symbols $(c : 1)$ for $0 \le c < N$; then the symbols $(1 : d)$ for $0 \le d < N$ and $\gcd(d, N) > 1$; and finally a pairwise inequivalent set of symbols $(c : d)$ with $c|N$, $c \neq 1, N$, $\gcd(c, d) = 1$ and $\gcd(d, N) > 1$. (The latter symbols do not arise when N is a prime power.)

To speed up the looking up of M-symbols in the list, we have found it extremely worthwhile to prepare at the start of the program a collection of lookup tables, containing for example a table of inverses modulo N. We also used a simple "hashing" system, so that given any particular symbol $(c : d)$ we could quickly determine to which symbol in our standard list it is equivalent. While this preparation of look-up tables may seem rather trivial, in practice it has had a dramatic effect, speeding up the mass computation of Hecke eigenvalues a_p (see Section 2.9) by a factor of up to 50.

Using the 2-term relations (2.2.6) we may identify the M-symbols in pairs, up to sign. This immediately halves the number of generators needed. Then the 3-term relations (2.2.7) are computed, each M-symbol being replaced by plus or minus one of the current generators, and the resulting equations solved using integer Gaussian elimination. At the end of this stage we have a list of r (say) "free generators" from the list of M-symbols, and a table expressing each of the M-symbols in the list as a $\mathbb{Q}$-linear combination of the generators. In practice, we store $\mathbb{Z}$-linear combinations, keeping a common denominator d_1 separately; however, by judicious choice of the order of elimination of symbols, in practice this denominator is very frequently 1.

Next we compute the boundary map δ on each of the free generators, using (2.2.5). We have a procedure based on Proposition 2.2.3 to test cusp equivalence. Hence we do not have to compute in advance a complete list of inequivalent cusps. Instead, we keep a cumulative list: each cusp we come across is checked for equivalence with those in the list already, and is added to the list if it represents a new equivalence class. We found this simpler to implement than using a standard set of pairwise inequivalent cusps, as in [37, Corollary 2.6].

We thus compute a matrix with integer entries for the linear map δ, and by a second step of Gaussian elimination can compute a basis for its kernel, which by definition is $H(N)$. This basis is stored as a list of $2g$ integer vectors in $\mathbb{Z}^r$ over a second common denominator d_2. (Here g is the genus of $X_0(N)$, so that $\dim H(N) = 2g$.) We may arrange (by reducing the basis suitably) that whenever a linear combination of M-symbols (represented as a vector in $\mathbb{Z}^r$) is in $\ker(\delta)$, then its coefficients with respect to the basis are given by (a subset of) $2g$ components of these vectors, divided by the cumulative common denominator $d_1 d_2$.

From now on we will regard elements of $H(N)$ as being given by vectors in $\mathbb{Z}^{2g}$ in this way.

2.3 Conversion between modular symbols and M-symbols

As noted above, each M-symbol $(c : d)$ has a representative with $\gcd(c, d) = 1$, and corresponds to the right coset representative $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ in Γ, where $ad - bc = 1$. The isomorphism of Theorem 2.1.4 thus becomes

$$(c : d) \mapsto \{b/d, a/c\}. \tag{2.3.1}$$

The modular symbol on the right of (2.3.1) is independent of the choice of a and b with $ad - bc = 1$, since $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in \Gamma_0(N)$ for all N, and so

$$\{\alpha, \beta\}_{\Gamma_0(N)} = \{\alpha + k, \beta + l\}_{\Gamma_0(N)}$$

for all k, l in $\mathbb{Z}$ and α, β in $\mathcal{H}^*$.

Conversely each modular symbol $\{\alpha, \beta\}$ with α and β in $\mathbb{Q} \cup \{\infty\}$ can be expressed, using continued fractions, as a sum of modular symbols of the special form $(M) = \{M(0), M(\infty)\}$ with $M \in \Gamma$, hence as a sum of M-symbols $(c : d)$, and finally as a linear combination of the generating M-symbols.

Using the notation introduced above in Subsection 2.1.6, if $q_0 = 1$, q_1, $\ldots$, q_k are the denominators of the continued fraction convergents to the rational number α as in (2.1.7), in terms of M-symbols we have

$$\{0, \alpha\} = \sum_{j=1}^{k} ((-1)^{j-1} q_j : q_{j-1}) \tag{2.3.2}$$

since the first two terms in (2.1.8) cancel out. Note that it is only the denominators of the continued fraction convergents which are used.

2.4 Action of Hecke and other operators

For each prime p not dividing N, the Hecke operator T_p acts on modular symbols $\{\alpha, \beta\}$ via

$$\begin{aligned} T_p\colon \quad \{\alpha, \beta\} &\mapsto \left[\begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix} + \sum_{r \bmod p} \begin{pmatrix} 1 & r \\ 0 & p \end{pmatrix} \right] \{\alpha, \beta\} \\ &= \{p\alpha, p\beta\} + \sum_{r \bmod p} \left\{ \frac{\alpha + r}{p}, \frac{\beta + r}{p} \right\}. \end{aligned} \tag{2.4.1}$$

This action induces a linear map from $H(N)$ to itself, provided that p does not divide N, which we again denote by T_p.

There are also Hecke operators, which we also denote T_p, acting on the space $S_2(N) = S_2(\Gamma_0(N))$ of cusp forms of weight 2 for $\Gamma_0(N)$. First recall that 2×2 matrices $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with $ad - bc > 0$ act on functions $f(z)$ on the right via

$$f \mapsto f\,|M \qquad \text{where} \qquad (f\,|M)(z) = \frac{ad - bc}{(cz + d)^2} f\left(\frac{az + b}{cz + d} \right).$$

A form of weight 2 for some group G will satisfy $f\,|M = f$ for all $M \in G$. This action extends by linearity to an action by formal linear combinations of matrices. The Hecke operator T_p is defined by

$$f\,|T_p = f\left|\left[\begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix} + \sum_{r=0}^{p-1} \begin{pmatrix} 1 & r \\ 0 & p \end{pmatrix}\right],\right.$$

so that

$$(f\,|T_p)\,(z) = p\,f(pz) + \frac{1}{p}\sum_{r=0}^{p-1} f\left(\frac{z+r}{p}\right).$$

A standard result is that T_p does act on $S_2(N)$, provided that $p \nmid N$. (There are similar operators U_p for primes p dividing N, but we will not need these).

These matrix actions on $S_2(N)$ and $H(N)$ are compatible, in the sense that they respect the duality between cusp forms and homology:

$$\langle \{\alpha, \beta\}, f\,|M \rangle = \langle \{M\alpha, M\beta\}, f \rangle,$$

since

$$\frac{d}{dz}\left(\frac{az+b}{cz+d}\right) = \frac{ad-bc}{(cz+d)^2},$$

and so

$$\int_\alpha^\beta (f\,|M)\,(z)dz = \int_\alpha^\beta \frac{ad-bc}{(cz+d)^2} f(M(z))dz = \int_{M\alpha}^{M\beta} f(w)dw.$$

Thus, in particular,

$$\langle \{\alpha, \beta\}, f\,|T_p \rangle = \langle T_p\{\alpha, \beta\}, f \rangle.$$

Secondly, for each prime q dividing N there is an involution operator W_q acting on $H(N)$ and $S_2(N)$. We recall the definition. Let q^α be the exact power of q dividing N, and let x, y, z, w be any integers satisfying $q^\alpha xw - (N/q^\alpha)yz = 1$. Then the matrix $W_q = \begin{pmatrix} q^\alpha x & y \\ Nz & q^\alpha w \end{pmatrix}$ has determinant q^α and normalizes $\Gamma_0(N)$ (modular scalar matrices). Thus W_q induces an action on $H(N)$ and $S_2(N)$, which is an involution since $W_q^2 \in \Gamma_0(N)$ (modulo scalars), and is independent of the values x, y, z, w chosen. The product of all the W_q for q dividing N is the Fricke involution W_N, coming from the transformation $z \mapsto -1/Nz$, with matrix $\begin{pmatrix} 0 & -1 \\ N & 0 \end{pmatrix}$.

The operators T_p for primes p not dividing N and W_q for primes q dividing N together generate a commutative $\mathbb{Q}$-algebra, called the Hecke algebra and denoted $\mathbb{T}$. Moreover, each operator is self-adjoint with respect to the so-called Petersson inner product on $S_2(N)$, and so there exist bases for $S_2(N)$ consisting of simultaneous eigenforms for all the T_p and W_q, with real eigenvalues. (See [1, Theorem 2] or [32, Corollary 2 to Theorem 4.2].) Similarly, the action of $\mathbb{T}$ on $H(N) \otimes \mathbb{R}$ can also be diagonalized.

Finally, recall from Subsection 2.1.3 that the transformation $z \mapsto z^* = -\overline{z}$ on $\mathcal{H}$ commutes with the action of $\Gamma_0(N)$ and hence also induces an involution on $H(N)$ which we denote $*$. This operator commutes with all the T_p and W_q, which thus preserve the eigenspaces $H^+(N)$ and $H^-(N)$. Moreover, $H^+(N)$ and $H^-(N)$ are isomorphic as modules for the Hecke algebra $\mathbb{T}$. It follows that in order to compute eigenvalues of Hecke operators, we can restrict our attention to $H^+(N)$. This has some practical significance, as we elaborate in the next section.

Implementation. To compute the matrices giving the action of each of these operators on $H(N)$ we may proceed as follows. We convert each of the generating M-symbols to a modular symbol as in Section 2.3. To compute a T_p, we apply (2.4.1) to each, reconvert each term on the right of (2.4.1) to a sum of M-symbols using (2.3.2), and hence express it as a $\mathbb{Z}^{2g}$-vector giving it as a linear combination of the generating M-symbols. This gives one column of the $2g \times 2g$ matrix. Similarly with W_q and W_N. Computing the matrix of $*$ is easier, as we can work directly with the M-symbols, on which $*$ acts via $(c : d) \mapsto (-c : d)$. These integer matrices are in fact $d_1 d_2$ times the actual operator matrices (where d_1 and d_2 are the denominators which may have arisen earlier as a result of the Gaussian elimination steps). Obviously this must be taken into account when we look for eigenvalues later; however, for simplicity of exposition we will assume from now on that this denominator $d_1 d_2$ is 1. We use the convention that the space is represented by column vectors, with operator matrices acting on the left.

Heilbronn matrices. There is an alternative approach to computing the T_p, based on so-called *Heilbronn matrices* of level p. These were described by Mazur in [38], and their application to give an algorithm for computing the Hecke action on homology in terms of M-symbols was given by Merel in his paper [42]. We will describe our own version of this method, which is easy to implement; our approach differs slightly from, and is a little simpler than, that of Merel's paper [42].

Since with this method one acts directly on the M-symbols, one avoids the conversion to and from modular symbols. This makes the method faster in practice, particularly as we may precompute the Heilbronn matrices for all the small primes p (say $p < 30$) for which we need to compute the matrix of T_p in order to split off one-dimensional eigenspaces from $H(N)$.

From the definition in (2.4.1), the action of T_p is expressed as the sum of the actions on modular symbols, on the left, of $p+1$ matrices of determinant p, namely $\begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & r \\ 0 & p \end{pmatrix}$ for r modulo p. The following result shows how each of these acts on M-symbols directly, via an action on the right.

PROPOSITION 2.4.1. *Let p be a prime not dividing N and $(c : d)$ an M-symbol for N. The action on $(c : d)$ of the $p + 1$ matrices appearing in* (2.4.1) *is as follows.*
(1)

$$\begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix} (c : d) = (c : pd) = (c : d) \begin{pmatrix} 1 & 0 \\ 0 & p \end{pmatrix}.$$

(2) *For $r \in \mathbb{Z}$, let $M_i \in \Gamma$ for $0 \le i \le k$ be the matrices constructed from the continued fraction convergents to r/p as in* (2.1.8)*, so that*

$$M_0(0) = \infty, \quad M_1(0) = M_0(\infty), \quad \dots, \quad M_k(0) = M_{k-1}(\infty), \quad M_k(\infty) = \frac{r}{p}.$$

Set $M_i' = \begin{pmatrix} p & -r \\ 0 & 1 \end{pmatrix} M_i S$ for $0 \le i \le k$. Then

$$\begin{pmatrix} 1 & r \\ 0 & p \end{pmatrix} (c : d) = \sum_{i=0}^{k} (c : d) M_i'.$$

PROOF. In each case we first solve $ad - bc = 1$ for integers a, b and apply the appropriate matrix to the modular symbol $\{b/d, a/c\}$.

(1) Since $p \nmid N$ we may assume by the Chinese Remainder Theorem that c is a multiple of p, say $c = pc'$. Now $M = \begin{pmatrix} a & pb \\ c' & d \end{pmatrix} = \begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix}^{-1} \in \Gamma$, and

$$(c:d)\begin{pmatrix} 1 & 0 \\ 0 & p \end{pmatrix} = (c:pd) = (c':d) = (M) = \left\{\frac{pb}{d}, \frac{a}{c'}\right\} = \begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix}\left\{\frac{b}{d}, \frac{a}{c}\right\}.$$

(2) By construction of the M_i, we have

$$\sum_{i=0}^{k}(M_i) = \sum_{i=0}^{k}\{M_i(0), M_i(\infty)\} = \left\{\infty, \frac{r}{p}\right\}.$$

Given an M-symbol $(c:d)$, we will show how to choose a and b which satisfy $ad - bc = 1$ and also

$$M = \begin{pmatrix} 1 & r \\ 0 & p \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & r \\ 0 & p \end{pmatrix}^{-1} \in \Gamma.$$

Then

$$\sum_{i=0}^{k}(MM_iS) = M\left\{\frac{r}{p}, \infty\right\} = \left(M\begin{pmatrix} 1 & r \\ 0 & p \end{pmatrix}\right) = \left(\begin{pmatrix} 1 & r \\ 0 & p \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right) = \begin{pmatrix} 1 & r \\ 0 & p \end{pmatrix}\left\{\frac{b}{d}, \frac{a}{c}\right\}.$$

Now we show that suitable values of a and b exist. Replacing d by $d + N$ if necessary, we may assume that $cr \not\equiv d \pmod{p}$. Given an arbitrary solution to $ad - bc = 1$, solve for t the congruence

$$(cr - d)t \equiv (b + dr) - r(a + cr) \pmod{p}.$$

Replacing (a, b) by $(a + ct, b + dt)$ we then still have $ad - bc = 1$ and now

$$(b + dr) - r(a + cr) = pb'$$

with $b' \in \mathbb{Z}$, and a simple calculation shows that $M = \begin{pmatrix} a + rc & b' \\ pc & d - rc \end{pmatrix}$ has the desired properties.

Since the bottom row of MM_iS is $(pc : d - rc)M_iS = (c:d)\begin{pmatrix} p & -r \\ 0 & 1 \end{pmatrix} M_iS = (c:d)M_i'$, the result follows. □

Hence for each prime p there exists a finite set R_p of matrices in $M_2(\mathbb{Z})$ with determinant p, called the Heilbronn matrices of level p, such that the Hecke operator T_p acts on M-symbols via

$$(c:d) \mapsto \sum_{M \in R_p} (c:d)M.$$

The usual definition of the set R_p (for an odd prime p not dividing N) is as follows: R_p is the set of matrices $\begin{pmatrix} x & -y \\ y' & x' \end{pmatrix} \in M_2(\mathbb{Z})$ with determinant $xx' + yy' = p$, and either (i) $x > |y| > 0$, $x' > |y'| > 0$, and $yy' > 0$; or (ii) $y = 0$, and $|y'| < x'/2$; or (iii) $y' = 0$, and $|y| < x/2$. This description, while closer to the original definition by Heilbronn and used by both Mazur and Merel, is not so easy to use in practice. One can show that the matrices in this definition may be constructed using the continued fraction expansions of r/p for r modulo p, and this leads to the presentation we have given here.

For example, for the first few primes we have

$$R_2 = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix} \right\},$$

$$R_3 = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 3 \end{pmatrix}, \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 3 & -1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 1 & -3 \end{pmatrix} \right\},$$

and

$$R_5 = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}, \begin{pmatrix} 5 & 2 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 3 & 5 \end{pmatrix}, \begin{pmatrix} 5 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 5 \end{pmatrix}, \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}, \right.$$
$$\left. \begin{pmatrix} 5 & -1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 1 & -5 \end{pmatrix}, \begin{pmatrix} 5 & -2 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -2 & 1 \\ 1 & -3 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ -3 & 5 \end{pmatrix} \right\}.$$

To compute the complete set R_p for any prime p we may use the following algorithm. Effectively, we are computing the continued fraction expansions of each rational r/p with denominator p, and recording the matrices denoted M_i' in the preceding Proposition. In line 3 of the algorithm, the loop is over a complete set of residues r modulo p, such as $-p/2 \le r < p/2$.

Algorithm for computing Heilbronn matrices

```
INPUT:        p (a prime).
OUTPUT:       the Heilbronn matrices of level p.
 1.  BEGIN
 2.  OUTPUT ( 1  0 );
            ( 0  p )
 3.  FOR r MODULO p DO
 4.  BEGIN
 5.       x1=p; x2=-r; y1=0; y2=1; a=-p; b=r;
 6.       OUTPUT ( x1  x2 );
                 ( y1  y2 )
 7.       WHILE b≠0 DO
 8.       BEGIN
 9.            q=nearest_integer(a/b);
10.            c=a-b*q; a=-b; b=c;
11.            x3=q*x2-x1; x1=x2; x2=x3;
12.            y3=q*y2-y1; y1=y2; y2=y3;
13.            OUTPUT ( x1  x2 )
                      ( y1  y2 )
14.       END
15.  END
16.  END
```

For example, take $p = 7$ and $r = 3$. The continued fraction convergents linking ∞ to $3/7$ are

$$\infty, 0, \frac{1}{2}, \frac{3}{7}$$

with associated unimodular matrices

$$M_0 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad M_1 = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}, \quad \text{and} \quad M_2 = \begin{pmatrix} -3 & 1 \\ -7 & 2 \end{pmatrix}.$$

The matrices M_i' constructed in Proposition 2.4.1 are then

$$M_0' = \begin{pmatrix} 7 & -3 \\ 0 & 1 \end{pmatrix}, \quad M_1' = \begin{pmatrix} -3 & -1 \\ 1 & -2 \end{pmatrix}, \quad \text{and} \quad M_2' = \begin{pmatrix} 1 & 0 \\ 2 & 7 \end{pmatrix}.$$

These are (up to sign) the same as the matrices output by lines 3–14 of the algorithm when $p = 7$ and $r = 3$.

2.5 Working in $H^+(N)$

Recall that $H^+(N)$ is the $+1$ eigenspace for the operator $*\colon z \mapsto -\overline{z}$ acting on $H(N) = H_1(X_0(N), \mathbb{Q})$. We would like to work in $H^+(N)$ to compute the action of the Hecke algebra $\mathbb{T}$, since there are obvious savings in computation time and storage space achieved by working in a space with half the dimension of $H(N)$. To do this, note that $H^+(N) \cong H(N)/H^-(N)$ (as vector spaces). We can thus compute $H^+(N)$ in terms of M-symbols by including extra 2-term relations

$$(c : d) = (-c : d) \tag{2.5.1}$$

between the M-symbols. We must also identify the cusp equivalence classes $[\alpha]$ and $[\alpha^*] = [-\alpha]$ for $\alpha \in \mathbb{Q}$.

Effectively we are replacing $\Gamma_0(N)$ by the larger group

$$\widetilde{\Gamma_0(N)} = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mid a, b, c, d \in \mathbb{Z}, ad - bc = \pm 1, c \equiv 0 \pmod{N} \right\} = \langle \Gamma_0(N), J \rangle$$

which still acts discretely on $\mathcal{H}^*$ via

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \colon \quad z \mapsto \begin{cases} \dfrac{az + b}{cz + d} & \text{if } ad - bc = +1, \\ \dfrac{a\overline{z} + b}{c\overline{z} + d} & \text{if } ad - bc = -1; \end{cases}$$

in particular, $J = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$ sends z to $z^* = -\overline{z}$, giving the action of $*$. Hence, in effect, $\widetilde{\Gamma_0(N)} = \langle \Gamma_0(N), * \rangle$, and $H^+(N) \cong H_1(\widetilde{\Gamma_0(N)} \backslash \mathcal{H}^*, \mathbb{Q})$. (A similar procedure is possible for other subgroups G of Γ of real type.)

As a further saving, use of the extra relation (2.5.1) enables us to cut out half the 3-term relations (2.2.7), as follows. Using (2.5.1) on the second and third terms of (2.2.7) yields

$$(c : d) + (c + d : c) + (d : c + d) = 0.$$

Also, (2.5.1) and (2.2.6) together imply

$$(d : c) = -(c : d).$$

Hence relation (2.2.7) for $(d : c)$ now gives the same information as (2.2.7) for $(c : d)$, and can be omitted. Geometrically, the triangles which determine the 3-term relations have been identified in pairs by the action of the larger group, since the effect of the transformation J is to fold the upper half-plane in two along the imaginary half-axis.

Implementation. We modify the procedure of Section 2.2 in three ways: taking the 2-term relations (2.2.6) and (2.5.1) together we may identify M-symbols in sets of four (instead of two), up to sign, at the first stage of elimination. Then in the second stage we have only half the number of 3-term relations to consider, as noted above, and each can be expressed in terms of half the number of current generators: so we have half the number of equations in half the number of variables to solve, giving a four-fold saving in space and time. Finally, in computing $\ker(\delta)$ we must use a wider notion of cusp equivalence, since for $\alpha, \beta \in \mathbb{Q}$,

$$\alpha \equiv \beta \pmod{\widetilde{\Gamma_0(N)}} \Longleftrightarrow \alpha \equiv \pm\beta \pmod{\Gamma_0(N)}.$$

Working in $H^+(N)$ is sufficient for the first stage of our algorithm, when we want to find certain cusp forms in $S_2(N)$, since $H^+(N) \otimes_{\mathbb{Q}} \mathbb{C} \cong S_2(N)$, both as vector spaces and as modules for the Hecke algebra $\mathbb{T}$. Hence eigenvectors in $H^+(N)$ correspond to eigenforms in $S_2(N)$. Since these eigenforms (or, more accurately, newforms—see the next section) have Fourier expansions in which the Fourier coefficients are determined by their Hecke eigenvalues, we can determine these coefficients indirectly by computing explicitly the action of the Hecke algebra $\mathbb{T}$ on $H^+(N)$.

2.6 Modular forms and modular elliptic curves

Let $S_2(N)$ denote, as above, the space of cusp forms of weight 2 on $\Gamma_0(N)$. Forms $f(z) \in S_2(N)$ have Fourier expansions of the form

$$f(z) = \sum_{n=1}^{\infty} a(n, f) e^{2\pi i n z},$$

with coefficients $a(n, f) \in \mathbb{C}$. The corresponding differentials $2\pi i f(z) dz$ are (the pullbacks of) holomorphic differentials on the Riemann surface $X_0(N)$. Hence $S_2(N)$ is a complex vector space of dimension g, where g is the genus of $X_0(N)$, and $2g = \dim H(N)$. Moreover, $S_2(N) = S_2(N)_{\mathbb{Q}} \otimes_{\mathbb{Q}} \mathbb{C}$ where $S_2(N)_{\mathbb{Q}}$ is the subset of $S_2(N)$ consisting of forms $f(z)$ with *rational* Fourier coefficients $a(n, f)$. This rational structure on $S_2(N)$ is a consequence of the deep fact that $X_0(N)$ may be viewed as the complex points of an algebraic curve defined over $\mathbb{Q}$; it may also be proved using Hecke operators and the duality with homology.

We are interested here in "rational newforms" f: that is, forms f which have rational Fourier coefficients $a(n, f)$, are simultaneous eigenforms for all the Hecke operators, and which are also "newforms" in the sense of Atkin and Lehner (see [1]). We briefly recall the definition.

For each proper divisor M of N and each $g \in S_2(M)$, the forms $g(Dz)$ for divisors D of N/M are in $S_2(N)$. The subspace $S_2^{\mathrm{old}}(N)$ of $S_2(N)$ spanned by all such forms is called the space of *oldforms*. There is also an inner product on $S_2(N)$, called the Petersson inner product, with respect to which the Hecke operators are self-adjoint (Hermitian). Define $S_2^{\mathrm{new}}(N)$ to be the orthogonal complement in $S_2(N)$ of $S_2^{\mathrm{old}}(N)$ with respect to the Petersson inner product. The restriction of the Hecke algebra $\mathbb{T}$ to $S_2^{\mathrm{new}}(N)$ is semisimple; $S_2^{\mathrm{new}}(N)$ has a basis consisting of simultaneous eigenforms, and these eigenforms are called *newforms*.

In general, newforms come in conjugate sets of $d \geq 1$ forms with eigenvalues generating an algebraic number field of degree d. The periods of such a set of conjugates $\{f\}$ form a lattice Λ of rank $2d$ in $\mathbb{C}^d$, and hence an abelian variety $A_f = \mathbb{C}^d/\Lambda$, which is defined over $\mathbb{Q}$. Here we will only be interested in the case $d = 1$, where the Hecke eigenvalues and hence Fourier coefficients of f are rational (in fact integers, being eigenvalues of integral matrices and hence algebraic integers). We will call such a form f a *rational newform*. Thus a rational newform f has an associated period lattice Λ_f:

$$\Lambda_f = \{\langle \{\alpha, \beta\}, f \rangle \mid \alpha, \beta \in \mathcal{H}^*, \alpha \equiv \beta \pmod{\Gamma_0(N)}\}$$

which is a discrete rank 2 subgroup of $\mathbb{C}$. Then $E_f = \mathbb{C}/\Lambda_f$ is an elliptic curve, the modular elliptic curve attached to f. Moreover it is known that E_f is defined over $\mathbb{Q}$, has conductor N, and has L-series $L(E_f, s) = \sum a(n, f)n^{-s}$ where $f = \sum a(n, f)\exp(2\pi inz)$. (See [64], [55], and [7] for proofs of these statements, and [28] for a fuller discussion.)

The Fourier coefficients $a(n, f)$ of a newform $f(z) = \sum a(n, f)\exp(2\pi inz)$ are obtained from the Hecke eigenvalues of f as follows (see [1]). Firstly, for a newform f we always have $a(1, f) \neq 0$, and we normalize so that $a(1, f) = 1$. Then:

If p is a prime not dividing N, and $f\,|T_p = a_p f$, then $a(p, f) = a_p$.

If q is a prime dividing N, and $f\,|W_q = \varepsilon_q f$ with $\varepsilon_q = \pm 1$, then

$$a(q, f) = \begin{cases} -\varepsilon_q & \text{if } q^2 \nmid N, \\ 0 & \text{if } q^2 | N. \end{cases} \tag{2.6.1}$$

For prime powers, we have the recurrence relation

$$a(p^{r+1}, f) = a(p, f)a(p^r, f) - \delta_N(p)pa(p^{r-1}, f) \qquad (r \geq 1) \tag{2.6.2}$$

where

$$\delta_N(p) = \begin{cases} 1 & \text{if } p \nmid N, \\ 0 & \text{if } p | N. \end{cases}$$

Finally, for composite indices we have multiplicativity: $a(mn, f) = a(m, f)a(n, f)$ when m and n are relatively prime.

With this background we may now make more precise what we mean by "computing the modular elliptic curves of conductor N". We do the following:

(1) Compute the space $H^+(N)$ in terms of M-symbols and their relations.
(2) Compute the action of sufficient Hecke operators W_q and T_p on $H^+(N)$ to determine the one-dimensional eigenspaces with rational eigenvalues; by duality, we now know the rational newforms in $S_2(N)$. Oldforms can be recognized, since in any systematic computation we will have already found them at some lower level M dividing N.
(3) Find a $\mathbb{Z}$-basis for the period lattice Λ_f, for each rational newform f, computing the generating periods to high precision.
(4) Given a $\mathbb{Z}$-basis for Λ_f, compute the coefficients of an equation for the attached elliptic curve E_f.
(5) As well as the period lattice of the curves E_f, we can also compute the rational number $L(E_f, 1)/\Omega(E_f)$ (exactly) and the real value $L(E_f, 1)$ (approximately). Also, when $L(E_f, 1) = 0$ we can also determine the order of vanishing of $L(E_f, s)$ at $s = 1$, giving the analytic rank r of E_f, and the value of the derivative $L^{(r)}(E_f, 1)$, which is important in view of the Birch–Swinnerton-Dyer conjecture; we will discuss the latter computations in a later section.

This is the program which we wish to carry out, and have in fact carried out for all $N \leq 5077$. In sections 2.7–2.14 we discuss steps (2)–(5) in more detail.

2.7 Splitting off one-dimensional eigenspaces

Having computed a representation of $H^+(N)$ in terms of M-symbols, we now wish to identify the one-dimensional eigenspaces with rational integer eigenvalues for all the Hecke operators. For each eigenspace we will later need a *dual* basis vector in order to compute the projection of an arbitrary vector onto the eigenspace. Explicitly, we identify $H^+(N)$ with $\mathbb{Q}^g$ via our M-symbol basis, representing each cycle as a column vector; each operator matrix acts on the left. Elements of the dual space will then be represented as row vectors. Projection onto a

one-dimensional eigenspace is then achieved by multiplying on the left by the appropriate row vector, which is defined up to scalar multiple by its being a simultaneous left eigenvector of each matrix. In our implementation, we do not distinguish between row and column vectors, and our linear algebra routines are designed to give right eigenvectors, so in practice all we do is find simultaneous eigenvectors for the transposes of the operator matrices. Projection (of a column vector) is then achieved by taking the dot product with the appropriate dual (row) vector. These remarks seem fairly trivial, but we need to be completely explicit if we are to implement these ideas successfully.

We wish to compute as few T_p as possible at this stage, to save time; we will have a much faster way of computing many Hecke eigenvalues later (see Section 2.9), once the eigenspaces have been found.

We also need to identify "oldclasses": these are also common eigenspaces for all the T_p (though not for all the W_q, see below) but have dimension greater than 1. In order to recognize and discard oldforms as early as possible, we can create a cumulative database of the number of newforms and the first few Hecke eigenvalues (including all W_q-eigenvalues) of each newform at each level. If we proceed systematically through the levels N in order, we will thus always know about the newforms at levels M dividing N but less than N.

An alternative approach might be possible here, in which we use further operators at level N, such as the U_q of [1], to eliminate all but newforms. We have not devised such a scheme which works in full generality; the advantage would be that each level could then be treated in isolation, independently of lower levels, but this was not necessary in our systematic investigations which resulted in the tables in this volume.

Before starting to split $H^+(N)$ we have the following data: the number of rational newforms g in $S_2(M)$ for proper divisors M of N; and for each such g, the W_q-eigenvalue ε_q for all primes q dividing M and the T_p-eigenvalue a_p for several primes p not dividing N. Each form g generates an "oldclass" in $S_2(N)$: a subspace of forms which have the same eigenvalue a_p for all primes p not dividing N. A basis for this oldclass consists of the forms $g(Dz)$ for all positive divisors D of N/M; hence its dimension is $d(N/M)$, the number of positive divisors of N/M. The forms in the oldclass do not necessarily, however, have the same W_q-eigenvalue for primes q dividing N. We now proceed to find these eigenvalues explicitly.

To simplify the following exposition, observe that the W_q operators may be defined for *all* primes q, not just those dividing the level N, using the matrices $W_q = \begin{pmatrix} q^\alpha x & y \\ Nz & q^\alpha w \end{pmatrix}$ of determinant q^α where $q^\alpha \mid\mid N$; for if in fact $q \nmid N$, then $\alpha = 0$, so that $W_q \in \Gamma_0(N)$ and $f\,|W_q = f$ for all $f \in S_2(N)$. Thus in such a case, W_q reduces to the identity.

We first consider the case where N/M is a prime power.

LEMMA 2.7.1. *Let g be a newform in $S_2(M)$, let l be a prime with $g\,|W_l = \varepsilon g$, and let $N = q^\beta M$ where q is also prime. Thus g determines an oldclass of dimension $\beta + 1$, spanned by the forms $g_i(z) = q^i g(q^i z) \in S_2(N)$, for $0 \le i \le \beta$.*

(1) *If $l \ne q$, then $g_i\,|W_l = \varepsilon g_i$ for all i;*

(2) *If $l = q$, then $g_i\,|W_q = \varepsilon g_{\beta - i}$.*

In case (1)*, all members of the oldclass have the same W_q-eigenvalue ε as g, so ε has multiplicity $\beta+1$ (as an eigenvalue of W_l acting on this oldclass). In case* (2)*, the ε-eigenspace for W_q has dimension $[(2+\beta)/2]$ (that is, $\frac{1}{2}(\beta+1)$ if β is odd, or $\frac{1}{2}(\beta+2)$ if β is even).*

PROOF. Suppose $l^\alpha \mid\mid M$. In case (1) we have $l^\alpha \mid\mid N$ also. Let $W_l^{(N)} = \begin{pmatrix} l^\alpha x & y \\ Nz & l^\alpha w \end{pmatrix}$, with $\det W_l^{(N)} = l^\alpha$. Then for $0 \le i \le \beta$ we have

$$\begin{pmatrix} q^i & 0 \\ 0 & 1 \end{pmatrix} W_l^{(N)} = W_l^{(M)} \begin{pmatrix} q^i & 0 \\ 0 & 1 \end{pmatrix}$$

where $W_l^{(M)} = \begin{pmatrix} l^\alpha x & q^i y \\ Mq^{\beta-i} z & l^\alpha w \end{pmatrix}$ also has determinant l^α. Hence

$$\begin{aligned} g_i \left| W_l^{(N)} \right. &= g \left| \begin{pmatrix} q^i & 0 \\ 0 & 1 \end{pmatrix} W_l^{(N)} \right. \\ &= g \left| W_l^{(M)} \begin{pmatrix} q^i & 0 \\ 0 & 1 \end{pmatrix} \right. \\ &= \varepsilon g \left| \begin{pmatrix} q^i & 0 \\ 0 & 1 \end{pmatrix} = \varepsilon g_i. \right. \end{aligned}$$

In case (2), when $q = l$, we have $q^{\alpha+\beta} \mid\mid N$. Let $W_q^{(N)} = \begin{pmatrix} q^{\alpha+\beta} x & y \\ Nz & q^{\alpha+\beta} w \end{pmatrix}$ with $\det W_q^{(N)} = q^{\alpha+\beta}$. Then for $0 \le i \le \beta$ we have

$$\begin{pmatrix} q^i & 0 \\ 0 & 1 \end{pmatrix} W_q^{(N)} = W_q^{(M)} \begin{pmatrix} q^{\beta-i} & 0 \\ 0 & 1 \end{pmatrix}$$

(modulo scalar matrices, which act trivially), where $W_q^{(M)} = \begin{pmatrix} q^{\alpha+i} x & y \\ Mz & q^{\alpha+\beta-i} w \end{pmatrix}$ has determinant q^α. Hence $g_i \left| W_q^{(N)} \right. = \varepsilon g_{\beta-i}$ as required. As a basis for the ε-eigenspace for $W_q^{(N)}$ we may take the forms $g_i + g_{\beta-i}$ for $0 \le i \le \beta/2$, and for the $(-\varepsilon)$-eigenspace, $g_i - g_{\beta-i}$ for $0 \le i < \beta/2$. Hence the multiplicities are as stated. □

Using this result we can easily extend to the general case by induction on the number of prime divisors of N/M, giving the following result.

PROPOSITION 2.7.2. *Let g be a newform in $S_2(M)$ where $M \mid N$. Write $N/M = \prod_{i=1}^k q_i^{\beta_i}$, so that the oldclass in $S_2(N)$ coming from g has dimension $d(N/M) = \prod(1+\beta_i)$.*
(1) *For every prime q not dividing N/M, the W_q-eigenvalue of every form in the oldclass is the same as that of g.*
(2) *Suppose $g \left| W_{q_i} \right. = \varepsilon_i g$ for $1 \le i \le k$. Let*

$$n_i^+ = \begin{cases} \frac{1}{2}(\beta_i + 1) & \text{if } \beta_i \text{ is odd,} \\ \frac{1}{2}(\beta_i + 2) & \text{if } \beta_i \text{ is even and } \varepsilon_i = +1, \\ \frac{1}{2}\beta_i & \text{if } \beta_i \text{ is even and } \varepsilon_i = -1, \end{cases} \tag{2.7.1}$$

and put $n_i^- = 1 + \beta_i - n_i^+$, so that $\prod(n_i^+ + n_i^-) = \prod(\beta_i + 1) = d(N/M)$. If $(\delta_1, \delta_2, \ldots, \delta_k)$ is any k-vector with each $\delta_i = \pm$, then the subspace of oldforms in the oldclass on which W_{q_i} has eigenvalue δ_i for $1 \le i \le k$ has dimension $\prod_{i=1}^k n_i^{\delta_i}$. □

Hence we are able to compute from our database a complete set of "sub-oldclasses"—that is, subspaces of oldclasses which have the same eigenvalues for *all* the operators T_p and W_q—with their dimensions.

Having thus computed a list of sub-oldclasses with their dimensions, W_q-eigenvalues and first few T_p-eigenvalues, we now proceed to find "new" one-dimensional rational eigenspaces of $H^+(N)$ as follows. We consider each prime in turn, starting with the q which divide N, then moving on to the p which do not divide N, computing W_q or T_p as appropriate. For each, we consider all possible integer eigenvalues ($\varepsilon_q = \pm 1$ for W_q, and a_p with $|a_p| < 2\sqrt{p}$ for T_p) and restrict all subsequent operations to each nonzero eigenspace in turn. At any given stage we have a subspace of $H^+(N)$ on which all the operators so far considered act as

scalars. Comparing with the oldform data we can tell whether this subspace consists entirely of oldforms: if so, we discard it. If not, and the subspace is one-dimensional, we have found a rational one-dimensional eigenspace corresponding to a newform. We then record a basis vector and a list of the (prime–eigenvalue) pairs needed to isolate this subspace. Otherwise we proceed recursively to the next prime and the next operator.

At the end of this stage of the computation in $H^+(N)$, we have found the number of rational one-dimensional "new" eigenspaces in $H^+(N)$, or equivalently, the number of rational newforms in $S_2(N)$. For each we have a dual (integer) eigenvector, which we will use to compute a large number of Hecke eigenvalues in Section 2.9.

Implementation. In preparation for splitting off the one-dimensional eigenspaces of $H^+(N)$ we compute the matrices of all the W-operators acting on $H^+(N)$, and store their transposes. We also collect from the "oldform database" information about the newforms at all levels M dividing and less than N. For each oldclass we must compute the eigenvalue multiplicities for each W_q using the formula (2.7.1) above.

The splitting itself is done recursively. At the general stage, at depth n, we have the following data:

- a particular subspace S of $H^+(N)$ (initially the whole of $H^+(N)$);
- a list of n primes (starting with the q dividing N, and initially empty);
- a list of eigenvalues, one for each of the primes in the list.

Here S is precisely the subspace of $H^+(N)$ on which the first n operators have the given eigenvalues.

Given this data, the recursive procedure does the following:

(1) check whether S consists entirely of oldforms, by comparing the list of eigenvalues which determine S with those of each "suboldclass"; if so, terminate this branch;
(2) otherwise, if $\dim S = 1$ then store the (single) basis vector for S in a cumulative list and terminate;
(3) otherwise, take the next operator T in sequence (computing and storing its matrix if it has not been used before) and compute the matrix T_S of its restriction to S; for all possible eigenvalues a of T, compute the kernel of $T_S - aI$; if non-trivial, pass the accumulated data, together with this kernel as a new working subspace, to the procedure at the next depth.

This procedure has been found to work extremely efficiently in practice. The only practical difficulty is the possibility of overflow during Gaussian elimination; it was found that the early use of W-operators was an efficient way of avoiding this for as long as possible. However, for larger values of N we were forced to abandon single-precision integer arithmetic for the linear algebra at this stage, and instead use a modular method, working in $\mathbb{Z}/P\mathbb{Z}$ for some large prime P, instead of in $\mathbb{Z}$. Alternatively, one could use multiprecision arithmetic, but this is likely to be slower.

In all subsequent calculations in $H^+(N)$, we will be interested only in the one-dimensional eigenspaces corresponding to rational newforms. To enhance the speed we now change the main M-symbol lookup tables: each vector in the table is replaced by the vector of its projections onto each of the subspaces, computed simply by taking the dot product with each dual eigenvector.

For each one-dimensional rational eigenspace found, we also compute the eigenvalue ε_N of the Fricke involution W_N, which is the product of all the W_q involutions. The significance of this is that $w = -\varepsilon_N$ is the sign of the functional equation of the L-series $L(f, s)$ attached to the newform f (see [64] and the next section).

2.8 $L(f,s)$ and the evaluation of $L(f,1)/\Omega(f)$

Attached to each newform f in $S_2(N)$ there is an L-function $L(f,s)$, defined as follows via Mellin transform:

$$L(f,s) = (2\pi)^s \Gamma(s)^{-1} \int_0^{i\infty} (-iz)^s f(z) \frac{dz}{z}. \tag{2.8.1}$$

This gives an entire function of the complex variable s. Substitute the Fourier expansion $f(z) = \sum_{n=1}^{\infty} a(n,f) \exp(2\pi i n z)$ and integrate term by term; provided that $\mathrm{Re}(s) > 3/2$ (for convergence), we obtain a representation of $L(f,s)$ as a Dirichlet series:

$$L(f,s) = \sum_{n=1}^{\infty} \frac{a(n,f)}{n^s}. \tag{2.8.2}$$

This L-function is one of the key links between the newform f and the modular elliptic curve E_f defined in Section 2.6 by its periods. First of all, the multiplicative relations satisfied by the coefficients $a(n,f)$, given above in Section 2.6, are equivalent to the statement that the Dirichlet series in (2.8.2) has an Euler product expansion:

$$\sum_{n=1}^{\infty} \frac{a(n,f)}{n^s} = \prod_{p \nmid N} \left(1 - a(p,f)p^{-s} + p^{1-2s}\right)^{-1} \prod_{p|N} \left(1 - a(p,f)p^{-s}\right)^{-1}. \tag{2.8.3}$$

This is exactly the form of the L-function of an elliptic curve of conductor N defined over $\mathbb{Q}$, and in fact the fundamental result (see [7], though partial results were known considerably earlier) is that

$$L(f,s) = L(E_f, s). \tag{2.8.4}$$

Thus (2.8.1) provides an analytic continuation to the entire plane of the L-function attached to the curve E_f, such as is conjectured to exist for all elliptic curves E defined over $\mathbb{Q}$.

Instead of the function $L(f,s)$ defined above by (2.8.1), it is sometimes convenient to use the variant with extra 'infinite' Euler factors:

$$\Lambda(f,s) = N^{s/2}(2\pi)^{-s}\Gamma(s)L(f,s) = \int_0^{\infty} f(iy/\sqrt{N}) y^{s-1} dy. \tag{2.8.5}$$

Thus for $\mathrm{Re}(s) > 3/2$ we have

$$\Lambda(f,s) = N^{s/2}(2\pi)^{-s}\Gamma(s) \sum_{n=1}^{\infty} \frac{a(n,f)}{n^s}.$$

The functions $L(f,s)$ and $\Lambda(f,s)$ also satisfy functional equations relating their values at s and $2-s$. For since f is an eigenform for the Hecke algebra $\mathbb{T}$, it is in particular an eigenform for the Fricke involution W_N. Suppose that $f|W_N = \varepsilon_N f$ with $\varepsilon_N = \pm 1$: that is, $f(-1/(Nz)) = \varepsilon_N N z^2 f(z)$. With $z = iy/\sqrt{N}$ this gives $f(i/y\sqrt{N}) = -\varepsilon_N y^2 f(iy/\sqrt{N})$. Hence the substitution of $1/y$ for y in (2.8.5) yields the functional equation

$$\Lambda(f, 2-s) = -\varepsilon_N \Lambda(f,s) \tag{2.8.6}$$

(note the change of sign). In view of (2.8.4), this gives a functional equation for $L(E_f,s)$ too, of the form conjectured for all elliptic curves over $\mathbb{Q}$.

From (2.8.6), we deduce that $L(f,1) = \Lambda(f,1) = 0$ when $\varepsilon_N = +1$; more generally, $L(f,s)$ has a zero of odd order when $\varepsilon_N = +1$, and a zero of even order (or no zero) when $\varepsilon_N = -1$. The significance of this is that the Birch–Swinnerton-Dyer conjectures predict that the order of the zero of $L(E,s)$ is equal to the rank of $E(\mathbb{Q})$, for an elliptic curve E defined over $\mathbb{Q}$. Thus we will be able to compare this order with the rank of the modular curves E_f, once we have found their equations explicitly.

The Birch–Swinnerton-Dyer conjectures also predict the value of $L(E,1)/\Omega(E)$, which in the case of our modular curve $E = E_f$ is $L(f,1)/\Omega(f)$, where $\Omega(f)$ is a certain period of f. We now discuss the relationship between $L(f,1)$ and the periods of f (by which we will always mean the periods of the differential $2\pi i f(z)dz$).

Substituting $s = 1$ into the Mellin transform formula (2.8.1), we obtain

$$L(f,1) = -2\pi i \int_0^{i\infty} f(z)dz = -\langle \{0,\infty\}, f\rangle . \tag{2.8.7}$$

The modular symbol $\{0,\infty\}$ is in the rational homology, so that $L(f,1)$ is a rational multiple of some period of f. To find the rational factor, we use the trick of "closing the path" (see [38, page 286] or [37]).

For each prime p not dividing N we have, by (2.4.1),

$$T_p(\{0,\infty\}) = \{0,\infty\} + \sum_{k=0}^{p-1} \{k/p,\infty\} = (1+p)\{0,\infty\} + \sum_{k=0}^{p-1} \{k/p, 0\},$$

and hence

$$(1+p-T_p)\cdot\{0,\infty\} = \sum_{k=0}^{p-1} \{0,k/p\}. \tag{2.8.8}$$

Let a_p be the T_p-eigenvalue of f, so that $T_p f = a_p f$. Integrating the differential $2\pi i f(z)dz$ along both sides of (2.8.8) gives

$$(1+p-a_p)\cdot\langle\{0,\infty\}, f\rangle = \sum_{k=0}^{p-1} \langle\{0,k/p\}, f\rangle . \tag{2.8.9}$$

Since p does not divide N, each modular symbol $\{0,k/p\}$ on the right of (2.8.9) is *integral*: that is, in $H_1(X_0(N),\mathbb{Z})$. Thus the right-hand side of (2.8.9) is a period of f. It is even a real period, since

$$\overline{\langle\{0,k/p\},f\rangle} = \langle\{0,-k/p\},f\rangle = \langle\{0,(p-k)/p\},f\rangle .$$

Let $\Omega_0(f)$ denote the least positive real period of f, and set

$$\Omega(f) = \begin{cases} 2\Omega_0(f) & \text{if the period lattice of } f \text{ is rectangular,} \\ \Omega_0(f) & \text{otherwise.} \end{cases}$$

Thus $\Omega(f)/\Omega_0(f)$ is the number of components of the real locus of the elliptic curve E_f. Also note that in each case, $\Omega(f)$ is twice the least real part of a period of f. This is useful since, as we are working in $H^+(N)$, we can only (at this stage) determine the projection of the period lattice Λ_f onto the real axis.

In both cases, (2.8.9) becomes

$$\frac{L(f,1)}{\Omega(f)} = \frac{n(p,f)}{2(1+p-a_p)}, \tag{2.8.10}$$

where $n(p,f)$ is an integer. Note that $1+p-a_p$ is non-zero, since, by well-known estimates, $|a_p| < 2\sqrt{p}$.

Formula (2.8.10) is significant in several ways. On the one hand, let E_f be the modular elliptic curve attached to f as above. Then $L(E_f,1) = L(f,1)$, and $\Omega_0(f) = \Omega_0(E_f)$, the least positive real period of E_f. Thus, once we know a_p and $n(p,f)$ for a single prime p, we can evaluate the rational number $L(E_f,1)/\Omega(E_f)$, whose value is predicted by the Birch–Swinnerton-Dyer conjecture for E_f. In particular, we should have $L(f,1) = 0$ if and only if $E_f(\mathbb{Q})$ is infinite. In the tables we give the value of $L(f,1)/\Omega(f)$ for each rational newform f computed, and observe that the value is consistent with the Birch–Swinnerton-Dyer conjecture in each case.

Secondly, having computed the right-hand side of (2.8.10) for a single prime p, we may (if $L(f,1) \neq 0$) use the fact that $n(p,f)/(1+p-a_p)$ is independent of p to compute the eigenvalue a_p quickly for other p, by computing $n(p,f)$. This is discussed in the next section.

2.9 Computing Fourier coefficients

For each one-dimensional rational eigenspace of $H^+(N)$ we will need to know many Fourier coefficients $a(n,f)$ of the corresponding newform $f(z) = \sum a(n,f)\exp(2\pi i n z)$. These are obtained from the Hecke eigenvalues by the recurrence formulae given in Section 2.6. We already have the eigenvalue ε_q of each W_q operator, and at least one eigenvalue a_{p_0} for the smallest prime p_0 not dividing N, which we recorded as we found the one-dimensional eigenspaces earlier.

It remains to compute a large number of the Hecke eigenvalues a_p for primes p not dividing N. If $L(f,1) \neq 0$ then the most efficient method is to use (2.8.10). First we compute $n(p_0,f)$ from the right-hand side of (2.8.8). (This integer is nonzero if and only if $L(f,1) \neq 0$, by (2.8.10)). For other primes p we then have

$$\frac{n(p,f)}{2(1+p-a_p)} = \frac{n(p_0,f)}{2(1+p_0-a_{p_0})},$$

and hence

$$a_p = 1 + p - \frac{n(p,f)(1+p_0-a_{p_0})}{n(p_0,f)}.$$

The integers $n(p,f)$ may be computed by expressing the right-hand side of (2.8.8) as a linear combination of the M-symbols which generate $H^+(N)$, and then projecting onto the one-dimensional subspace corresponding to f: here we take the dot product with the dual eigenvector computed previously, normalized so that its components are relatively prime integers. The integer this produces is then actually too big by a scaling factor $d_1 d_2$, where d_1 and d_2 are the denominators defined in Section 2.2; this factor can be ignored at this stage, where it cancels out in the computation of a_p, but must be included when we need the actual ratio $L(f,1)/\Omega(f)$ from (2.8.10).

If $L(f,1) = 0$ then a variation of this method may be used. For $\alpha \in \mathbb{Q}$ we have

$$\begin{aligned}(1+p-T_p)\{\alpha,\infty\} &= \{\alpha, p\alpha\} + \sum_{k=0}^{p-1}\left\{\alpha, \frac{\alpha+k}{p}\right\} \\ &= \{0, p\alpha\} + \sum_{k=0}^{p-1}\left\{0, \frac{\alpha+k}{p}\right\} - (p+1)\{0,\alpha\}.\end{aligned} \tag{2.9.1}$$

If p does not divide N and $\alpha = n/d$ with $\gcd(d, N) = 1$ then $[0] = [p\alpha] = [(\alpha + k)/p]$ for all k, so that the right-hand side of (2.9.1) lies in the integral homology $H_1(X_0(N), \mathbb{Z})$. Hence we can express it as an integral linear combination of the generating M-symbols. Projecting onto the rational one-dimensional subspace of $H^+(N)$ corresponding to f, we find that

$$\frac{\operatorname{Re}\langle\{\alpha, \infty\}, f\rangle}{\Omega(f)} = \frac{n(\alpha, p, f)}{2(1 + p - a_p)} \tag{2.9.2}$$

for some integer $n(\alpha, p, f)$, where the left-hand side is independent of p. Thus we can compute each a_p from $n(\alpha, p, f)$, given a_{p_0} and $n(\alpha, p_0, f)$, provided that the latter in nonzero.

It is slightly simpler to use a modular symbol of the form $\{0, \alpha\}$ here instead of $\{\alpha, \infty\}$, since (for suitable α) this will be integral. However the formula analogous to (2.9.1) has more terms of the form $\{0, \beta\}$ on the right, so this is slower in practice.

Remark. Equation (2.9.1) and the remarks following it show that the modular symbol $\{\alpha, \infty\}$ lies in the rational homology $H_1(X_0(N), \mathbb{Q})$ provided that the denominator of α is coprime to N. More generally, for an arbitrary rational number α, the right-hand side of (2.9.1) will be integral provided that $p \equiv 1 \pmod{N}$; this proves that $\{\alpha, \infty\} \in H_1(X_0(N), \mathbb{Q})$ in all cases, which is the Manin-Drinfeld Theorem (Theorem 2.1.3) for $\Gamma_0(N)$.

Implementation. In practice we only use the first method if $L(f, 1) \neq 0$ for *all* the rational newforms f in $S_2(N)$. Otherwise we find a rational α such that $n(\alpha, p_0, f) \neq 0$ for all f, where p_0 is the smallest prime not dividing N.

We have already discussed computation of the integers $n(p, f)$. The $n(\alpha, p, f)$ are computed similarly by expressing the right-hand side of (2.9.1) in terms of the generating M-symbols and projecting onto each eigenspace. Note that the term $\{0, \alpha\}$ of (2.9.1) need only be computed once.

The Hecke eigenvalues which we have computed are stored in a data file for use both in subsequent steps of the calculations at level N, and also as part of the cumulative database which will be accessed when levels which are multiples of N are reached.

The exact number of a_p needed depends on N, and on the form f, and will not be known until the numerical calculation of periods is carried out in the next phase. Our strategy here was first to compute a_p for all p up to some predetermined bound (we used all $p < 1000$ for $N \leq 200$, $p < 2000$ for $200 < N \leq 400$, and $p < 3000$ for $401 < N \leq 1000$). We may also store extra information, so that if more eigenvalues are needed later, these can be computed without having to repeat the time-consuming steps described in Sections 2.1–2.7. Specifically, we may store the following: the M-symbols which generate $H^+(N)$; a table giving each M-symbol as a linear combination of these generators; a basis for $\ker(\delta)$; and a (dual) basis vector for each rational one-dimensional eigenspace.

Recapitulation. At this point we have completed the first phase of the computation at level N, in which we have been working in the space $H^+(N)$. To summarize, we know

(1) the number of rational newforms f in $S_2(N)$; and, for each f,
(2) the sign w of the functional equation for $L(f, s)$;
(3) the ratio $L(f, 1)/\Omega(f)$;
(4) all W_q-eigenvalues ε_q of f;
(5) a large number of T_p-eigenvalues a_p of f.

In particular, we know the number of modular elliptic curves E_f of conductor N (up to isogeny); for each curve, we know the sign of its functional equation and whether or not its L-series vanishes at $s = 1$.

All computations carried out so far are exact and algebraic. In addition, we can also in this first phase compute approximations to the value $L(f, 1)$ (when it is non-zero) and to the

period $\Omega(f)$, though we do not know at this stage whether $\Omega(f)/\Omega_0(f) = 1$ or 2. In other words, we can compute the projection of the period lattice Λ_f onto the real axis. Of course, this is insufficient information from which to construct the curve E_f.

In the second phase, which we describe in Sections 2.10–2.14, we compute the period lattice Λ_f of each rational newform f, and hence obtain an (approximate) equation for the curve $\mathbb{C}/\Lambda_f$.

These "analytic" quantities (periods) will necessarily be computed approximately, by summing certain infinite series whose coefficients involve the Fourier coefficients of f (see below). In order to achieve sufficient accuracy, we may have to compute many thousands of these Fourier coefficients, and it is therefore necessary to have efficient ways of doing this, such as the method described in this section.

2.10 Computing periods I

In order to compute the full period lattice Λ_f for each rational newform f found earlier, we have to work in the full space $H(N)$. By working in $H^+(N)$ we could only compute the real period $\Omega_0(f)$. Although we could also compute the least imaginary period $\Omega_{\mathrm{im}}(f)$ by working similarly in $H^-(N)$ (which would be slightly faster), the lattice spanned by $\Omega_0(f)$ and $\Omega_{\mathrm{im}}(f)$ may have index 2 in Λ_f. Hence from now on we work in $H(N)$.

We begin by computing $H(N)$ using M-symbols as in Section 2.2 (omitting relations (2.5.1)). Let $\gamma_1, \gamma_2, \dots, \gamma_{2g}$ be a $\mathbb{Z}$-basis for $H_1(X_0(N), \mathbb{Z})$ (and hence also a $\mathbb{Q}$-basis for $H(N)$). Using this basis we will identify $H(N)$ with the space of rational column vectors, and dual vectors will be represented by row vectors. Next we read from the data file (created during the first phase) the number of rational newforms and, for each, the eigenvalues a_p and ε_q. For each form f we now compute two integer dual (row) eigenvectors with eigenvalues a_p and ε_q for all p and q: one, v^+, with eigenvalue $+1$ for the $*$ operator, and one, v^-, with eigenvalue -1. This is much faster than repeating the splitting step described in Section 2.7, since we already know the eigenvalues which determine each one-dimensional eigenspace. As before, the eigenvectors $v^\pm$ we compute must be dual eigenvectors, since we will use them for projecting onto the eigenspaces in question.

Let $\gamma^\pm \in H^\pm(N)$ (respectively) be eigenvectors with the same eigenvalues as $v^\pm$, such that $v^+\gamma^+ = v^-\gamma^- = 1$. We view $\gamma^\pm$ as column vectors in $\mathbb{Q}^{2g}$ by expressing them as linear combinations of the basis $\gamma_1, \gamma_2, \dots, \gamma_{2g}$ for $H(N)$. Thus the product $v^+\gamma^+$ is the product of a row vector by a column vector: essentially a dot product. Set $x = \langle \gamma^+, f\rangle$ and $y = -i\,\langle \gamma^-, f\rangle$ (so that $x, y \in \mathbb{R}$). We do not actually compute these vectors $\gamma^\pm$ in practice; they are only needed for this exposition, as they determine the real numbers x and y. Moreover, although the eigenvectors $v^\pm$ which we do use are only determined up to a scalar multiple, we shall see that this choice does not (as it should not) affect the specific period lattice we obtain.

Let $\gamma = \sum_{j=1}^{2g} c_j\gamma_j$ be an arbitrary integral cycle in $H(N)$. We identify γ with the column vector with component c_j. Then we have

$$\langle \gamma, f\rangle = \big\langle (v^+\gamma)\gamma^+ + (v^-\gamma)\gamma^-, f\big\rangle = (v^+\gamma)x + (v^-\gamma)yi. \tag{2.10.1}$$

The period lattice Λ_f is the set of all such integral periods $\langle \gamma, f\rangle$. To determine a $\mathbb{Z}$-basis for Λ_f we proceed as follows. Write $v^+ = (a_1, a_2, \dots, a_{2g})$ and $v^- = (b_1, b_2, \dots, b_{2g})$ with $a_j, b_j \in \mathbb{Z}$. Then as a special case of (2.10.1) we have

$$\langle \gamma_j, f\rangle = a_j x + b_j yi,$$

since $v^+\gamma_j = a_j$ and $v^-\gamma_j = b_j$. Hence Λ_f is spanned over $\mathbb{Z}$ by the $2g$ periods $\langle \gamma_j, f\rangle = a_j x + b_j yi$. Let Λ be the $\mathbb{Z}$-span in $\mathbb{Z}^2$ of the $2g$ pairs (a_j, b_j), and let (λ_1, μ_1), (λ_2, μ_2) be a

$\mathbb{Z}$-basis for Λ. Then we find that

$$\Lambda_f = \{\langle \gamma, f\rangle \mid \gamma \in H(N)\} = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2,$$

where

$$\omega_j = \lambda_j x + \mu_j y i \qquad (j = 1, 2). \tag{2.10.2}$$

Thus ω_1 and ω_2 form a $\mathbb{Z}$-basis for Λ_f.

We may compute (λ_1, μ_1) and (λ_2, μ_2) from v^+ and v^- using the Euclidean algorithm in $\mathbb{Z}$. In fact it is easy to see that there are only two possibilities, since $v^\pm$ are determined within the subspace they generate by being the $+1$ and -1 eigenvectors for an involution. Normalize $v^\pm$ so that each is primitive in $\mathbb{Z}^{2g}$; that is, $\gcd(a_1, \dots, a_{2g}) = \gcd(b_1, \dots, b_{2g}) = 1$. In the first case (which we will call "Type 1"), $v^+ \equiv v^- \pmod 2$, and we may take $(\lambda_1, \mu_1) = (2, 0)$ and $(\lambda_2, \mu_2) = (1, 1)$, so that $\omega_1 = 2x$ and $\omega_2 = x + yi$. In this case $\Omega(f) = \Omega_0(f)$ and the elliptic curve has negative discriminant.

In the second case ("Type 2"), when v^+ and v^- are independent modulo 2, we will be able to take $(\lambda_1, \mu_1) = (1, 0)$ and $(\lambda_2, \mu_2) = (0, 1)$, so that $\omega_1 = x$ and $\omega_2 = yi$. In this case the period lattice is rectangular, $\Omega(f) = 2\Omega_0(f)$, and the elliptic curve has positive discriminant.

It remains to compute the real numbers x and y. We describe two methods: the first computes periods directly, while the second computes them indirectly by computing $L(f \otimes \chi, 1)$ for suitable quadratic characters χ. The latter method is in certain cases more accurate (in that fewer a_p are needed for the same accuracy) but cannot be used when N is a perfect square, as we shall see below.

Observe that the cycles $\gamma^\pm$ do not enter into the calculations directly, but are merely used to define x and y. Also, if either v^+ or v^- is replaced by a scalar multiple of itself, then γ^+ and γ^- (and hence x and y) are scaled down by the same amount, but λ_j and μ_j are scaled up. In particular, it is no loss of generality to assume that $v^\pm$ are primitive integer vectors. Thus (2.10.2) defines ω_1 and ω_2 unambiguously, as generators of the full period lattice of f.

Direct method. The simplest method is essentially the same as that used by Tingley in [67]. Using a recent improvement (see [18]), this method can now be made to converge as well as the indirect method described later.

From (2.10.1) it suffices to compute $\langle \gamma, f\rangle$ for a single cycle γ such that $v^+\gamma$ and $v^-\gamma$ are both nonzero; then by taking real and imaginary parts we can solve (2.10.1) for x and y and compute the periods ω_1 and ω_2 from (2.10.2). (In some cases it may be better in practice to use two different cycles, one for the real period and one for the imaginary period, but for simplicity we will assume that this is not the case.)

We denote by $I_f(\alpha, \beta)$ the integral $I_f(\alpha, \beta) = \int_\alpha^\beta 2\pi i f(z)dz$, and set $I_f(\alpha) = I_f(\alpha, \infty)$. Let $M \in \Gamma_0(N)$; since f is holomorphic, the period integral $I_f(\alpha, M(\alpha))$ is independent of the basepoint α, and can be expressed as $I_f(\alpha) - I_f(M(\alpha))$. We will denote this period of f by $P_f(M)$. Note that any cycle $\gamma \in H(N)$ can be expressed as $\{\alpha, M(\alpha)\}$ for a suitable matrix $M \in \Gamma_0(N)$, and then $\langle \gamma, f\rangle = P_f(M)$. The map $M \mapsto P_f(M)$ is (by Corollary 2.1.2) a group homomorphism from $\Gamma_0(N)$ to the additive group of complex numbers, whose image is the period lattice Λ_f.

Our basic tool for computing periods is the following easy result.

PROPOSITION 2.10.1. *Let $z_0 = x_0 + iy_0 \in \mathcal{H}$, so that $y_0 > 0$. Let f be a cusp form of weight 2 with Fourier coefficients $a(n, f)$. Then*

$$I_f(z_0) = \int_{z_0}^\infty 2\pi i f(z)dz = -\sum_{n=1}^\infty \frac{a(n, f)}{n} e^{2\pi i n x_0} e^{-2\pi n y_0}. \tag{2.10.3}$$

PROOF. Using the Fourier expansion $f(z) = \sum_{n=1}^{\infty} a(n,f)\exp(2\pi inz)$, we can integrate term-by-term over a vertical path from z_0 to ∞ to obtain the result. The term-by-term integration is justified since the series converges absolutely, since $|a(n,f)| \ll n$ and $y_0 > 0$. $\square$

We can sum the series (2.10.3) to obtain an approximation to $I_f(z_0)$, provided that we have sufficiently many Fourier coefficients $a(n,f)$. The important point to notice is that this series is a power series in $\exp(-2\pi y_0)$ (with bounded coefficients since $|a(n,f)| < n$ for all n), so will converge best when y_0 is large (or at least, not too small).

Suppose we are given a matrix $M = \begin{pmatrix} a & b \\ cN & d \end{pmatrix} \in \Gamma_0(N)$, where $a,b,c,d \in \mathbb{Z}$, and we wish to compute the associated period $P_f(M) = I_f(\alpha) - I_f(M(\alpha))$ of f. How should we choose α? If α has large imaginary part, then $M(\alpha)$ will tend to have a small imaginary part; we would like to maximize both of these simultaneously. The simplest solution, used by Tingley in his thesis [67] for the original computations of modular elliptic curves[2], is to choose

$$\alpha = \frac{-d+i}{cN}, \qquad \text{so that} \qquad M(\alpha) = \frac{a+i}{cN}.$$

Thus both α and $M(\alpha)$ have imaginary part $(cN)^{-1}$. (Note that, by replacing M by $-M$ if necessary, we may assume that $c > 0$; we are not interested in M with $c = 0$ since these are parabolic, and hence have zero period.) Hence we obtain the following.

PROPOSITION 2.10.2. *Let $f \in S_2(N)$. Then, for all $M = \begin{pmatrix} a & b \\ cN & d \end{pmatrix} \in \Gamma_0(N)$, the period $P_f(M)$ is given by*

$$P_f(M) = I_f(\alpha) - I_f(M(\alpha)), \tag{2.10.4}$$

where $\alpha \in \mathcal{H}$ is arbitrary. Taking $\alpha = \frac{-d+i}{cN}$, we have:

$$\begin{aligned} P_f(M) &= I_f\left(\frac{-d+i}{cN}\right) - I_f\left(\frac{a+i}{cN}\right) \\ &= \sum_{n=1}^{\infty} \frac{a(n,f)}{n} e^{-2\pi n/cN}\left(e^{2\pi ina/cN} - e^{-2\pi ind/cN}\right). \end{aligned} \tag{2.10.5}$$

To use this result, we take a rational number b/d with denominator d coprime to N, solve $ad - bcN = 1$ for a and c, and set $M = \begin{pmatrix} a & b \\ cN & d \end{pmatrix}$. The integral cycle $\gamma = \{0, b/d\}$ should have the properties that $v^+\gamma$ and $v^-\gamma$ are both nonzero; also, since $y_0 = 1/(Nc)$ with $c > 0$ we should also choose b/d so that c is as small as possible, to speed convergence in the series (2.10.5). This series converges adequately quickly for small N, but when N increases we require too many terms in order to obtain the periods to sufficient precision. (Not only does it take longer to sum the series when we use more terms, but more significantly, computing the coefficients $a(n,f)$ by modular symbols becomes more expensive as n increases.)

The series (2.10.5) is a power series in $\exp(-2\pi/cN)$ for some small positive integer c; at best we might hope to use $c = 1$ and have a power series in $\exp(-2\pi/N)$. We can improve this, however, to give a better formula which involves power series in $\exp(-2\pi/d\sqrt{N})$ for a small positive integer d. This greatly improves the convergence of the series.

[2]and also used in the first edition of this book

In order to do this, we make use of the fact that the newform f is an eigenform for the Fricke involution W_N, which for brevity we will here denote simply W. Thus (as in Section 2.8 above) we have

$$f(z) = \varepsilon\ (f \mid W)(z)$$

where $\varepsilon = \pm 1$ is the Fricke eigenvalue. By changing variables in the integrals, we see that

$$I_f(W(\alpha), W(\beta)) = I_{f|W}(\alpha, \beta) = \varepsilon\ I_f(\alpha, \beta). \tag{2.10.6}$$

In particular, if $\beta = W(\alpha)$ we obtain $I_f(\alpha, W(\alpha)) = -\varepsilon I_f(\alpha, W(\alpha))$, so that when $\varepsilon = +1$ we have $I_f(\alpha, W(\alpha)) = 0$ for all α.

Assume we are in this case ($\varepsilon = +1$). Then in any period integral, we may replace an endpoint α with $W(\alpha)$ without affecting the value of the integral. In particular,

$$P_f(M) = I_f(\alpha, M(\alpha)) = I_f(W(\alpha), M(\alpha)).$$

Setting $\alpha = di/(\sqrt{N} - cNi)$ we find that

$$M(\alpha) = \frac{b}{d} + \frac{i}{d\sqrt{N}} \qquad \text{and} \qquad W(\alpha) = \frac{c}{d} + \frac{i}{d\sqrt{N}},$$

which both have the same imaginary part $1/d\sqrt{N}$. (We may assume that $d > 0$, again by replacing M by $-M$ if necessary.) Hence $P_f(M) = I_f(W(\alpha)) - I_f(M(\alpha)) = I_f(\frac{c}{d} + \frac{i}{d\sqrt{N}}) - I_f(\frac{b}{d} + \frac{i}{d\sqrt{N}})$, where both integrals converge relatively well.

When $\varepsilon = -1$, we can obtain a slightly more complicated result which is just as good in practice. Combining both cases gives the following.

PROPOSITION 2.10.3. *Let $f \in S_2(N)$, such that $f \mid W = \varepsilon f$ with $\varepsilon = \pm 1$. Then for all $M = \begin{pmatrix} a & b \\ cN & d \end{pmatrix} \in \Gamma_0(N)$ the period $P_f(M)$ is given by*

$$P_f(M) = (1 - \varepsilon) I_f(i/\sqrt{N}) + \varepsilon I_f(W(\alpha)) - I_f(M(\alpha)), \tag{2.10.7}$$

where $\alpha \in \mathcal{H}$ is arbitrary. Taking $\alpha = M^{-1}(\frac{b}{d} + \frac{i}{d\sqrt{N}})$, so that $W(\alpha) = \frac{c}{d} + \frac{i}{d\sqrt{N}}$, we have

$$\begin{aligned} P_f(M) &= (1 - \varepsilon) I_f(i/\sqrt{N}) + \varepsilon I_f\left(\frac{c}{d} + \frac{i}{d\sqrt{N}}\right) - I_f\left(\frac{b}{d} + \frac{i}{d\sqrt{N}}\right) \\ &= \sum_{n=1}^{\infty} \frac{a(n,f)}{n} \left((\varepsilon - 1)e^{-2\pi n/\sqrt{N}} + e^{-2\pi n/d\sqrt{N}} \left(e^{2\pi i n b/d} - \varepsilon e^{2\pi i n c/d}\right)\right). \end{aligned} \tag{2.10.8}$$

PROOF. Using $W(i/\sqrt{N}) = i/\sqrt{N}$, we simply compute:

$$\begin{aligned} I_f(\alpha, M(\alpha)) &= I_f(\alpha, i/\sqrt{N}) + I_f(i/\sqrt{N}, W(\alpha)) + I_f(W(\alpha), M(\alpha)) \\ &= \varepsilon I_f(W(\alpha), i/\sqrt{N}) + I_f(i/\sqrt{N}, W(\alpha)) + I_f(W(\alpha), M(\alpha)) \\ &= (1 - \varepsilon)(I_f(i/\sqrt{N}) - I_f(W(\alpha))) + I_f(W(\alpha)) - I_f(M(\alpha)) \\ &= (1 - \varepsilon) I_f(i/\sqrt{N}) + \varepsilon I_f(W(\alpha)) - I_f(M(\alpha)) \end{aligned}$$

which establishes (2.10.7). Then (2.10.8) follows from (2.10.3), using the value of α defined before. $\square$

Note that the term $(1-\varepsilon)I_f(i/\sqrt{N})$, which appears in (2.10.7), is equal to $-L(f,1)$, by (2.11.1) below. Hence this term is zero unless the analytic rank of f is zero.

When we use this method for computing the periods, before proceeding to the next stage we store the following data:

$$\texttt{type}, M, v^+\gamma, v^-\gamma.$$

Here $\texttt{type} = 1$ or 2 denotes the lattice type, M is a matrix in $\Gamma_0(N)$ such that $\gamma = \{0, M(0)\}$, and the integers $v^+\gamma$ and $v^-\gamma$ are nonzero. Then we will be able to compute the periods from stored data quickly without having to recompute $H(N)$ or the eigenvectors $v^{\pm}$. We compute the period $P_f(M)$ using (2.10.8), set $x = \mathrm{Re}(P_f(M))/v^+\gamma$ and $y = \mathrm{Im}(P_f(M))/v^-\gamma$ from (2.10.1), and take the period lattice Λ_f to be the lattice with $\mathbb{Z}$-basis $2x$, $x+yi$ (if type 1) or x, yi (if type 2). If we later find that we need greater accuracy here, then after computing more a_p, we can obtain more accurate values for the periods ω_1 and ω_2 very quickly, without having to repeat the expensive calculation in $H(N)$.

2.11 Computing periods II: Indirect method

The idea here is to compute $\Omega(f)$ indirectly by computing $L(f,1)$ and dividing by the ratio $L(f,1)/\Omega(f)$, which we know from (2.8.10). If $L(f,1) = 0$, and in any case to find the imaginary period, we can use the technique of twisting by a quadratic character χ, since the value $L(f\otimes\chi, 1)$ is a rational multiple of a real or imaginary period of f (depending on whether $\chi(-1) = +1$ or -1), and is non-zero for suitable χ.

We are also interested in the value of $L(f,1)$ for its own sake, in relation to the Birch–Swinnerton-Dyer conjecture for the modular curve E_f. We will return to this, and the method of computing $L^{(r)}(f,1)$ for $r > 0$, in Section 2.13.

If $L(f,1) \neq 0$, then we may compute $L(f,1)$ accurately from (2.8.7) as follows. Let $\varepsilon_N = \pm 1$ be the eigenvalue of the Fricke involution W_N on f. Then in the notation of the previous section, using (2.10.6) and $W_N(i/\sqrt{N}) = i/\sqrt{N}$:

$$\begin{aligned} L(f,1) = -\int_0^{i\infty} 2\pi i f(z)dz &= I_f(\infty, 0) \\ &= I_f(\infty, i/\sqrt{N}) + I_f(i/\sqrt{N}, 0) \\ &= I_f(\infty, i/\sqrt{N}) + \varepsilon_N I_f(i/\sqrt{N}, \infty) \\ &= (\varepsilon_N - 1) I_f(i/\sqrt{N}). \end{aligned} \tag{2.11.1}$$

Thus if $L(f,1) \neq 0$, then necessarily $\varepsilon_N = -1$, and in this case $L(f,1) = -2I_f(i/\sqrt{N})$. Using Proposition 2.10.1 then gives the following result.

Proposition 2.11.1. *If $f(z) = \sum_{n=1}^{\infty} a(n,f)\exp(2\pi i n z) \in S_2(N)$ and $f|W_N = -f$ then*

$$L(f,1) = 2\sum_{n=1}^{\infty} \frac{a(n,f)}{n}\exp(-2\pi n/\sqrt{N}). \tag{2.11.2}$$

Remark. If in (2.11.1) we split the range of integration at $Ai/\sqrt{N}$ for some positive real number A (instead of taking $A = 1$) then we obtain the more general formula

$$L(f,1) = \sum_{n=1}^{\infty} \frac{a(n,f)}{n}\left(\exp(-2\pi An/\sqrt{N}) - \varepsilon_N \exp(-2\pi n/A\sqrt{N})\right),$$

where the right-hand side is independent of A. This can be useful in situations where we do not know the value of ε_N, since we can evaluate this expression for two values of A, say $A = 1$ and $A = 1.1$, and check that the values obtained are approximately the same. For only one of the two possible values of ε_N will this happen. This idea is due to H. Cohen (see [9, Section 7.5]).

More generally, let l be an odd prime not dividing N, and χ the quadratic character modulo l. Define

$$(f \otimes \chi)(z) = \sum_{n=1}^{\infty} \chi(n) a(n, f) \exp(2\pi i n z)$$

and

$$L(f \otimes \chi, s) = (2\pi)^s \Gamma(s)^{-1} \int_0^{i\infty} (-iz)^s (f \otimes \chi)(z) \frac{dz}{z};$$

then for $\mathrm{Re}(s) > 3/2$ we can integrate term-by-term to obtain

$$L(f \otimes \chi, s) = \sum_{n=1}^{\infty} \chi(n) a(n, f) n^{-s}.$$

Suppose, as above, that $f|W_N = \varepsilon_N f$. Then $f \otimes \chi$ is in $S_2(Nl^2)$, and

$$(f \otimes \chi)|W_{Nl^2} = \chi(-N)\varepsilon_N f \otimes \chi$$

(special case of equation (14) in [64]). Hence we can immediately generalize Proposition 2.11.1 to obtain the following.

PROPOSITION 2.11.2. *Let f be as above. Let l be an odd prime not dividing N. If $\chi(-N) = \varepsilon_N$ then $L(f \otimes \chi, 1) = 0$, while if $\chi(-N) = -\varepsilon_N$, then*

$$L(f \otimes \chi, 1) = 2 \sum_{n=1}^{\infty} \frac{\chi(n) a(n, f)}{n} \exp(-2\pi n / l\sqrt{N}). \tag{2.11.3}$$

The values $L(f \otimes \chi, 1)$ are related to the periods of f by a formula similar to (2.8.10). Let $g(\chi)$ be the Gauss sum attached to χ: if $l \equiv 1 \pmod 4$ then $\chi(-1) = +1$ and $g(\chi) = \sqrt{l}$, while if $l \equiv 3 \pmod 4$ then $\chi(-1) = -1$ and $g(\chi) = i\sqrt{l}$. If we set $l^* = \chi(-1)l$ then in all cases we have $g(\chi) = \sqrt{l^*}$. By [64, equation(12)] we have

$$f \otimes \chi = \frac{g(\chi)}{l} \sum_{k=0}^{l-1} \chi(-k) f \left| \begin{pmatrix} l & k \\ 0 & l \end{pmatrix} \right..$$

Hence

$$\begin{aligned} L(f \otimes \chi, 1) &= -\langle \{0, \infty\}, f \otimes \chi \rangle \\ &= -\frac{g(\chi)}{l} \sum \chi(-k) \left\langle \{0, \infty\}, f \left| \begin{pmatrix} l & k \\ 0 & l \end{pmatrix} \right. \right\rangle \\ &= -\frac{g(\chi)}{l} \sum \chi(-k) \langle \{k/l, \infty\}, f \rangle \\ &= \chi(-1) \frac{g(\chi)}{l} \langle \gamma_l, f \rangle \\ &= \frac{1}{\sqrt{l^*}} \langle \gamma_l, f \rangle, \end{aligned}$$

where

$$\gamma_l = \sum_{k=0}^{l-1} \chi(k)\,\{0, k/l\}\,.$$

Here we have used the identity $\sum \chi(k) = 0$. Since l does not divide N, the cycle γ_l is in the integral homology. Thus for each prime l not dividing $2N$ we can define an integral period

$$P(l, f) = \langle \gamma_l, f \rangle\,,$$

and we have shown that

$$P(l, f) = \sqrt{l^*} L(f \otimes \chi, 1).$$

Clearly $(\gamma_l)^* = \chi(-1)\gamma_l$, since $\{0, k/l\}^* = \{0, -k/l\}$. So, if $\chi(-1) = +1$, then $\gamma_l \in H^+(N)$, hence $P(l, f)$ is an integer multiple of the real period $\Omega_0(f)$, and thus of the form $m^+(l, f)x$ for some integer $m^+(l, f)$. So, provided that $m^+(l, f) \neq 0$, we have

$$x = \sqrt{l}\,\frac{L(f \otimes \chi, 1)}{m^+(l, f)} = \frac{P(l, f)}{m^+(l, f)}. \tag{2.11.4}$$

In practice, if we express γ_l as a linear combination of the basis cycles γ_j and thus view it as a column vector, then $m^+(l, f) = v^+\gamma_l$.

Similarly, if $\chi(-1) = -1$ then $\gamma_l \in H^-(N)$, and $P(l, f) = m^-(l, f)yi$, where $m^-(l, f) = v^-\gamma_l$ is an integer, so that if $m^-(l, f) \neq 0$ then

$$y = \sqrt{l}\,\frac{L(f \otimes \chi, 1)}{m^-(l, f)} = \frac{P(l, f)}{im^-(l, f)}. \tag{2.11.5}$$

Assuming that N is not a perfect square, we find the smallest primes $l^+ \equiv 1 \pmod 4$ and $l^- \equiv 3 \pmod 4$ (not dividing N) such that $m^+ = m^+(l^+, f)$ and $m^- = m^-(l^-, f)$ are nonzero. A necessary (but not sufficient) condition for this to be true is that for the associated quadratic characters, $\chi_1(-N) = \chi_2(-N) = -\varepsilon_N$; for if $\chi(-N) = \varepsilon_N$ then the sign of the functional equation for $L(f \otimes \chi, s)$ is -1, and hence $L(f \otimes \chi, 1) = 0$. Suitable primes always exist, provided that N is not a perfect square, by a theorem of Murty and Murty (see [44]). We then compute $L(f \otimes \chi_j, 1)$ for $j = 1, 2$ from (2.11.3), obtain x and y from (2.11.4) and (2.11.5), and finally substitute in (2.10.2) as before to obtain the periods ω_1 and ω_2.

If N is a square, however, then $\chi(-N) = \chi(-1)$ for all primes l not dividing $2N$; hence we will only be able to find the real period this way if $\varepsilon_N = -1$, and only the imaginary period if $\varepsilon_N = +1$. Rather than seek a way round this difficulty we always use the "direct" method to compute the periods when N is square.

To assist convergence in (2.11.3) we clearly want to choose l as small as possible. It is a simple matter to estimate the error obtained in truncating the series (2.11.3) for $L(f \otimes \chi, 1)$ at a certain point $n = n_{\max}$. In practice we may use this to estimate the number of eigenvalues a_p needed to obtain the desired accuracy. However, to save time, we did not in all cases compute this many a_p, if the computed values of c_4 and c_6 (see Section 2.14) were close to integers, and when rounded led us to the coefficients of an elliptic curve of conductor N.

Note that, apart from the numerical evaluation of the periods $P(l^\pm, f)$ (using the series (2.11.3) for $L(f \otimes \chi, 1)$), all these computations are purely algebraic: we express the cycles γ_l in terms of our homology basis using continued fractions, and take the dot products of the resulting column vectors with our dual eigenvectors $v^\pm$ to obtain the integers $m^\pm$.

The result of this algebraic computation consists of the following data for each rational newform f: primes $l^\pm$ congruent respectively to ± 1 modulo 4; nonzero integers $m^\pm$; and the

type (1 or 2) of the lattice. As in the direct method, before proceeding we store the following data for each newform f:

$$\texttt{type}, l^+, m^+, l^-, m^-.$$

To compute the lattice from this data set of five integers, we compute the periods $P(l^\pm, f)$ using formula (2.11.3), divide by $m^\pm$ respectively to obtain x and y, and take Λ_f to be the lattice with $\mathbb{Z}$-basis $2x$, $x+yi$ (if type 1) or x, yi (if type 2). In practice we store just these five integers, and recompute the periods when we need them. In particular, if at the first attempt we are unable to compute the integer invariants c_4, c_6 of the curve E_f to sufficient precision to recognize them, then we will return to $H^+(N)$ in order to compute more Hecke eigenvalues, and then recompute the periods to greater precision without having to recompute $H(N)$.

Tricks and shortcuts.

In fact, the data l^+ and m^+ can be computed earlier in the first $H^+(N)$ phase, since they only depend on the real projection of the period lattice. Hence we can already compute the real period x from the data we have from the first phase. Moreover, it is easy to find a suitable prime l^- once we know the Hecke eigenvalues of f, by numerically computing $P(l, f)$ for several primes $l \equiv -1 \pmod 4$ until we find a value which is clearly non-zero.

It follows that the only purpose of the extremely expensive second phase of the computation, working in $H(N)$, is to determine the integer factor m^- and the type of the lattice. An alternative approach, which we have used systematically for larger levels ($N > 3200$), is simply to guess the value of m^- by trying each positive integer m in turn. For each $m \geq 1$ we set $y = P(l^-, f)/m$ and test the two possible lattices (one of each type). If either lattice has approximate integer invariants c_4 and c_6, and the rounded integral values are valid invariants of an elliptic curve over $\mathbb{Q}$, and the resulting curve has conductor N, then we store for later use the successful value m^- of m and the type, and consider the curve E'_f we have found as a possible candidate for the actual modular elliptic curve E_f.

The curves E_f and E'_f are certainly isogenous; they even have the same real period. In many cases, the curve E'_f has no rational isogenies; in such a case we can conclude that $E_f = E'_f$ with no ambiguity. In any case, we can compute the isogeny class of curves isogenous to E'_f via rational isogenies, and the only loss is that we do not always know exactly which curve in the class is the "strong Weil curve" E_f. (A further disadvantage is that we cannot compute the degree of the modular parametrization of E_f, as this requires knowledge of $H(N)$: see Section 2.15 below.)

The great advantage of this method is that in only a few seconds computation time, as soon as we have a rational newform, we can (almost always) write down an associated curve E'_f; before this was implemented, it could take many hours of computation time to determine $H(N)$, find the eigenvectors $v^\pm$, and hence determine the factor m^- and the lattice type, before we could compute E_f.

Finally we discuss some variants of the trick just described.

1. We may use the same trick to find l^+ and m^+ if we have not computed them earlier. Then we are obtaining the period lattice and equation of the curve using only the Fourier coefficients of f (i.e. the coefficients of the L-series of the curve); the sign of the functional equation; and the conductor N. No modular symbol information at all is needed in this case. In fact, one may even guess the sign of the functional equation if all one has is the L-series; see the remark after Proposition 2.11.1.

2. Let l_1 and l_2 be two primes $\equiv -1 \pmod 4$ for which $-N$ has the correct quadratic character, so that $P(l_1, f)$ and $P(l_2, f)$ are both not trivially zero. We may compute the periods $P(l_j, f)$; assume that these are nonzero (or use different primes l_j). We know that there exist nonzero integers m_j such that $P(l_j, f) = m_j yi$ for $j = 1, 2$. Therefore $P(l_2, f)/P(l_1, f) =$

m_2/m_1, and we may compute a floating point approximation to this rational number. In practice (provided we have many Fourier coefficients, and the primes l_j are small) we will be able to recognize this rational number (say by using continued fractions). Its denominator is a factor of the unknown integer m_1. If we do this for several different values of l_2 (with the same l_1) then the least common multiple of the denominators may give us a nontrivial factor of m_1, and then in our search for the exact value we may restrict to multiples of this factor. This is useful in practice.

3. Another possibility, which we have not implemented, is to compute $H^-(N)$ in order to determine m^- exactly, as we do m^+ from $H^+(N)$. This would be no harder than the original computation of $H^+(N)$, and in fact it would be easier to find the eigenvector corresponding to each newform f, since we already know its eigenvalues. The result would be that we would have computed exactly all the data we need in a shorter time than would be required for computing $H(N)$, except for the type of the lattice. Then the only ambiguity is that we would not know the type, and would have to try both types to see which results in a curve with integral coefficients. If both types succeeded (as does happen), we would only know the curve E_f up to a 2-isogeny.

2.12 Computing periods III: Evaluation of the sums

The results of the previous two sections express the periods of a rational newform $f(z) = \sum a(n,f)\exp(2\pi inz)$, and the value $L(f,1)$, in terms of various infinite series, each of the form $\sum a(n,f)c(n)$. In each case the factor $c(n)$ is a simple function of n, but the coefficient $a(n,f)$ must be computed more indirectly from the $a(p,f)$ for prime p as in Section 2.9.

In practice we will know $a(p,f)$ for the first few primes, say $p \le$ `pmax`. An elegant and efficient recursive procedure for summing a series of the form $\sum a(n)c(n)$ over

$$\{n : 1 \le n \le \texttt{nmax}, \text{ and } p|n \Rightarrow p \le \texttt{pmax}\},$$

with $a(n)$ defined in a similar recursive manner, was described in [5, pages 27–28]. This method has the advantage of minimizing the number of multiplications involved and the number of $a(n)$ which must be stored. Also, if some $a(n) = 0$ then there is a whole class of integers m for which $a(m) = 0$ that the procedure avoids automatically. Although in our program this part of the computation was not critical for either time or storage space, we found this algorithm to be very useful. It may also be applied in other similar situations for other kinds of modular forms: we have ourselves used it in [14], with cusp forms of weight 2 for $\Gamma_1(N)$, and also in our work over imaginary quadratic fields.

To evaluate such a sum, assume that the array `p[i]` hold the first `pmax` primes p_i, and that the array `ap[i]` holds the coefficients `ap[i]` $= a(p_i)$ for $p_i \le$ `pmax`. We can evaluate the sum $a(n)c(n)$ over all $n \le$ `nmax` all of whose prime divisors are less than or equal to `pmax` with the following pseudo-code.

Algorithm for recursively computing a multiplicative sum

```
1.  BEGIN
2.  Sum = c(1);
3.  FOR i WHILE p[i]  ≤  pmax DO
4.  BEGIN
5.       add(p[i],i,ap[i],1)
6.  END
7.  END
```

(Subroutine to add the terms dependent on p)

```
subroutine add(n,i,a,last_a)
 1.  BEGIN
 2.  IF a=0 THEN j0 = i ELSE Sum = Sum + a*c(n); j0 = 1 FI;
 3.  FOR j FROM j0 TO i WHILE p[j]*n ≤ nmax DO
 4.  BEGIN
 5.       next_a = a*ap[j];
 6.       IF j=i AND (N ≢ 0 (mod p[j])) THEN
 7.            next_a = next_a - p[j]*last_a
 8.       FI;
 9.       add(p[j]*n,j,next_a,a)
10.  END
11.  END
```

Here the recursive function `add(n,i,a,last_a)` is always called under the following conditions: (i) $p_i =$ `p[i]` is the smallest prime dividing $n =$ `n`; (ii) `a` $= a(n)$; (iii) `last_a` $= a(n/p_i)$. The procedure for n calls itself with pn in place of n, for all primes $p \le p_i$, having first computed `next_a` $= a(pn)$ using the recurrence formulae from Section 2.6; if $a(n) = 0$ then only $p = p_i$ need be used, since then $a(pn) = a(p)a(n) = 0$ for all $p < p_i$.

2.13 Computing $L^{(r)}(f, 1)$

In investigating the Birch–Swinnerton-Dyer conjecture for the modular curves E_f we will need to compute the numerical value of the rth derivative $L^{(r)}(E_f, 1) = L^{(r)}(f, 1)$, where r is the order of $L(f, s)$ at $s = 1$. This integer r is sometimes called the 'analytic rank' of the curve E_f, since it is also, according the the Birch–Swinnerton-Dyer conjecture, the rank of $E_f(\mathbb{Q})$. Following earlier work with examples of rank 0 and 1, this computation was carried out by Buhler, Gross and Zagier in [6], for the curve of conductor 5077 and rank 3. Their method works in general, and we describe it here.

Recall the definition of $\Lambda(f, s)$ from Section 2.8:

$$\Lambda(f,s) = N^{s/2}(2\pi)^{-s}\Gamma(s)L(f,s) = \int_0^\infty f(iy/\sqrt{N})y^{s-1}dy. \tag{2.8.5}$$

Let the W_N-eigenvalue of f be ε. Using $f(-1/Nz) = \varepsilon N z^2 f(z)$ we obtain

$$\Lambda(f,s) = \int_1^\infty f(iy/\sqrt{N})\left(y^{s-1} - \varepsilon y^{1-s}\right)dy$$

(from which the functional equation (2.8.6) follows immediately). Differentiating k times with respect to s gives

$$\Lambda^{(k)}(f,s) = \int_1^\infty f(iy/\sqrt{N})(\log y)^k\left(y^{s-1} - \varepsilon(-1)^k y^{1-s}\right)dy,$$

so at $s = 1$ we have

$$\Lambda^{(k)}(f,1) = (1 - (-1)^k\varepsilon)\int_1^\infty f(iy/\sqrt{N})(\log y)^k dy.$$

Trivially this gives $\Lambda^{(k)}(f,1)=0$ if $\varepsilon=(-1)^k$. In particular, since $\Lambda^{(r)}(f,1)\neq 0$, by definition of r, we must have $(-1)^r=-\varepsilon$ so that r is even if and only if $\varepsilon=-1$. Hence setting $k=r$, we have

$$\begin{aligned}\Lambda^{(r)}(f,1) &= 2\int_1^\infty f(iy/\sqrt{N})(\log y)^r dy \\ &= 2\sum_{n=1}^\infty a(n,f)\int_1^\infty \exp(-2\pi ny/\sqrt{N})(\log y)^r dy. \qquad (2.13.1)\end{aligned}$$

If $r=0$, of course, we recover the formula

$$\Lambda(f,1)=\frac{\sqrt{N}}{\pi}\sum_{n=1}^\infty \frac{a(n,f)}{n}\exp(-2\pi n/\sqrt{N})$$

which agrees with (2.11.2) since $\Lambda(f,1)=(\sqrt{N}/2\pi)L(f,1)$. Now assume that $r\geq 1$. Integrating (2.13.1) by parts gives

$$\Lambda^{(r)}(f,1)=\frac{r\sqrt{N}}{\pi}\sum_{n=1}^\infty \frac{a(n,f)}{n}\int_1^\infty \exp(-2\pi ny/\sqrt{N})(\log y)^{r-1}\frac{dy}{y}.$$

Since $\Lambda(f,s)$ vanishes to order r at $s=1$ we have $L^{(r)}(f,1)=(2\pi/\sqrt{N})\Lambda^{(r)}(f,1)$, and hence the following result.

PROPOSITION 2.13.1. *Let f be a newform in $S_2(N)$ with W_N-eigenvalue ε, and suppose that the order of $L(f,s)$ at $s=1$ is at least r, where $\varepsilon=(-1)^{r-1}$. Then*

$$L^{(r)}(f,1)=2r!\sum_{n=1}^\infty \frac{a(n,f)}{n}G_r\left(\frac{2\pi n}{\sqrt{N}}\right) \qquad (2.13.2)$$

where

$$G_r(x)=\frac{1}{(r-1)!}\int_1^\infty e^{-xy}(\log y)^{r-1}\frac{dy}{y}.$$

In order to evaluate the series in (2.13.2) we may use the summation procedure of the preceding section, provided that we are able to compute the function $G_r(x)$. When $r=1$, $G_1(x)$ is the exponential integral $\int_1^\infty e^{-xy}dy/y$, which may be evaluated for small x (say $x<3$) by the power series

$$G_1(x)=\left(\log\frac{1}{x}-\gamma\right)-\sum_{n=1}^\infty\frac{(-x)^n}{n\cdot n!} \qquad (2.13.3)$$

where γ is Euler's constant $0.577\ldots$. For larger x (say $x>2$) it is better to use the continued fraction expansion

$$G_1(x)=\cfrac{e^{-x}}{x+\cfrac{1}{1+\cfrac{1}{x+\cfrac{2}{1+\cfrac{2}{x+\cfrac{3}{1+\ldots}}}}}}.$$

To generalize the series (2.13.3) for $G_1(x)$, we observe that the functions $G_r(x)$ satisfy the functional equations $G_r'(x) = (-1/x)G_{r-1}(x)$, with $G_0(x) = e^{-x}$. It follows that

$$G_r(x) = P_r\left(\log\frac{1}{x}\right) + \sum_{n=1}^{\infty}\frac{(-1)^{n-r}}{n^r\cdot n!}x^n$$

where $P_r(t)$ is a polynomial of degree r satisfying $P_r'(t) = P_{r-1}(t)$ and $P_0(t) = 0$. From our earlier expression for $G_1(x)$ we see that $P_1(t) = t - \gamma$. In general $P_r(t) = Q_r(t-\gamma)$ where

$$\begin{aligned}
Q_1(t) &= t;\\
Q_2(t) &= \frac{1}{2}t^2 + \frac{\pi^2}{12};\\
Q_3(t) &= \frac{1}{6}t^3 + \frac{\pi^2}{12}t - \frac{\zeta(3)}{3};\\
Q_4(t) &= \frac{1}{24}t^4 + \frac{\pi^2}{24}t^2 - \frac{\zeta(3)}{3}t + \frac{\pi^4}{160};\\
Q_5(t) &= \frac{1}{120}t^5 + \frac{\pi^2}{72}t^3 - \frac{\zeta(3)}{6}t^2 + \frac{\pi^4}{160}t - \frac{\zeta(5)}{5} - \frac{\zeta(3)\pi^2}{36}.
\end{aligned}$$

For $N < 5077$ we always found that $r \le 2$, and determining the value of r in such cases is easy. Certainly $r = 0$ if and only if $L(f,1) \ne 0$, which can be determined algebraically by (2.8.10). When $L(f,1) = 0$ and $\varepsilon = +1$ we know that r is odd; by computing $L'(f,1)$ to sufficient precision using (2.13.2) we could verify that $L'(f,1) \ne 0$, so that $r = 1$. Similarly, when $L(f,1) = 0$ and $\varepsilon = -1$, we know that r is even and at least 2, and we could check that $r = 2$ by computing $L''(f,1)$ to sufficient precision to be certain that $L''(f,1) \ne 0$.

In higher rank cases we have the problem of deciding whether $L^{(k)}(f,1) = 0$, since no approximate calculation can by itself determine this. The first case where this occurs is for $N = 5077$, the rank 3 case considered in [6]. Here one finds that $L'(f,1) = 0$ to 13 decimal places using (2.13.2) with 250 terms; then it is possible to conclude that $L'(f,1) = 0$ exactly, by applying the theorem of Gross and Zagier concerning modular elliptic curves of rank 1 (see [25] or [26]) which relates the value of $L'(f,1)$ to the height of a certain Heegner point on E_f. In this case no point on E_f has sufficiently small positive height, and one can therefore deduce that $L'(f,1) = 0$, so that $r \ge 3$. Finally the value of $L^{(3)}(f,1)$ can be computed numerically and hence shown to be non-zero (approximately 1.73 in this case). See [6] for more details. Using more recent work of Kolyvagin (see [29]) this argument can be simplified, since it is now known that when $L(f,s)$ has a simple zero at $s = 1$, the curve E_f has rank exactly 1. But in this case E_f has rank 3 (computed via two-descent, though finding three independent points of infinite order is easy and shows that the rank is at least 3), so again the analytic rank must be at least 3, and is therefore exactly 3 as before.

The results of Kolyvagin in [29] imply that when $L(f,s)$ has a zero of order $r = 0$ or 1 at $s = 1$ then[3] the rank of E_f is exactly r. For the tables we also verified that the rank of $E_f(\mathbb{Q})$ was r directly in almost all cases (the exceptions being curves where the coefficients were so large that the two-descent algorithm, described in the next chapter, would have taken too long to run). These results apply to all but 18 of the rational newforms f we found at levels up to 1000. The remaining cases all had $r = 2$ (determined as above) and we verified that the rank of $E_f(\mathbb{Q})$ was 2 in each case. For a summary of the ranks found in the extended computations to $N = 5077$, see Chapter IV.

[3]In fact, Kolyvagin's result in the rank 0 case was conditional on a certain technical hypothesis, which was later proved independently by Murty and Murty and by Bump, Friedberg and Hoffstein. See [44]. The analogous hypothesis in the rank 1 case was already known as a consequence of a theorem of Waldspurger. The rank 0 result was previously proved in the case of complex multiplication by Coates and Wiles.

2.14 Obtaining equations for the curves

So far we have described how to compute, to a certain precision, the periods ω_1 and ω_2 which generate the period lattice Λ_f of the modular curve $E_f = \mathbb{C}/\Lambda_f$ attached to each rational newform f in $S_2(N)$. Now we turn to the question of finding an equation for E_f.

Set $\tau = \omega_1/\omega_2$. Interchanging ω_1 and ω_2 if necessary, we may assume that $\mathrm{Im}(\tau) > 0$. By applying the well-known algorithm for moving a point in the upper half-plane $\mathcal{H}$ into the usual fundamental region for $\mathrm{SL}(2,\mathbb{Z})$ we may assume that $|\mathrm{Re}(\tau)| \leq 1/2$ and $|\tau| \geq 1$, so that $\mathrm{Im}(\tau) \geq \sqrt{3}/2$. One merely replaces (ω_1, ω_2) by $(\omega_1 - n\omega_2, \omega_2)$ for suitable $n \in \mathbb{Z}$ and (ω_1, ω_2) by $(-\omega_2, \omega_1)$ until both conditions are satisfied. In practice one must be careful about rounding errors, as it is quite possible to have both $|\tau| < 1$ and $|-1/\tau| < 1$ after rounding, which is liable to prevent the algorithm from terminating.

Set $q = \exp(2\pi i\tau)$. Then the lattice invariants $c_4 (= 12g_2)$ and $c_6 (= 216g_3)$ are given by

$$c_4 = \left(\frac{2\pi}{\omega_2}\right)^4 \left(1 + 240\sum_{n=1}^{\infty} \frac{n^3 q^n}{1-q^n}\right) \quad \text{and} \quad c_6 = \left(\frac{2\pi}{\omega_2}\right)^6 \left(1 - 504\sum_{n=1}^{\infty} \frac{n^5 q^n}{1-q^n}\right) \tag{2.14.1}$$

(see, for example, [31, p.47]). Since $|q| = \exp(-2\pi\mathrm{Im}(\tau)) \leq \exp(-\pi\sqrt{3}) < 0 \cdot 005$, these series converge extremely rapidly. Thus, assuming that ω_1 and ω_2 are known to sufficient precision, we can compute c_4 and c_6 as precisely as required.

Since E_f is defined over $\mathbb{Q}$, the numbers c_4 and c_6 are rational, but there is no simple reason why they should be integral. Fortunately, a result of Edixhoven (see [21]) states that in fact they are integral. Hence, provided that we have computed the periods and then c_4 and c_6 to sufficient precision, we will be able to recognize the corresponding exact integer values.

This only presents practical difficulties when c_4 and c_6 are large, since standard double precision arithmetic only yields around 16 decimal places. In several cases this means that we can recognize c_4, but the last digit or digits of c_6 are undetermined. One obvious way round these difficulties is to use multiprecision arithmetic, though the resulting programs are slower, which can be an important consideration when large numbers of curves are being processed. In these situations, we are helped by the fact that we know that c_4 and c_6 are the invariants of an elliptic curve of conductor N. This implies the following congruence conditions (see [30], [10] or Section 3.2 below):

(1) $c_4^3 - c_6^2 = 1728\Delta$, where Δ is a non-zero integer divisible by the primes dividing N;
(2) for primes $p \geq 5$ which divide N, we have $p \mid c_4 \iff p \mid c_6 \iff p^2 \mid N$;
(3) $c_6 \not\equiv 9 \pmod{27}$;
(4) either $c_6 \equiv -1 \pmod 4$, or $c_4 \equiv 0 \pmod{16}$ and $c_6 \equiv 0, 8 \pmod{32}$.

Note that in condition (1) we should not assume that Δ is only divisible by the "bad primes" which divide N, since we do not know that c_4 and c_6 are the invariants of a minimal model. However, Edixhoven's result does bound the non-minimality, and in practice all the equations of curves we have constructed are minimal, verifying the conjecture (Manin's "$c = 1$" conjecture) that this should always be the case. Conditions (2)–(4) do assume minimality at the relevant primes.

Since c_4 tends to be smaller than c_6, the common situation is that we know c_4, but may need to use the above congruence conditions to help us find c_6 in case it has more than 16 digits.

Given integral invariants c_4, c_6 satisfying (1), (3) and (4) above, the coefficients of a standard Weierstrass equation for the curve may be obtained as follows (see Section 3.1), where all divisions are exact:

$$
\begin{aligned}
b_2 &= -c_6 \mod 12 \in \{-5, \ldots, 6\};\\
b_4 &= (b_2^2 - c_4)/24;\\
b_6 &= (-b_2^3 + 36 b_2 b_4 - c_6)/216;\\
a_1 &= b_2 \mod 2 \in \{0, 1\};\\
a_3 &= b_6 \mod 2 \in \{0, 1\};\\
a_2 &= (b_2 - a_1)/4;\\
a_4 &= (b_4 - a_1 a_3)/2;\\
a_6 &= (b_6 - a_3)/4.
\end{aligned}
$$

Having the coefficients $[a_1, a_2, a_3, a_4, a_6]$ of a curve E, we may apply Tate's algorithm (see Section 3.2 below) to check that E has conductor N. We also check whether this model for E is minimal. These conditions do hold for all the cases we have computed to date ($N \leq 5077$). We also verify in each case that the traces of Frobenius of E_f and E for all primes under 1000 agree in each case, and in nearly all cases (see the previous section) that the rank of E, computed via two-descent, agrees with the 'analytic rank' of E_f. Finally, we can compute all curves isogenous to E over $\mathbb{Q}$: see Section 3.8 for one way to do this. This final list of curves will, according to the Shimura–Taniyama–Weil conjectures, contain all elliptic curves defined over $\mathbb{Q}$ with conductor N (up to isomorphism). At the time of writing[4] this has been proved (by Wiles and Taylor, following Ribet, Frey and others) provided that N is divisible by neither 4 nor 25.

Our computations do not give any verification of the Shimura–Taniyama–Weil conjecture, since if there did exist elliptic curves over $\mathbb{Q}$ which were not modular, then we would simply not find them. We could only verify the conjecture if we had an independent method for listing all curves of conductor N, up to isogeny. For example, this has been done

- when N is a power of 2 (Ogg) or of the form $2^a 3^b$ (Coghlan, see [2, Table 4]);
- when $N = 11$ (by Agrawal, Coates, Hunt and Van der Poorten, using the theory of Baker; and independently by Serre, using a variant of Faltings's method based on quartic fields [54]);
- for certain prime values of N (see [4]).

Our results are compatible with those of Brumer and Kramer in [4] for curves of prime conductor under 1000.

The algorithms we used to study these curves E further will be the subject of the next chapter.

2.15 Computing the degree of a modular parametrization

The modular elliptic curves we have shown how to construct in this chapter can be parametrized by modular functions for the subgroup $\Gamma_0(N)$ of the modular group $\Gamma = \mathrm{PSL}(2, \mathbb{Z})$. Equivalently, there is a non-constant map φ from the modular curve $X_0(N)$ to E. In the paper [17], we presented a method of computing the degree of such a map φ for arbitrary N. Our method is derived from a method of Zagier in [69]; by using those ideas, together with the modular symbol and M-symbol techniques which have been used above, we are able to derive an explicit formula for $\deg(\varphi)$ which is in general much simpler to implement than Zagier's, for arbitrary subgroups of finite index in Γ. To implement this formula one needs to have explicit coset representatives for the subgroup, but it is not necessary to determine an explicit fundamental domain for its action on the upper half-plane $\mathcal{H}$. In particular, it is

[4] October 1996

simple to implement for $\Gamma_0(N)$ for arbitrary N, in contrast with Zagier's formula which is only completely explicit for N prime.

In this section we present the algorithm described in [17]. For more details and proofs, see [17]. Worked examples are given in the appendix to this chapter, and results for $N \leq 1000$ may be found in Chapter IV.

2.15.1. Modular Parametrizations.

Let G be a congruence subgroup of the modular group $\Gamma = \mathrm{PSL}(2,\mathbb{Z})$. The quotient $X = X_G = G\backslash\mathcal{H}^*$ is a Riemann surface, and an algebraic curve defined over a number field, and is called a modular curve.

An elliptic curve E defined over $\mathbb{Q}$ is called a modular elliptic curve if there is a non-constant map $\varphi\colon X_G \to E$ for some modular curve X_G. The pull-back of the (unique up to scalar multiplication) holomorphic differential on E is then of the form $2\pi i f(z)dz$, where $f \in S_2(G)$. According to the Shimura–Taniyama–Weil conjecture, this should be the case for every elliptic curve defined over $\mathbb{Q}$, with $G = \Gamma_0(N)$, where N is the conductor of E. Moreover, the cusp form f should be a newform in the usual sense.

We will suppose that we are given a cusp form $f \in S_2(G)$. Since the differential $f(z)dz$ is holomorphic, the function

$$z_0 \mapsto I_f(z_0) = 2\pi i \int_{z_0}^{\infty} f(z)dz$$

is well-defined for $z_0 \in \mathcal{H}^*$ (independent of the path from z_0 to ∞). Also, for $M \in G$, the function

$$M \mapsto P_f(M) = I_f(z_0) - I_f(M(z_0)) = 2\pi i \int_{z_0}^{M(z_0)} f(z)dz$$

is independent of z_0, and defines a function $P_f\colon G \to \mathbb{C}$ which is a group homomorphism. The image Λ_f of this map will, under suitable hypotheses on f which we will assume to hold, be a lattice of rank 2 in $\mathbb{C}$, so that $E_f = \mathbb{C}/\Lambda_f$ is an elliptic curve. Hence I_f induces a map

$$\begin{aligned} \varphi\colon X = G\backslash\mathcal{H}^* &\to E_f = \mathbb{C}/\Lambda_f \\ z \mod G &\mapsto I_f(z) \mod \Lambda_f. \end{aligned}$$

The period map $P_f\colon G \to \Lambda_f$ is surjective (by definition) and its kernel contains all elliptic and parabolic elements of G. We may write $\Lambda_f = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$ with $\mathrm{Im}(\omega_2/\omega_1) > 0$. Then

$$P_f(M) = n_1(M)\omega_1 + n_2(M)\omega_2$$

where $n_1, n_2\colon G \to \mathbb{Z}$ are homomorphisms. These functions are explicitly computable in terms of modular symbols as seen in earlier sections. Alternatively, given sufficiently many Fourier coefficients of the cusp form $f(z)$ we may evaluate the period integrals $I_f(z)$ (using the formula (2.10.8), for example) to sufficient precision that (assuming that the fundamental periods ω_1 and ω_2 are also known to some precision) one can determine the integer values of $n_1(M)$ and $n_2(M)$ for any given $M \in G$. The latter approach is used in [69]. The advantage of the modular symbol approach is that exact values are obtained directly, and that it is not necessary to compute (or even know) any Fourier coefficients of f. On the other hand, it becomes computationally infeasible to carry out the modular symbol computations when the index of G in Γ is too large, whereas the approximate approach can still be used, provided that one has an explicit equation for the curve E to hand, from which one can compute the periods and the Fourier coefficients in terms of traces of Frobenius (assuming that E is modular and defined over $\mathbb{Q}$). This method was used, for example, to compute $\deg(\varphi)$ for the curve of rank 3

with conductor 5077, in [69]; we verified the value obtained (namely 1984) using our modular symbol implementation.

The special case we are particularly interested in is where $G = \Gamma_0(N)$ and $f(z)$ is a normalized newform for $\Gamma_0(N)$. Then the periods of $2\pi i f(z)$ do form a suitable lattice Λ_f, and the modular elliptic curve $E_f = \mathbb{C}/\Lambda_f$ is defined over $\mathbb{Q}$ and has conductor N.

In order to compute the degree of the map $\varphi\colon X \to E_f$, the idea used in [69] is to compute the Petersson norm $||f||$ in two ways. The first way involves $\deg(\varphi)$ explicitly, while the second expresses it as a sum of terms involving periods, which can be evaluated as above.

PROPOSITION 2.15.1. *Let $f(z)$ be a cusp form of weight 2 for G as above, and $\varphi\colon X \to E_f$ the associated modular parametrization. Then*

$$4\pi^2||f||^2 = \deg(\varphi)\mathrm{Vol}(E_f).$$

REMARK. In terms of the fundamental periods ω_1, ω_2 of E_f, the volume is given by $\mathrm{Vol}(E_f) = |\mathrm{Im}\,(\overline{\omega_1}\omega_2)|$. More generally, if ω, $\omega' \in \Lambda_f$, with $\omega = n_1(\omega)\omega_1 + n_2(\omega)\omega_2$ and $\omega' = n_1(\omega')\omega_1 + n_2(\omega')\omega_2$, then (up to sign) we have

$$\mathrm{Im}\,(\overline{\omega}\omega') = \mathrm{Vol}(E_f) \cdot \begin{vmatrix} n_1(\omega) & n_1(\omega') \\ n_2(\omega) & n_2(\omega') \end{vmatrix}.$$

2.15.2. Coset representatives and Fundamental Domains.

Let $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ and $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ be the usual generators for Γ, so that S has order 2 and TS has order 3. Let $\mathcal{F}$ be the usual fundamental domain for Γ defined above in (2.1.1), and $\mathcal{T}$ the "ideal triangle" with vertices at 0, 1 and ∞. Recall from Section 2.1 that $\langle M \rangle$ denotes the transform of $\mathcal{T}$ by M for $M \in \Gamma$, which is the ideal triangle with vertices at the cusps $M(0)$, $M(1)$ and $M(\infty)$. These triangles form a triangulation of the upper half-plane $\mathcal{H}$, whose vertices are precisely the cusps $\mathbb{Q} \cup \{\infty\}$. Recall that

$$\langle M \rangle = \langle MTS \rangle = \left\langle M(TS)^2 \right\rangle$$

but that otherwise the triangles are distinct. The triangle $\langle M \rangle$ has three (oriented) edges; these are the modular symbols (M), (MTS) and $(M(TS)^2)$.

Assume, for simplicity, that G has no non-trivial elements of finite order, *i.e.*, no conjugates of either S or TS. (This assumption is merely for ease of exposition; in fact, it is easy to see that elliptic elements of G contribute nothing to the formula in Theorem 2.15.4 below in any case.) Choose, once and for all, a set $\mathcal{S}$ of right coset representatives for G in Γ, such that $M \in \mathcal{S} \Rightarrow MTS \in \mathcal{S}$; this is possible since, by hypothesis, G contains no conjugates of TS.

Let $\mathcal{S}'$ be a subset of $\mathcal{S}$ which contains exactly one of each triple M, MTS, $M(TS)^2$, so that $\mathcal{S} = \mathcal{S}' \cup \mathcal{S}'TS \cup \mathcal{S}'(TS)^2$. Then a fundamental domain for the action of G on $\mathcal{H}$ is given by

$$\mathcal{F}_G = \bigcup_{M \in \mathcal{S}'} \langle M \rangle .$$

In general, this set need not be connected, but this does not matter for our purposes: it can be treated as a disjoint union of triangles, whose total boundary is the sum of the oriented edges (M) for $M \in \mathcal{S}$.

The key idea in the algebraic reformulation of Zagier's method is to make use of the coset action of Γ on the set $\mathcal{S}$. We now introduce notation for the actions of the generators S and T.

Action of S. For each $M \in \mathcal{S}$ we set $MS = s(M)\sigma(M)$, where $s\colon \mathcal{S} \to G$ is a function and $\sigma\colon \mathcal{S} \to \mathcal{S}$ is a permutation. Since S^2 is the identity, the same is true of σ, and $s(\sigma(M)) = s(M)^{-1}$. For brevity we will write $M^* = \sigma(M)$, so that $M^{**} = M$ for all $M \in \mathcal{S}$. (This conflicts with an earlier use of the notation M^* in Section 2.1, but this should not cause confusion.)

Note that the triangles $\langle M\rangle$ and $\langle MS\rangle$ are adjacent in the triangulation of $\mathcal{H}$, since they share the common side $(M) = \{M(0), M(\infty)\} = -(MS)$. However, since in general we do not have $MS \in \mathcal{S}$, in the fundamental domain $\mathcal{F}_G$ for G it is the triangles $\langle M\rangle$ and $\langle M^*\rangle$ which are glued together by the element $s(M) \in G$ which takes (M^*) to $-(M)$ (the orientation is reversed).

Action of T. Similarly, for $M \in \mathcal{S}$ we set $MT = t(M)\tau(M)$ with $t(M) \in G$ and $\tau(M) \in \mathcal{S}$. The permutation τ of $\mathcal{S}$ plays a vital part in what follows. The following lemma will not be used later, but is included for its own interest as it explains the geometric significance of this algebraic permutation.

Lemma 2.15.2.
(a) *Two elements M and M' of $\mathcal{S}$ are in the same τ-orbit if and only if the cusps $M(\infty)$ and $M'(\infty)$ are G-equivalent; hence the number of τ-orbits on $\mathcal{S}$ is the number of G-equivalence classes of cusps.*
(b) *The length of the τ-orbit containing $M \in \mathcal{S}$ is the width of the cusp $M(\infty)$ of G.*

Proof. (a) M and M' are in the same τ-orbit if and only if $M_0 = M'T^jM^{-1} \in G$ for some j, which is if and only if $M_0M(\infty) = M'(\infty)$, since the stabilizer of ∞ in Γ is the subgroup generated by T.

(b) The length of the orbit of M is the least $k > 0$ such that $MT^kM^{-1} = \left(MTM^{-1}\right)^k \in G$, which is the width of the cusp $M(\infty)$, since the stabilizer of $M(\infty)$ in Γ is generated by MTM^{-1}. □

Thus there is a one–one correspondence between the orbits of τ on $\mathcal{S}$ and the classes of G-inequivalent cusps, with the length of each orbit being the width of the corresponding cusp.

In each τ-orbit in $\mathcal{S}$, we choose an arbitrary base-point M_1, and set $M_{j+1} = \tau(M_j)$ for $1 \le j \le k$, where k is the length of the orbit and $M_{k+1} = M_1$. Thus $M_jT = t(M_j)M_{j+1}$, so that

$$M_1T^j = t(M_1)t(M_2)\dots t(M_j)M_{j+1}.$$

In particular, $M_1T^k = M_0M_1$, where $M_0 = t(M_1)t(M_2)\dots t(M_k) \in G$. Since M_0 is parabolic and P_f is a homomorphism, we obtain the following.

Lemma 2.15.3.

$$\sum_{j=1}^{k} P_f(t(M_j)) = 0.$$

Write $M \prec M'$ if M and M' are in the same τ-orbit in $\mathcal{S}$, and M precedes M' in the fixed ordering determined by choosing a base-point for each orbit. In the notation above, $M \prec M'$ if and only if $M = M_i$ and $M' = M_j$ where $1 \le i < j \le k$.

We can now state the main results of this section.

Theorem 2.15.4. *Let f be a cusp form of weight 2 for G with associated period function $P_f : G \to \mathbb{C}$. Then (the square of) the Petersson norm of f is given by*

$$||f||^2 = \frac{1}{8\pi^2} \sum_{M \prec M'} \mathrm{Im}(P_f(t(M))\overline{P_f(t(M'))}).$$

Here the sum is over all ordered pairs $M \prec M'$ in $\mathcal{S}$ which are in the same orbit of the permutation τ of $\mathcal{S}$ induced by right multiplication by T.

Combining this result with Proposition 2.15.1, we immediately obtain our explicit formula for the degree of the modular parametrization.

THEOREM 2.15.5. *With the above notation,*

$$\deg(\varphi) = \frac{1}{2\mathrm{Vol}(E_f)} \sum_{M \prec M'} \mathrm{Im}(\overline{P_f(t(M))} P_f(t(M'))) = \frac{1}{2} \sum_{M \prec M'} \begin{vmatrix} n_1(t(M)) & n_1(t(M')) \\ n_2(t(M)) & n_2(t(M')) \end{vmatrix}.$$

Hence to compute $\deg(\varphi)$, we only have to compute the right coset action of T on an explicit set $\mathcal{S}$ of coset representatives for G in Γ, and evaluate the integer-valued functions n_1 and n_2 on each of the matrices $t(M)$ for $M \in \mathcal{S}$. In the case of $\Gamma_0(N)$, these steps can easily be carried out using M-symbols, and we will give some further details below.

REMARKS. 1. The formula given in Theorem 2.15.5 expresses $\deg(\varphi)$ explicitly as a sum which can be grouped as a sum of terms, one term for each cusp, by collecting together the terms for each τ-orbit. It is not at all clear what significance, if any, can be given to the individual contributions of each cusp to the total.

2. The form of our formula is identical to the one in [69]. However, we stress that in [69], the analogue of our coset action τ is defined not algebraically, as here, but geometrically, as a permutation of the edges of a fundamental polygonal domain for G (and dependent on the particular fundamental domain used). Then it becomes necessary to have an explicit picture of such a fundamental domain, including explicit matrices which identify the edges of the domain in pairs. This is only carried out explicitly in [69] in the case $G = \Gamma_0(N)$ where N is a prime. In our formulation, the details are all algebraic rather than geometric, which makes the evaluation of the formula more practical to implement. Also, we have the possibility of evaluating the functions n_1 and n_2 exactly using modular symbols, instead of using numerical evaluation of the periods, which reduces the computation of $\deg(\varphi)$ entirely to linear algebra and integer arithmetic.

2.15.3. Implementation for $\Gamma_0(N)$.

We now discuss the case $G = \Gamma_0(N)$ in greater detail, using M-symbols to represent the coset representatives. The right coset action of Γ on $P^1(N)$ is given by (2.2.4), so we have $\sigma(c:d) = (c:d)S = (d:-c)$ and $\tau(c:d) = (c:d)T = (c:c+d)$.

LEMMA 2.15.6. *The length of the τ-orbit of $(c:d)$ is $N/\gcd(N, c^2)$.*

PROOF. $\tau^k(c:d) = (c:d) \iff (c:kc+d) = (c:d) \iff cd \equiv c(kc+d) \pmod N \iff kc^2 \equiv 0 \pmod N \iff k \equiv 0 \pmod{N/\gcd(N,c^2)}$. □

In earlier sections, it was immaterial exactly which coset representatives were used, or in practice which pair $(c,d) \in \mathbb{Z}^2$ was used to represent the M-symbol $(c:d)$. For the application of Theorem 2.15.5, however, we must ensure that our set is closed under right multiplication by TS, where $(c:d)TS = (c+d:-c)$, unless $(c:d)$ is fixed by TS, which is if and only if $c^2 + cd + d^2 \equiv 0 \pmod N$. Thus each M-symbol $(c:d)$ will be represented by a specific pair $(c,d) \in \mathbb{Z}^2$ with $\gcd(c,d) = 1$, in such a way that our set $\mathcal{S}$ of representatives contains the pairs $(c+d,-c)$ and $(-d,c+d)$ whenever it contains (c,d), unless $(c:d)$ is fixed by TS. (Even when working with pairs $(c,d) \in \mathbb{Z}^2$ we will identify (c,d) and $(-c,-d)$.)

Fixing these triples of pairs (c,d) corresponds to fixing the triangles $\langle M \rangle$ which form a (possibly disconnected) fundamental domain for $\Gamma_0(N)$. If $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, the pair (c,d)

corresponds to the directed edge $\{M(0), M(\infty)\} = \{b/d, a/c\}$. For this reason, we will refer to the pairs (c, d) as edges, and the triples of pairs as triangles. Right multiplication by TS corresponds geometrically to moving round to the next edge of the triangle, while right multiplication by S corresponds to moving across to the next triangle $\langle M^* \rangle$ adjacent to the current one. The τ-action is given by composing these, taking $(c : d)$ (or edge $\{b/d, a/c\}$) to the symbol $(c : d)T = (c : c + d)$ with corresponding edge $\{(a + b)/(c + d), a/c\}$, up to translation by an element of $\Gamma_0(N)$. Note how in this operation the endpoint at the cusp $M(\infty) = a/c$ is fixed, in accordance with Lemma 2.15.2 above.

We may therefore proceed as follows. For each orbit, start with a standard pair (c, d), chosen in an M-symbol class $(c : d)$ not yet handled. Apply T to obtain the pair $(c, c + d)$. If this pair is the standard representative for the class $(c : c + d)$, we need take no action and may continue with the orbit. But if $(c, c + d) \equiv (r, s)$, say, with $(r, s) \in \mathcal{S}$, then we must record the "gluing matrix"

$$M = \begin{pmatrix} a & a+b \\ c & c+d \end{pmatrix} \begin{pmatrix} p & q \\ r & s \end{pmatrix}^{-1} \in \Gamma_0(N),$$

where $ad - bc = ps - qr = 1$, whose period $P_f(M)$ will contribute to the partial sum for this orbit. When this happens, we say that the orbit has a "jump" at this point. Different choices for a, b, p and q only change M by parabolic elements, and so do not affect the period $P_f(M)$. We continue until we return to the starting pair, and then move to another orbit, until all M-symbols have been used. As checks on the computation we may use Lemmas 2.15.2 and 2.15.6: the length of the orbit starting at (c, d) can be precomputed as $N/\gcd(N, c^2)$, and the number of orbits is the number of $\Gamma_0(N)$-inequivalent cusps.

A worked example for the case $N = 11$ is included in the appendix to this chapter.