Canadian Mathematical Society
Conference Proceedings
Volume 17, 1995

Modular forms and modular curves

Fred Diamond and John Im

1. Introduction

The theory of modular forms has its roots in the work of 19th century mathematicians including Jacobi and Eisenstein. In the 1920's and 30's much of the foundation for the modern theory was created by Hecke [**Hec1**], [**Hec2**], [**Hec3**]. In addition to establishing the analytic continuation and functional equation for L-functions associated to modular forms, he showed that for a special class of modular forms, the L-functions have Euler product expressions. This special class consists of forms which are simultaneous eigenvectors for certain linear operators, now called Hecke operators.

The work of Eichler and Shimura greatly advanced the role of modular forms and their L-functions in number theory. One achievement [**Shi1**] was the construction of abelian varieties over $\mathbf{Q}$ whose L-functions were those studied by Hecke. Shimura also proposed a partial converse, namely that every elliptic curve over $\mathbf{Q}$ arises this way. This conjecture grew out of an idea of Taniyama [**Shi8**] and became well-known through work of Weil [**Weil**]. A large part of the Shimura-Taniyama-Weil conjecture has now been proved by Wiles [**Wil2**] (see also [**Diam**]), with a key ingredient supplied by the work of Taylor and Wiles [**TaWi**].

In light of the recent work of Wiles, it is evident that two major developments in the theory began to unfold around 1970, building on the work and insight of Shimura.

One of these was the introduction of tools of modern algebraic geometry. Deligne [**Del1**] generalized the Eichler-Shimura construction to higher weight using ℓ-adic cohomology; Deligne-Rapoport [**DeRa**] and Drinfeld [**Drin**] studied the arithmetic of modular curves; the study of congruences between modular forms was placed in the algebraic-geometric context by work of Serre [**Ser2**], Swinnerton-Dyer [**SwDy**] and Katz [**Katz1**]. This development has been a rich source of techniques, results and ideas in the field and figures prominently in Mazur's bounding of the number of rational torsion points on an elliptic curve over $\mathbf{Q}$ [**Maz1**], as well as in the recent work of Ribet [**Rib4**] and Wiles [**Wil2**].

The other development was the beginning of the Langlands program. The work of Jacquet and Langlands [**JaLa**] on automorphic representations placed the theory

1991 *Mathematics Subject Classification.* Primary 11F11; Secondary 11F32, 11G18.
The second author was supported in part by NSERC of Canada.

in a broader context and added insight from representation theory. According to Langlands' conjectures, these representations should correspond in a natural way to algebraic-geometric objects. Roughly speaking, class field theory is the special case of the correspondence for GL_1. The Eichler-Shimura construction and Deligne's generalization provide special cases of one direction for GL_2; special cases of the other direction were established by Langlands [**Lngl2**] and Tunnell [**Tunn**], and now by Wiles [**Wil2**].

This article was intended to be a survey of results on modular forms and modular curves. In our attempt, and failure, to keep the work a reasonable length, we chose to ignore many important aspects of the theory and instead to emphasize those which play a role in the work of Ribet and Wiles. None of the results we present here are ours, and we have no doubt often failed to properly attribute them. We apologize in advance for these and other shortcomings, which are due largely to our ignorance. We can hardly claim to be experts on many of the topics we included; indeed we learned a great deal in preparing this article.

We have aimed the article at advanced or recent graduate students specializing in the field, though we hope that others will find it a useful reference. Parts of the paper vary in the amount of background assumed. Beginning with §8, we usually take for granted graduate courses in number theory and algebraic geometry based for example on the material found in Lang [**Lang1**], Silverman [**Sil1**] and Hartshorne [**Hart**].

The article is divided into three parts.

Part I is a rapid introduction to modular forms, focusing on the theory of Hecke operators and newforms. More detailed treatments of most of the topics we cover can be found in a number of valuable texts, such as those of Shimura [**Shi1**], Lang [**Lang2**], Miyake [**Miy2**], Knapp [**Kna2**] and Hida [**Hida3**].

In Part II, we turn our attention to modular curves. We begin with their description as Riemann surfaces and moduli-theoretic interpretation. Then we go on to explain some of the algebraic geometric methods used to study their arithmetic and that of their Jacobians. Much of the material can be found in Deligne-Rapoport [**DeRa**], but much is scattered in the literature.

Part III returns to the subject of modular forms from a more sophisticated point of view. We first give a brief introduction to modular forms in the context of automorphic representations, mainly following Jacquet and Langlands [**JaLa**]. Then we approach from the perspective of the geometry of modular curves, often following Shimura [**Shi1**] and Deligne-Rapoport [**DeRa**].

ACKNOWLEDGMENTS: Work on the article was begun while one of the authors (FD) was a Ritt Assistant Professor at Columbia University, and continued during visits to the Institute for Advanced Study and Princeton University.

We would like to thank Henri Darmon and Richard Taylor for their many valuable comments and suggestions during the preparation of this article. We have also benefited from conversations and correspondence with Brian Conrad, Bas Edixhoven, Dipendra Prasad, Jacob Sturm and Jacques Tilouine. We are grateful to Kumar Murty for encouraging us to write this survey, and being patient regarding its completion. Finally, we are especially indebted to Goro Shimura and Andrew Wiles for guiding us into this fascinating subject and continuing to be a source of inspiration.

CONTENTS

Part I. Modular forms

2. Definitions and examples

We begin by recalling the definition of a modular form and listing some examples.

2.1. Definitions.

PRIMARY REFERENCES:
[**Shi1**, §2.1], [**Lang2**, §I.2, VII.1], [**Miy2**, §2.1, 4.3], [**Kna2**, §VIII.2, IX.2] and [**Hida3**, §5.1].

Let $\mathfrak{H}$ denote the complex upper half-plane, and $\mathrm{GL}_2^+(\mathbf{R})$ the subgroup of $\mathrm{GL}_2(\mathbf{R})$ consisting of elements with positive determinant. Then $\mathrm{GL}_2^+(\mathbf{R})$ acts on $\mathfrak{H}$ via Mobius transformations. For any integer k, any $\mathbf{C}$-valued function f on $\mathfrak{H}$ and $\alpha \in \mathrm{GL}_2^+(\mathbf{R})$, we define a new function $f|[\alpha]_k$ on $\mathfrak{H}$ by

$$(f|[\alpha]_k)(z) = \det(\alpha)^{k-1}(cz+d)^{-k} f(\alpha(z)), \qquad z \in \mathfrak{H}$$

where $\alpha = \begin{pmatrix} * & * \\ c & d \end{pmatrix}$. A subgroup Γ of $\mathrm{SL}_2(\mathbf{Z})$ is a congruence subgroup if it contains $\Gamma(N)$ for some positive integer N, where

$$\Gamma(N) = \left\{ \gamma \in \mathrm{SL}_2(\mathbf{Z}) \,\middle|\, \gamma \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \pmod{N} \right\};$$

$\Gamma(N)$ itself is called the principal congruence subgroup (of level N). For example,

$$\begin{aligned} \Gamma_0(N) &= \left\{ \gamma \in \mathrm{SL}_2(\mathbf{Z}) \,\middle|\, \gamma \equiv \begin{pmatrix} * & * \\ 0 & * \end{pmatrix} \pmod{N} \right\}, \\ \Gamma_1(N) &= \left\{ \gamma \in \Gamma_0(N) \,\middle|\, \gamma \equiv \begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix} \pmod{N} \right\} \end{aligned}$$

are congruence subgroups of $\mathrm{SL}_2(\mathbf{Z})$ containing $\Gamma(N)$.

Let k be a non-negative integer, and Γ a congruence subgroup. By a modular form of weight k with respect to Γ, we mean a function $f : \mathfrak{H} \to \mathbf{C}$ satisfying

(i) f is holomorphic on $\mathfrak{H}$;
(ii) $f|[\gamma]_k = f$ for all $\gamma \in \Gamma$;
(iii) f is holomorphic at the cusps.

We need to explain (iii). The group Γ contains a matrix $\begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}$ for some positive integer h. Hence $f(z+h) = f(z)$ for all $z \in \mathfrak{H}$, and thus f has a Fourier expansion at ∞ of the form

$$f(z) = \sum_{n=-\infty}^{\infty} a_n q_h^n, \quad q_h = e^{2\pi i z/h}.$$

To say that f is holomorphic (resp. vanishes) at ∞, we must have $a_n = 0$ for all $n < 0$ (resp. $n \leq 0$); this condition is independent of the choice of h. If $\alpha \in \mathrm{SL}_2(\mathbf{Z})$ then $f|[\alpha]_k|[\gamma]_k = f|[\alpha]_k$ for all $\gamma \in \alpha^{-1}\Gamma\alpha$, so that for any $\alpha \in \mathrm{SL}_2(\mathbf{Z})$, $f|[\alpha]_k$ also has a Fourier expansion at ∞. We say that f is holomorphic (resp. vanishes) at the cusps if $f|[\alpha]_k$ is holomorphic (resp. vanishes) at ∞ for all $\alpha \in \mathrm{SL}_2(\mathbf{Z})$.

The space (over $\mathbf{C}$) of all such functions will be denoted $\mathcal{M}_k(\Gamma)$; its dimension is finite for any congruence subgroup Γ of $\mathrm{SL}_2(\mathbf{Z})$. If an element f, in addition to being a modular form, vanishes at (all) the cusps then it is called a cusp form; the space of cusp forms on Γ of weight k will be denoted $\mathcal{S}_k(\Gamma)$. The finite set $\Gamma\backslash(\mathbf{Q} \cup \{\infty\})$ can be viewed as the set of cusps of the modular curve associated to Γ, whence the terminology holomorphic at the cusps and vanishing at the cusps.

We will discuss this in greater detail below in §12.1. We will also return there to the topic of the dimensions of the spaces $\mathcal{M}_k(\Gamma)$ and $\mathcal{S}_k(\Gamma)$.

Now, let $\varepsilon : (\mathbf{Z}/N\mathbf{Z})^\times \to \mathbf{C}^\times$ be a Dirichlet character mod N; we also write ε for the (completely) multiplicative map on $\mathbf{Z}$ where, by convention, $\varepsilon(m) = 0$ for m not prime to N. A modular form of weight k, level N, character ε, or simply of type (k, N, ε), is a modular form of weight k with respect to $\Gamma_1(N)$ which transforms under the bigger group $\Gamma_0(N)$ by the character ε, i.e., it is an element $f \in \mathcal{M}_k(\Gamma_1(N))$ satisfying

$$(f|[\gamma]_k)(z) = \varepsilon(d_\gamma) f(z), \qquad \forall \gamma \in \Gamma_0(N)$$

where d_γ denotes the d-entry of γ. Such a modular form has the q-expansion at ∞ of the form

$$f(z) = \sum_{n=0}^{\infty} a_n q^n, \quad q = e^{2\pi i z} \tag{2.1.1}$$

since $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right) \in \Gamma_1(N)$. The space of all modular forms of type (k, N, ε) is denoted $\mathcal{M}_k(N, \varepsilon)$ or $\mathcal{M}_k(\Gamma_0(N), \varepsilon)$. The Dirichlet character ε mod N is called the Nebentypus of any element in this space.

Equivalently, consider the action of $d \in (\mathbf{Z}/N\mathbf{Z})^\times$ on $\mathcal{M}_k(\Gamma_1(N))$ given by $\langle d \rangle_k : f \mapsto f|[\sigma_d]_k$, where σ_d is any element of $\mathrm{SL}_2(\mathbf{Z})$ such that

$$\sigma_d \equiv \left(\begin{smallmatrix} \bar{d} & 0 \\ 0 & d \end{smallmatrix}\right) \pmod{N}; \tag{2.1.2}$$

here, $\bar{d}$ is the multiplicative inverse of d mod N. The action depends only on d (mod N) and not on the choice of defining matrix σ_d. The space $\mathcal{M}_k(N, \varepsilon)$ is then the ε-eigenspace with respect to this action. In particular, we have a direct sum decomposition

$$\mathcal{M}_k(\Gamma_1(N)) = \bigoplus_{\varepsilon} \mathcal{M}_k(N, \varepsilon)$$

where ε runs over all Dirichlet characters mod N such that $\varepsilon(-1) = (-1)^k$. Letting $\mathcal{S}_k(N, \varepsilon)$ denote the space of cusp forms in $\mathcal{M}_k(N, \varepsilon)$, we obtain a similar decomposition of $\mathcal{S}_k(\Gamma_1(N))$.

2.2. Examples.

PRIMARY REFERENCES:

[**Shi1**, §2.2], [**Ser1**, Ch. VII], [**Kobl**, Ch. III], [**Miy2**, Ch. VII], [**Kna2**, §VIII.2, IX.3] and [**Hida3**, §5.1].

Note that for weight $k = 0$, $\mathcal{M}_0(\Gamma) = \mathbf{C}$ for any congruence subgroup of $\mathrm{SL}_2(\mathbf{Z})$ and $\mathcal{M}_0(N, \varepsilon) = 0$ unless ε is the trivial character. We list here an assortment of examples of modular forms of positive weight.

EXAMPLE 2.2.1. Let k be an even integer > 2. For $z \in \mathfrak{H}$, consider the function

$$G_k(z) = \sum_{(m,n)}{}' \frac{1}{(mz+n)^k} \tag{2.2.1}$$

where $'$ denotes that the sum is over pairs of integers (m, n) not equal to $(0, 0)$. The reader can check that it is a modular form on $\mathrm{SL}_2(\mathbf{Z})$ of weight k, and that its

q-expansion is given by

$$G_k(z) = 2\zeta(k)\big(1 - \frac{2k}{B_k}\sum_{n=1}^{\infty}\sigma_{k-1}(n)q^n\big), \qquad q = e^{2\pi iz}$$

where $\sigma_{k-1}(n) = \sum_{d|n} d^{k-1}$ and B_k are the Bernoulli numbers defined by

$$\frac{te^t}{e^t - 1} = \sum_{k=0}^{\infty} B_k \frac{t^k}{k!};$$

see e.g. [**Ser1**, §VII.4]. Restricting the double sum in (2.2.1) over relatively prime pairs (m, n) we obtain the normalized Eisenstein series

$$E_k(z) = \frac{1}{2}\sum \frac{1}{(mz+n)^k} = 1 - \frac{2k}{B_k}\sum_{n=1}^{\infty}\sigma_{k-1}(n)q^n \tag{2.2.2}$$

where the first sum is over the integers m, n with $(m, n) = 1$.

Before proceeding with more examples, let us introduce some notation and recall a few facts about Dirichlet L-functions and generalized Bernoulli numbers.

For a Dirichlet character ε modulo N its L-function is defined as usual by the analytic continuation of the series

$$L_N(s, \varepsilon) = \sum_{n=1}^{\infty}\varepsilon(n)n^{-s} = \prod_{p\nmid N}\big(1 - \varepsilon(p)p^{-s}\big)^{-1};$$

here the subscript N is written only to emphasize the modulus of the character, and will be dropped if it is clear from the context. If ε mod N is primitive, then its functional equation can be given (e.g. [**Lang2**, §XIV, Theorem 2.2(ii)]) in the form

$$L(1-s, \bar{\varepsilon}) = L(s, \varepsilon)(N/2\pi)^s\Gamma(s)(e^{\pi is/2} + \varepsilon(-1)e^{-\pi is/2})/W(\varepsilon) \tag{2.2.3}$$

where $W(\varepsilon) = \sum_{j=0}^{N-1}\varepsilon(j)e^{2\pi ij/N}$ denotes the Gauss sum of ε.

For a Dirichlet character ε mod N (not necessarily primitive) the generalized Bernoulli numbers $B_{k,\varepsilon}$ are defined by the formula

$$\sum_{a=1}^{N}\varepsilon(a)\frac{te^{at}}{e^{Nt} - 1} = \sum_{k=0}^{\infty}\frac{B_{k,\varepsilon}}{k!}t^k.$$

It is known (e.g. [**Lang2**, §XIV, Theorem 2.3]) that we have $B_{k,\varepsilon} = -kL_N(1-k, \varepsilon)$ for $k \geq 1$.

EXAMPLE 2.2.2. For an integer $k \geq 0$ and a Dirichlet character ε mod N such that $\varepsilon(-1) = (-1)^k$, consider the series (in two variables $z \in \mathfrak{H}$, $s \in \mathbf{C}$)

$$E_{k,N,\varepsilon}(z, s) = \sum_{\gamma\in\Gamma_\infty\backslash\Gamma_0(N)}\bar{\varepsilon}(d_\gamma)j(\gamma, z)^{-k}|j(\gamma, z)|^{-2s} \tag{2.2.4}$$

where we have put $j(\gamma, z) = c_\gamma z + d_\gamma$ for any $\gamma = \begin{pmatrix} * & * \\ c_\gamma & d_\gamma \end{pmatrix} \in \mathrm{GL}_2(\mathbf{R})$ and $\Gamma_\infty = \big\{\pm\begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix} \big| m \in \mathbf{Z}\big\}$ is the stabilizer of ∞ in $\Gamma_0(N)$. This series is uniformly and absolutely convergent for $\Re(2s) \geq 2 - k + \epsilon$ (for any $\epsilon > 0$) and satisfies the transformation property

$$E_{k,N,\varepsilon}(\alpha(z), s) = \varepsilon(d)j(\alpha, z)^k|j(\alpha, z)|^{2s}E_{k,N,\varepsilon}(z, s)$$

under $\alpha = \begin{pmatrix} * & * \\ c & d \end{pmatrix}$ in $\Gamma_0(N)$ (whenever the series converges uniformly). The function

$$\Gamma(s+k)L_N(2s+k,\bar{\varepsilon})E_{k,N,\varepsilon}(z,s)$$

can be continued to a meromorphic function on the whole s-plane which is entire except when $k = 0$ and ε is trivial. If $k = 0$ and $N = 1$, then it is analytic everywhere except for simple poles at $s = 0$ and 1; if $k = 0$, $N > 1$ and ε is trivial, then there is only one simple pole at $s = 1$; see [**Hec1**, §1,2], and also [**Miy2**, §7.2], where the series (2.2.4) is denoted $E^*_{k,N}(z,s;\bar{\varepsilon})$. We discuss its functional equation in Remark 2.2.3 below.

Put, for $k \geq 1$,

$$E_{k,N,\varepsilon}(z) = E_{k,N,\varepsilon}(z,0).$$

Then, as Hecke [**Hec1**] showed, $E_{k,N,\varepsilon}$ belongs to $\mathcal{M}_k(N,\varepsilon)$ except when $k = 2$ and ε is trivial.

In the case that ε is primitive, the Fourier expansion of $E_{k,N,\varepsilon}$ is given by

$$E_{k,N,\varepsilon}(z) = 1 + A^{-1} \cdot \sum_{n=1}^{\infty} \Big(\sum_{d|n} \varepsilon(d) d^{k-1}\Big) q^n, \tag{2.2.5}$$

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

$$A = \frac{L(k,\bar{\varepsilon})N^k(k-1)!}{W(\bar{\varepsilon})(-2\pi i)^k} = -2L_N(1-k,\varepsilon) = -2kB^{-1}_{k,\varepsilon};$$

see [**Hec1**, §1,2], [**Shi5**, (3.4)]. For $N = 1$ (so ε is trivial) and k even > 2, note that $E_{k,N,\varepsilon}$ is the normalized Eisenstein series E_k introduced in Example 2.2.1. This can be seen either from their definitions or by comparing their q-expansions (2.2.2) and (2.2.5).

If ε mod N is not primitive, then $E_{k,N,\varepsilon}(z)$ can be written as a linear combination of the forms $E_{k,C,\varepsilon_0}(dz)$ over divisors d of N/C, where C is the conductor of ε_0, the primitive character associated to ε. (See [**Hec1**], also [**Shi5**, (3.3)].)

REMARK 2.2.3. We digress briefly to discuss the functional equation for the series $E_{k,N,\varepsilon}(z,s)$ of (2.2.4).

In the case of $N = 1$ (k even, $\varepsilon = 1$), the Eisenstein series $E_k(z,s) = E_{k,1,1}(z,s)$ satisfies the functional equation

$$y^s E_k(z,s) = \Phi_k(s) y^{1-k-s} E_k(z, 1-k-s)$$

sending s to $1-k-s$, where

$$\Phi_k(s) = (-1)^{k/2} 2^{2-k-2s} \pi \cdot \frac{\Gamma(2s+k-2)\zeta(2s+k-1)}{\Gamma(s)\Gamma(s+k)\zeta(2s+k)};$$

see e.g. [**Kubo**]. It follows from the functional equation that $\Phi_k(s)\Phi_k(1-k-s) = 1$, which can also be checked directly.

For general $E_{k,N,\varepsilon}$, the functional equation is more complicated. We give here only a vague indication of its general shape and refer the reader to [**Kubo**], [**Huxl**] and [**Shi7**] for more details. One can consider instead a vector-valued function $\mathcal{E}(s)$ whose components include the series $E_{k,N,\chi}$ for characters χ mod N (satisfying $\chi(-1) = (-1)^k$). The components also include certain "companion series" for which the stabilizers in the defining sums (cf. (2.2.4)) are those of cusps inequivalent to ∞. The functional equation then relates the values of $\mathcal{E}(s)$ and $\mathcal{E}(1-s)$ (with suitable normalization in s). See also [**Hida3**, §9.3] for an adelic version.

EXAMPLE 2.2.4. The following are special cases of Example 2.2.2 and deal explicitly with weights one and two for which the defining series (2.2.4) do not converge absolutely at $s = 0$. The weight one case also provides an Eisenstein series which is important in studying congruences between modular forms (see Remark 2.2.5 below).

Let ε be an *odd* character mod N, i.e., $\varepsilon(-1) = -1$; we shall assume that it is primitive. Then

$$G_{1,\varepsilon}(z) = B_{1,\varepsilon} - 2\sum_{n=1}^{\infty}\Big(\sum_{d|n}\varepsilon(d)\Big)q^n$$

defines a modular form of type $(1, N, \varepsilon)$. Its constant term

$$B_{1,\varepsilon} = \sum_{a=1}^{N-1}\varepsilon(a)a/N = -L(0,\varepsilon)$$

is non-zero, and the normalized Eisenstein series $E_{1,\varepsilon} = G_{1,\varepsilon}/B_{1,\varepsilon}$ is precisely $E_{1,N,\varepsilon}$ of (2.2.5) in view of the fact that the values $L(0,\varepsilon)$ and $L(1,\bar{\varepsilon})$ are related via the functional equation (2.2.3) with $s = 1$ and $\bar{\varepsilon}$ in place of ε.

Similarly, in the case of weight $k = 2$, take ε to be a primitive non-trivial *even* character. Consider the series defined by

$$G_{2,\varepsilon}(z) = \frac{1}{2}B_{2,\varepsilon} - 2\sum_{n=1}^{\infty}\Big(\sum_{d|n}\varepsilon(d)d\Big)q^n.$$

Its constant term $B_{2,\varepsilon}/2 = -L(-1,\varepsilon)$ is non-zero, and again from the functional equation (2.2.3) relating $L(-1,\varepsilon)$ and $L(2,\bar{\varepsilon})$ (with $s = 2$), we see that the normalized function $E_{2,\varepsilon} = 2G_{2,\varepsilon}/B_{2,\varepsilon}$ is precisely $E_{2,N,\varepsilon}$ of (2.2.5). Thus $G_{2,\varepsilon}$ belongs to $\mathcal{M}_2(N,\varepsilon)$.

REMARK 2.2.5. Now, take N to be an odd prime ℓ and fix a prime divisor λ of $\mathbf{Q}(\mu_{\ell-1})$ lying above (ℓ) where $\mu_{\ell-1}$ denotes the set of $(\ell-1)$-th roots of unity. Let ε be the Dirichlet character mod ℓ such that $\varepsilon(a)a \equiv 1 \pmod{\lambda}$ for $a \in (\mathbf{Z}/\ell\mathbf{Z})^{\times}$. Here, $\varepsilon(a)$ belongs to $\mu_{\ell-1}$ and the congruence is in the ring of integers $\mathcal{O}$ of $\mathbf{Q}(\mu_{\ell-1})$. We have that $E_{1,\varepsilon}$ satisfies the congruence [**Koike**, §1]

$$E_{1,\varepsilon} \equiv 1 \pmod{\lambda};$$

indeed, all the coefficients of $E_{1,\varepsilon}$ (except of course the constant term) are in $b\mathcal{O}$ where $b = \ell/(\sum_1^{\ell-1}\varepsilon(a)a)$, and the denominator $\sum\varepsilon(a)a$ does not belong to the prime λ so that $b \equiv 0 \pmod{\lambda}$.

We remark here that the Eisenstein series $E_{\ell-1}$ of weight $\ell-1$ and level 1 (see Example 2.2.1) satisfies the similar congruence

$$E_{\ell-1} \equiv 1 \pmod{\ell} \quad \text{if } \ell \geq 5$$

which is essentially the von Staudt congruence; see e.g. [**Lang2**, §X.2].

Both $E_{1,\varepsilon}$ and $E_{\ell-1}$ play important roles in the theory of congruences between modular forms, for they provide congruences between modular forms of different weights. For example, if we take $\ell = 3$ (and so ε is the non-trivial character mod 3) then $E_{1,\varepsilon}$ has integer Fourier coefficients and satisfies $E_{1,\varepsilon} \equiv 1 \pmod{3}$. If f is a modular form of type $(1, M, \varepsilon')$ for some $M > 0$ and character ε', then $fE_{1,\varepsilon}$ is of type $(2, 3M, \varepsilon\varepsilon')$. Moreover if f has coefficients in the ring of integers $\mathcal{O}$ of

some number field, then so does $fE_{1,\varepsilon}$ and the Fourier expansions of f and $fE_{1,\varepsilon}$ are congruent mod $3\mathcal{O}$.

EXAMPLE 2.2.6. When $k = 2$ and ε is the trivial character mod N, there is a well known trick of Hecke to construct such modular forms (e.g. [**Hec1**, §2], [**Kna2**, §IX.3]). Note that the right-hand side of (2.2.2) makes sense even for $k = 2$. Let us denote it by E_2, i.e.,

$$E_2(z) = 1 - 24\sum_{n=1}^{\infty}\sigma_1(n)q^n$$

($B_2 = 1/6$); it is obtained by choosing the order of summation in (2.2.1) to be $\sum_m(\sum'_n \dots)$. Then E_2 is holomorphic on $\mathfrak{H}$ and at ∞, but fails to have the (weight 2) modularity property with respect to $\mathrm{SL}_2(\mathbf{Z})$ (in fact $\mathcal{M}_2(\mathrm{SL}_2(\mathbf{Z})) = 0$). Let

$$F_2(z) = \lim_{s\to 0^+} E_2(z,s),$$

where $E_2(z,s)$ is the Eisenstein series in Example 2.2.2 with $N = 1$ (and so ε is the trivial character). This time $F_2(z)$ is not holomorphic, for

$$F_2(z) = E_2(z) + c(\pi y)^{-1}$$

with some $c \neq 0$, but this nearly holomorphic function has the modularity property of weight 2 under $\mathrm{SL}_2(\mathbf{Z})$. Therefore, for any integer $N > 0$, the function

$$F_2(z) - NF_2(Nz) = E_2(z) - NE_2(Nz)$$

belongs to $\mathcal{M}_2(\Gamma_0(N))$. More generally, given numbers $c_d \in \mathbf{C}$ for $d|N$ such that $\sum_{d|N} c_d/d = 0$, the function $\sum_{d|N} c_d F_2(dz) = \sum_{d|N} c_d E_2(dz)$ is a modular form of weight 2 on $\Gamma_0(N)$.

In particular, if $N = p$ is a prime then

$$E_2(z) - pE_2(pz) = (1-p) - 24\sum_{n=1}^{\infty}\Big(\sum_{d|n}\varepsilon(d)d\Big)q^n$$

is a weight 2 modular form of level p with trivial character.

EXAMPLE 2.2.7. Let

$$\Delta = \frac{1}{1728}(E_4^3 - E_6^2).$$

Then Δ is a modular form on $\Gamma_1(1) = \mathrm{SL}_2(\mathbf{Z})$ of weight 12. It vanishes at ∞ since the constant term in its q-expansion is 0, as can be seen from the q-expansions of E_4 and E_6. As $\Delta|[\alpha]_{12} = \Delta$ for all $\alpha \in \mathrm{SL}_2(\mathbf{Z})$, we have that Δ vanishes at all the cusps. Hence, $\Delta \in \mathcal{S}_{12}(\Gamma_1(1))$. Its q-expansion is given by

$$\Delta(z) = q\prod_{n=1}^{\infty}(1-q^n)^{24}$$

(e.g. [**Ser1**, §VII.4]), and the coefficients define the Ramanujan function $\tau(n)$. There are no cusp forms on $\mathrm{SL}_2(\mathbf{Z})$ with weight smaller than 12; see e.g. [**Ser1**, §VII.3], [**Shi1**, §2.6] (and also §12.1).

EXAMPLE 2.2.8. On smaller congruence subgroups of $\mathrm{SL}_2(\mathbf{Z})$ there may be cusp forms of low weight. For example, $\mathcal{S}_2(\Gamma_0(11)) = \mathbf{C}f$, where

$$f(z) = \big(\Delta(z)\Delta(11z)\big)^{1/12} = q\prod_{n=1}^{\infty}[(1-q^n)(1-q^{11n})]^2.$$

In the terminology of §6.3, f is a newform of level or conductor 11 with trivial (or no) character. For more of such examples, let N be one of the integers $\{2, 3, 5, 11\}$ and let $k = 24/(N+1)$. Then $(\Delta(z)\Delta(Nz))^{1/(N+1)}$ is a cusp form on $\Gamma_0(N)$ and spans $\mathcal{S}_k(\Gamma_0(N))$; see [**Shi1**, Example 2.28], and also [**Birch**].

3. Hecke operators

Hecke operators [**Hec3**] arise in many contexts. They give rise to modular correspondences, they act on modular forms and on the integral homology of modular curves; roughly speaking, they act on objects arising from GL_2 by certain natural representation-theoretic and algebro-geometric constructions. Though they can be realized in various ways, it is the consistency with which they act that makes them so useful to study. We begin with a description, following [**Shi1**], of the abstract Hecke ring using double cosets. Then we explain how these double cosets give rise to modular correspondences, a subject to which we return in Part II. Then, as an important and concrete instance of how Hecke operators act in a particular setting, we shall discuss the representation of the Hecke ring on the space of modular forms. In particular, we consider the eigenforms, eigenvalues and eigenspaces for the Hecke operators, a subject to which we return in §6 and Part III.

Throughout this section, we shall fix a positive integer N.

3.1. Double coset description.

PRIMARY REFERENCES:
[**Shi1**, §3.1–3.3] and [**Miy2**, §2.7, 4.5].

With N as above, let

$$\Delta_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbf{Z}) \middle| \det > 0,\ c \equiv 0 \pmod N,\ (a, N) = 1 \right\},$$

$$\Delta_1(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbf{Z}) \middle| \det > 0,\ c \equiv a - 1 \equiv 0 \pmod N \right\}.$$

Put $\Gamma = \Gamma_1(N)$ and $\Delta = \Delta_1(N)$; though the notation (Γ, Δ) will be reserved for this pair in this section, our discussion is valid for $(\Gamma_0(N), \Delta_0(N))$ verbatim.

Let $R(\Gamma, \Delta)$ denote the $\mathbf{Z}$-module generated by the double cosets $\Gamma\alpha\Gamma$, $\alpha \in \Delta$. Note that $\Delta = \bigcup_{n \in \mathbf{N}} \Delta^n$, where $\Delta^n = \{\alpha \in \Delta | \det \alpha = n\}$. This can be made into a ring by defining multiplication between two double cosets $u = \Gamma\alpha\Gamma$ and $v = \Gamma\beta\Gamma$ as follows. Consider their coset decompositions $\Gamma\alpha\Gamma = \coprod_i \Gamma\alpha_i$ and $\Gamma\beta\Gamma = \coprod_j \Gamma\beta_j$. Then $\Gamma\alpha\Gamma\beta\Gamma = \bigcup_{i,j} \Gamma\alpha_i\beta_j$ (not necessarily disjoint), and so $\Gamma\alpha\Gamma\beta\Gamma$ is a finite union of double cosets of the form $\Gamma\gamma\Gamma$. Define

$$u \cdot v = \sum_w m(u, v; w) w$$

where the sum is extended over all double cosets $w = \Gamma\gamma\Gamma \subset \Gamma\alpha\Gamma\beta\Gamma$, and

$$m(u, v; w) = \#\{(i, j) | \Gamma\alpha_i\beta_j = \Gamma\gamma\} \tag{3.1.1}$$

for $w = \Gamma\gamma\Gamma$. One can check that these definitions depend only on u, v and w, and not on the choices of representatives $\{\alpha_i\}$, $\{\beta_j\}$, γ.

Equipped with the above multiplication law extended linearly, $R(\Gamma, \Delta)$ becomes an associative, and in fact commutative, ring with $\Gamma = \Gamma \cdot 1 \cdot \Gamma$ as the unit element. It is called the Hecke ring with respect to (Γ, Δ).

3.2. Modular correspondences.

PRIMARY REFERENCES:
[**Shi1**, §3.4, §7.2] and [**Miy2**, §2.8].

Now we explain how the double cosets we have defined give rise to correspondences on modular curves. Though modular curves and correspondences will be one of the central topics of Part II, we give a brief introduction here.

For a congruence subgroup Γ we call the quotient space $\Gamma\backslash\mathfrak{H}$ the modular curve associated to Γ. We are especially interested in the modular curves

$$\begin{aligned} Y_0(N) &= \Gamma_0(N)\backslash\mathfrak{H} \\ Y_1(N) &= \Gamma_1(N)\backslash\mathfrak{H} \end{aligned}$$

associated to $\Gamma_0(N)$ and $\Gamma_1(N)$ respectively.

For a pair of modular curves X and Y, it will suffice for the moment to view a "correspondence on $X \times Y$" as a homomorphism $\mathrm{Div}\,(X) \to \mathrm{Div}\,(Y)$ where $\mathrm{Div}\,X$ denotes the free abelian group generated by the elements of X. In particular, a function $f : X \to Y$ extends to a correspondence which we denote by the same symbol. Note that the the correspondences on $X \times X$ form an associative ring, where multiplication is given by composition of correspondences.

Let Γ be the congruence subgroup $\Gamma_0(N)$ or $\Gamma_1(N)$, let Y be the curve $Y_0(N)$ or $Y_1(N)$, and let Δ be $\Delta_0(N)$ or $\Delta_1(N)$, respectively. For any α such that $\alpha^{-1}\Gamma\alpha$ and Γ are commensurable, e.g. for $\alpha \in \Delta$, put

$$\Gamma_\alpha = \Gamma \cap \alpha^{-1}\Gamma\alpha \qquad \text{and} \qquad Y_\alpha = \Gamma_\alpha\backslash\mathfrak{H}\,.$$

Let $\varphi : \mathfrak{H} \to Y$ and $\varphi_\alpha : \mathfrak{H} \to Y_\alpha$ be the canonical projections, and consider the (possibly branched) coverings $\pi, \pi^\alpha : Y_\alpha \to Y$ defined by $\pi \circ \varphi_\alpha = \varphi$ and $\pi^\alpha \circ \varphi_\alpha = \varphi \circ \alpha$. These are induced from the obvious maps id and α on $\mathfrak{H}$, i.e., π is the natural projection, and π^α is the composition of the natural projection $\Gamma_\alpha\backslash\mathfrak{H} \to \alpha^{-1}\Gamma\alpha\backslash\mathfrak{H}$ followed by the isomorphism $\alpha^{-1}\Gamma\alpha\backslash\mathfrak{H} \to \Gamma\backslash\mathfrak{H}$ obtained from $z \mapsto \alpha(z)$. Using these coverings, we get a correspondence $\tau_\alpha = \pi^\alpha \circ {}^t\pi$ from Y to itself where ${}^t\pi$, the transpose of π, is defined as follows: If $\Gamma = \coprod_{i=1}^{e} \Gamma_\alpha\epsilon_i$ is a (finite) coset decomposition of $\Gamma_\alpha\backslash\Gamma$ then ${}^t\pi$ sends a point $\varphi(z) \in Y$, $z \in \mathfrak{H}$, to the formal sum $\sum_i \varphi_\alpha(\epsilon_i z)$ of points in its preimage $\pi^{-1}(\varphi(z))$ (counted with multiplicity). Thus, $\tau_\alpha(\varphi(z))$ is the divisor $\sum_i \varphi(\alpha\epsilon_i(z))$. Since $\varphi(\beta(z))$ depends only on the coset $\Gamma\beta$, we have: $\tau_\alpha(\varphi(z)) = \sum_i \varphi(\alpha_i(z))$ if $\Gamma\alpha\Gamma = \coprod_i \Gamma\alpha_i$. (The coset decomposition $\Gamma = \coprod_i \Gamma_\alpha\epsilon_i$ gives a disjoint union $\Gamma\alpha\Gamma = \coprod_i \Gamma\alpha\epsilon_i$, so that the divisor is recovered with $\alpha_i = \alpha\epsilon_i$.) One can check that τ_α depends only on the double coset $\Gamma\alpha\Gamma$, and that $\Gamma\alpha\Gamma \mapsto \tau_\alpha$ defines a homomorphism from the Hecke ring $R(\Gamma, \Delta)$ to the ring of correspondences on $(\Gamma\backslash\mathfrak{H}) \times (\Gamma\backslash\mathfrak{H})$.

3.3. Hecke rings.

PRIMARY REFERENCES:
[**Shi1**, §3.1–3.3] and [**Miy2**, §4.5].

For each positive integer n, denote by $T(n)$ the formal sum of all double cosets $\Gamma\alpha\Gamma$ with $\alpha \in \Delta^n$ in $R(\Gamma, \Delta)$. For example, $T(p) = \Gamma\left(\begin{smallmatrix} 1 & 0 \\ 0 & p \end{smallmatrix}\right)\Gamma$ for every prime p. Further, for two positive integers a, d such that $a|d$ and $(d, N) = 1$, let $T(a, d)$ denote the double coset $\Gamma\sigma_a\left(\begin{smallmatrix} a & 0 \\ 0 & d \end{smallmatrix}\right)\Gamma$, where σ_a is as in (2.1.2). Note that $T(1, p) = T(p)$. Let us write $m|N^\infty$ if every prime factor of m divides N. The structure of $R(\Gamma, \Delta)$ in terms of the Hecke operators $T(a, d)$ and $T(m)$ is given by the following [**Shi1**, Theorem 3.34]

PROPOSITION 3.3.1. 1. *$R(\Gamma,\Delta)$ is a polynomial ring over $\mathbf{Z}$ in the variables $T(p,p)$ for all primes $p\nmid N$ and $T(p)$ for all primes p.*
2. *Every element $\Gamma\alpha\Gamma$ with $\alpha\in\Delta$ is uniquely expressed as a product*

$$T(m)T(a,d)=T(a,d)T(m),$$

where $m|N^\infty$, $a|d$ and $(d,N)=1$.
3. *If $(m,n)=1$ or $m|N^\infty$ or $n|N^\infty$, then $T(mn)=T(m)T(n)$.*
4. *$R(\Gamma,\Delta)\otimes_{\mathbf{Z}}\mathbf{Q}$ is generated over $\mathbf{Q}$ by $T(n)$ for all n.*

REMARK 3.3.2. The last assertion follows from the first assertion together with the equation

$$pT(p,p)=T(p)^2-T(p^2),$$

which is valid for every prime p not dividing N.

Let $\Gamma(1)$, $\Delta(1)$ be Γ, Δ with $N=1$, i.e., $\Gamma(1)=\mathrm{SL}_2(\mathbf{Z})$ and $\Delta(1)$ the set of 2×2 integral matrices with positive determinant. Let $\tilde{T}(n)$, $\tilde{T}(a,d)$ with $a|d$ temporarily denote the Hecke operators of level 1, i.e., with respect to $(\Gamma(1),\Delta(1))$. Then $R(\Gamma(1),\Delta(1))$ is $\mathbf{Z}[\tilde{T}(p),\tilde{T}(p,p);\forall p]$ and $R(\Gamma,\Delta)$ is its homomorphic image via the map

$$\begin{array}{lcll}\tilde{T}(p) & \mapsto & T(p) & \forall \text{ prime } p\ ,\\ \tilde{T}(p,p) & \mapsto & T(p,p) & \forall \text{ prime } p\nmid N\ ,\\ \tilde{T}(p,p) & \mapsto & 0 & \forall \text{ prime } p|N\ .\end{array}$$

Thus, any algebraic relation amongst the Hecke operators of $R(\Gamma(1),\Delta(1))$ can be translated to the corresponding relation for the Hecke operators in $R(\Gamma,\Delta)$, where the only change is that we put 0 in place of $\tilde{T}(p,p)$ for $p|N$. An example of such a relation is

$$T(m)T(n)=\sum_d dT(d,d)T(mn/d^2)$$

where the sum is over positive divisors d of (m,n) which are relatively prime to N.

The element $T(p,p)$ for $p\nmid N$, p prime, is often denoted $S(p)$ in literature; if N is the level, we put $S(p)=0$ for $p|N$.

3.4. Action on modular forms.

PRIMARY REFERENCES:
[**Shi1**, §3.4, 3.5], [**Ser1**, §VII.5], [**Lang2**, §VII.2, VII.3] and [**Miy2**, §2.8, 4.5].

Thus far, we have discussed the Hecke operators as elements in an abstract ring. We now turn to considering how they are realized on the space of modular forms. For this we first describe the action of a double coset on modular forms on Γ.

For $f\in\mathcal{M}_k(\Gamma)$ and $\alpha\in\Delta$, the action of the double coset $\Gamma\alpha\Gamma$ is

$$f|[\Gamma\alpha\Gamma]_k=\sum_i f|[\alpha_i]_k,$$

where $\{\alpha_i\}$ is the set of representatives for $\Gamma\backslash\Gamma\alpha\Gamma$. This gives a well-defined action of $[\Gamma\alpha\Gamma]_k$ on $\mathcal{M}_k(\Gamma)$ which preserves the subspace $\mathcal{S}_k(\Gamma)$. Extending by linearity gives an action of $R(\Gamma,\Delta)$ on $\mathcal{M}_k(\Gamma)$ and $\mathcal{S}_k(\Gamma)$. To make this action more explicit, we use the following ([**Shi1**, Proposition 3.36]).

LEMMA 3.4.1. *For every $(a, N) = 1$, fix an element σ_a of $\mathrm{SL}_2(\mathbf{Z})$ as defined by (2.1.2). Then for every $n \in \mathbf{N}$, we have*

$$\Delta^n = \coprod_a \coprod_b \Gamma \sigma_a \begin{pmatrix} a & b \\ 0 & d \end{pmatrix},$$

where the disjoint union is over $a > 0$ with $ad = n$, $(a, N) = 1$ and over $b \pmod d$.

Recall that $\Delta^n = \{\alpha \in \Delta \mid \det \alpha = n\}$. From this coset decomposition, we have the action of the n-th Hecke operator on $\mathcal{M}_k(\Gamma)$, denoted $T(n)_k$, given by

$$f|T(n)_k = \sum_{a,\, b} f|[\sigma_a \begin{pmatrix} a & b \\ 0 & d \end{pmatrix}]_k, \tag{3.4.1}$$

the sum being over a, b as in Lemma 3.4.1.

Put $\Gamma_0 = \Gamma_0(N)$, $\Delta_0 = \Delta_0(N)$. The above discussion and the lemma are also valid when $(\Gamma, \Delta) = (\Gamma_0, \Delta_0)$.

More generally, the Hecke operators act on the space of modular forms of type (k, N, ε). Observe that the operators $T(n)_k$ and $T(n, n)_k = n^{k-2}\langle n \rangle_k$ on $\mathcal{M}_k(\Gamma_1(N))$ preserve the subspaces $\mathcal{M}_k(N, \varepsilon)$ as they commute with the operations of $d \in (\mathbf{Z}/N\mathbf{Z})^\times$ via $\langle d \rangle_k$. (Recall that $\langle d \rangle_k$ was defined before (2.1.2).) The map $\Gamma\beta\Gamma \mapsto \Gamma_0\beta\Gamma_0$ defines a surjective homomorphism $R(\Gamma, \Delta) \to R(\Gamma_0, \Delta_0)$ and the restriction of $[\Gamma\beta\Gamma]_k$ to $\mathcal{M}_k(N, \varepsilon)$ depends only on $\Gamma_0\beta\Gamma_0$. Therefore $R(\Gamma_0, \Delta_0)$ acts on $\mathcal{M}_k(N, \varepsilon)$ and the action is given by

$$f|[\Gamma_0\alpha\Gamma_0]_{k,\varepsilon} = \sum_\nu \varepsilon(a(\alpha_\nu)) f|[\alpha_\nu]_k, \qquad f \in \mathcal{M}_k(N, \varepsilon)$$

where $\alpha \in \Delta_0$, $\Gamma_0\alpha\Gamma_0 = \coprod_\nu \Gamma_0\alpha_\nu$ and $a(\alpha)$ denotes the a-entry of the matrix α. Using this and Lemma 3.4.1, the action of the Hecke operators can be made explicit. The lemma gives $\Gamma\alpha\Gamma = \coprod_\nu \Gamma\alpha_\nu$ with α_ν of the form $\sigma_a \begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$, which yields $\Gamma_0\alpha\Gamma_0 = \coprod_\nu \Gamma_0\alpha_\nu$ with the same α_ν's. Also, for $f \in \mathcal{M}_k(N, \varepsilon)$ we have $f|[\sigma_a]_k = \varepsilon(a)f$ for every integer a prime to N. Thus, if we denote by $T(n)_{k,\varepsilon}$ and $T(a, d)_{k,\varepsilon}$ the corresponding restricted actions of $T(n)_k$ and $T(a, d)_k$ on $\mathcal{M}_k(N, \varepsilon)$ as above, then

$$f|T(n)_{k,\varepsilon} = n^{k-1} \sum_a \sum_{b=0}^{d-1} \varepsilon(a) d^{-k} f\left(\frac{az + b}{d}\right) \qquad (a > 0,\ ad = n). \tag{3.4.2}$$

(Recall that $\varepsilon(a) = 0$ for $(a, N) \neq 1$.) Also,

$$f|T(d, d)_{k,\varepsilon} = d^{k-2}\varepsilon(d)f$$

for $f \in \mathcal{M}_k(N, \varepsilon)$ and $(d, N) = 1$. This yields

$$T(m)_{k,\varepsilon}T(n)_{k,\varepsilon} = \sum_{d|(m,n)} d^{k-1}\varepsilon(d)T(mn/d^2)_{k,\varepsilon}. \tag{3.4.3}$$

For n with $(n, N) = 1$ we may also view $T(n)_{k,\varepsilon}$ as an operator on $\mathcal{M}_k(\Gamma)$ by setting it equal to 0 on the eigenspaces $\mathcal{M}_k(N, \varepsilon')$ for $\varepsilon' \neq \varepsilon$, i.e., $T(n)_{k,\varepsilon} = T(n)_k \circ \mathrm{pr}_\varepsilon$, where

$$\mathrm{pr}_\varepsilon = \frac{1}{\#((\mathbf{Z}/N\mathbf{Z})^\times)} \sum_a \bar{\varepsilon}(a)\langle a \rangle_k$$

is the projection of $\mathcal{M}_k(\Gamma)$ onto $\mathcal{M}_k(N, \varepsilon)$.

REMARK 3.4.2. Let k and N be fixed. We have encountered, for each $n \in \mathbf{N}$, several actions of the n-th Hecke operator $T(n)$. One is the action $T(n)_k$ on $\mathcal{M}_k(\Gamma_1(N))$ where $T(n)$ is viewed as an element of the ring $R(\Gamma_1(N), \Delta_1(N))$; another is the action of $T(n)_{k,\varepsilon}$ on the space $\mathcal{M}_k(N, \varepsilon)$; yet another is the action $T_0(n)_k$ on $\mathcal{M}_k(\Gamma_0(N))$ where $T(n)$ is viewed as an element of $R(\Gamma_0(N), \Delta_0(N))$. But whenever a modular form f lies in the intersection of any two of these spaces these Hecke actions on f are exactly the same. In addition to all previous notation, we shall therefore use a looser notation T_n to denote any of these actions and write $T_n f$ or $f|T_n$ whenever the action on f is defined and k and N are clear from the context. Similarly, we write $\langle n \rangle f$ or $f|\langle n \rangle$ in such a situation. Moreover, we shall occasionally restrict our attention to the operators acting on a space of forms of type (k, N, ε) with a given Nebentypus ε, because the forms on $\Gamma_0(N)$ yield such a space (with ε the trivial character mod N) while the space $\mathcal{S}_k(\Gamma_1(N))$ is a direct sum of the space $\mathcal{S}_k(N, \varepsilon)$, where ε ranges over all characters mod N satisfying $\varepsilon(-1) = (-1)^k$.

Returning to the Hecke actions on modular forms, equation (3.4.3) can be summarized by the formal identity

$$\sum_{n=1}^{\infty} T_n n^{-s} = \prod_p (1 - T_p p^{-s} + \varepsilon(p) p^{k-1-2s})^{-1}$$

on the space of modular forms of type (k, N, ε). Also, if formula (3.4.2) for the action of $T(n)_{k,\varepsilon}$ is unravelled in terms of the q-expansion of f at ∞ we obtain (e.g. [**Shi1**, (3.5.11)]):

PROPOSITION 3.4.3. *Let $\sum_0^\infty a_n q^n$ be the q-expansion of $f \in \mathcal{M}_k(N, \varepsilon)$, and let $\sum_0^\infty b_n q^n$ be the q-expansion of $T_m f$. Then the coefficients b_n are given by*

$$b_n = \sum_{d|(m,n)} \varepsilon(d) d^{k-1} a_{mn/d^2}.$$

This formula provides a characterization of the Hecke operators which is quite practical from a computational point of view.

Consider the operators U_m, V_m defined on $\mathbf{C}[[q]]$ by

$$U_m(\sum_n a_n q^n) = \sum_n a_{mn} q^n, \qquad V_m(\sum_n a_n q^n) = \sum_n a_n q^{mn}.$$

They satisfy $U_{m_1 m_2} = U_{m_1} \circ U_{m_2}$, $V_{m_1 m_2} = V_{m_1} \circ V_{m_2}$, and $U_{p_1} \circ V_{p_2} = V_{p_2} \circ U_{p_1}$ for primes $p_1 \neq p_2$. Also, $U_m \circ V_m$ is the identity, while $V_m \circ U_m$ is the projection on the part of the power series with powers of q divisible by m. In terms of these operators, we have

$$T_n = \sum_{d|n} \varepsilon(d) d^{k-1} V_d \circ U_{n/d}.$$

Equivalently, this is captured in the formal identity

$$\sum_{n=1}^{\infty} T_n n^{-s} = \Big(\sum_{n=1}^{\infty} \varepsilon(n) n^{k-1} V_n n^{-s}\Big)\Big(\sum_{n=1}^{\infty} U_n n^{-s}\Big).$$

(See Chapter VII, Theorem 3.2 of [**Lang2**].) Note that $T_p = U_p$ for primes p dividing N, since $\varepsilon(p) = 0$ for such p by convention.

3.5. Hecke eigenforms, eigenvalues and eigenspaces.

PRIMARY REFERENCES:
[**Shi1**, §3.5], [**Ser1**, §VII.5], [**Lang2**, §VII.3] and [**Kna2**, §IX.6].

Let N be a positive integer, and denote by $\mathbf{T}_N$ the polynomial ring over $\mathbf{Z}$ generated by indeterminates T_p for all primes p and indeterminates S_p for all primes p not dividing N. It is the full Hecke algebra of level N, and is isomorphic to the Hecke ring $R(\Gamma, \Delta)$ by the first assertion of Proposition 3.3.1. Also, denote by $\mathbf{T}^{(N)}$ the subring generated by T_p, S_p for all primes p not dividing N. Then the spaces of modular or cusp forms we have discussed previously, such as $\mathcal{M}_k(\Gamma_1(N))$, $\mathcal{S}_k(N, \varepsilon)$, etc., are modules over $\mathbf{T}_N$ and $\mathbf{T}^{(N)}$ via the usual Hecke action of these indeterminates. To study the Hecke action on a space of modular forms, we need only consider the images of $\mathbf{T}_N$ and $\mathbf{T}^{(N)}$ in $\operatorname{End} \mathcal{M}_k(\Gamma_1(N))$, the ring of endomorphisms of $\mathcal{M}_k(\Gamma_1(N))$. We remark that in the literature, slightly different sets of Hecke operators are often chosen, but they yield the same subring of $\operatorname{End} \mathcal{M}_k(\Gamma_1(N))$.

PROPOSITION 3.5.1. 1. *Let $\tilde{\mathbb{T}}$ be the subring of* $\operatorname{End} \mathcal{M}_k(\Gamma_1(N))$ *generated by $\{T_n\}$ for all $n \in \mathbf{N}$, and $\tilde{\mathbb{T}}'$ the subring generated by $\{T_p, \langle q \rangle_k\}$ for all primes p and all primes $q \not| N$. Then, $\tilde{\mathbb{T}} = \tilde{\mathbb{T}}'$.*

2. *For $k \geq 2$, this ring is precisely the image of $\mathbf{T}_N$ in* $\operatorname{End} \mathcal{M}_k(\Gamma_1(N))$. *For $k = 1$, $\tilde{\mathbb{T}} (= \tilde{\mathbb{T}}')$ is contained in the image, and we have equality after tensoring with $\mathbf{Q}$. For $k = 0$, all these rings are just $\mathbf{Z}$ with all primes not dividing N inverted.*

3. *Similar statements hold for $\mathbf{T}^{(N)}$ (with the corresponding subrings generated by the elements "away from N").*

Indeed, the formula

$$p^{k-1} \langle p \rangle_k = T_p^2 - T_{p^2}$$

shows that for $k \geq 1$, $\{T_n\}$ and $\{T_p, \langle q \rangle_k\}$ generate the same subrings of endomorphisms. One inclusion is obvious and for the other, apply the formula with two primes q and r congruent mod N (noticing that q^{k-1}, r^{k-1} are relatively prime and $\langle q \rangle_k = \langle r \rangle_k$). A similar argument using

$$q^{k-2} \langle q \rangle_k = S_q$$

shows that this ring is precisely the image of $\mathbf{T}_N$ if $k \geq 2$. For $k = 1$ this formula reads $\langle q \rangle_1 = qS_q$, yielding only one inclusion of the subrings, but an equality after tensoring with $\mathbf{Q}$. The same argument is valid for $\mathbf{T}^{(N)}$. For $k = 0$, we have $T_p = 1$ if p is a prime dividing N; for primes p not dividing N, we have $T_p = 1 + p^{-1}$, $\langle p \rangle_0 = 1$ and $S_p = p^{-2}$. The formula $p^{-1} = T_p - 1$ (when $p \not| N$) shows that $\tilde{\mathbb{T}} = \tilde{\mathbb{T}}'$, and this ring is as described in the assertion. Similarly, as $S_p = (p^{-1})^2 = (T_p - 1)^2$, it coincides with the image of $\mathbf{T}_N$ in the endomorphism ring; it is also the image of $\mathbf{T}^{(N)}$ in this case.

For $\mathbf{T} = \mathbf{T}_N$ or $\mathbf{T}^{(N)}$, we call an element of $\mathcal{M}_k(\Gamma_1(N))$ a $\mathbf{T}$-eigenform if it is a common eigenvector under all $T \in \mathbf{T}$. A $\mathbf{T}^{(N)}$-eigenform is not necessarily a $\mathbf{T}_N$-eigenform. For instance, the Ramanujan $\Delta \in \mathcal{S}_{12}(\Gamma_0(N))$ of Example 2.2.7 is a $\mathbf{T}^{(N)}$-eigenform, but never a $\mathbf{T}_N$-eigenform for $N > 1$.

REMARK 3.5.2. All of the examples given in §2.2 are $\mathbf{T}^{(N)}$-eigenforms. They are even $\mathbf{T}_N$-eigenforms, provided in Example 2.2.2 that the character ε is primitive, and in Example 2.2.6 that N is prime.

Observe that a (non-zero) $\mathbf{T}^{(N)}$-eigenform in $\mathcal{M}_k(\Gamma_1(N))$ has to have a (unique) character, i.e., it is necessarily of type (k, N, ε) for some Nebentypus ε mod N. This is because the image of $(\mathbf{Z}/N\mathbf{Z})^\times$ in the endomorphism ring is contained in that of $\mathbf{T}^{(N)}$ by Proposition 3.5.1. Also, the subring generated by $\{\langle q\rangle_k\}$ for all primes q not dividing N is precisely the image of $(\mathbf{Z}/N\mathbf{Z})^\times$ (under $d \mapsto \langle d\rangle_k$). In fact Proposition 3.5.1 yields several equivalent definitions of a $\mathbf{T}$-eigenform. For instance, a modular form on $\Gamma_1(N)$ is a $\mathbf{T}^{(N)}$-eigenform if and only if it is of type (k, N, ε) for some ε and is a common eigenvector under T_p for all primes p not dividing N; it is a $\mathbf{T}_N$-eigenform if and only if it is a simultaneous eigenvector under all T_n.

For each non-zero $\mathbf{T}$-eigenform f, we may consider the $\mathbf{T}$-eigenspace consisting of $\mathbf{T}$-eigenforms g with the same eigenvalue as f under each operator in $\mathbf{T}$. Note that by commutativity of the operators involved, a $\mathbf{T}^{(N)}$-eigenform will be a $\mathbf{T}_N$-eigenform if the $\mathbf{T}^{(N)}$-eigenspace to which it belongs is one-dimensional. While this fails in general, we shall see in §6.3 that this holds for certain forms called newforms.

Let us next observe that every $\mathbf{T}_N$-eigenspace is (at most) one-dimensional: Suppose $f \in \mathcal{M}_k(N, \varepsilon)$ is a (non-zero) $\mathbf{T}_N$-eigenform and let $\sum_0^\infty a_n q^n$ be the q-expansion of f. Then its Fourier coefficients a_n can be read off in terms of the eigenvalues. If λ_n denotes the n-th eigenvalue, i.e., $f|T_n = \lambda_n f$, then it follows from Proposition 3.4.3 that

- $a_n = \lambda_n a_1$ for all $n \in \mathbf{N}$;
- $a_1 \neq 0$ if $k \neq 0$ (so, $a_1 = 0 \Rightarrow k = 0$ and $f = a_0$);
- if $a_0 \neq 0$ then $\lambda_n = \sum_{d|n} \varepsilon(d) d^{k-1}$.

Thus, if two forms of type (k, N, ε) are common eigenforms of T_n for all n with the same system $\{\lambda_n\}$ of eigenvalues then one is a scalar multiple of the other. Such a form is said to be normalised if $a_1 = 1$.

REMARK 3.5.3. Let f be such an eigenform of weight $k \geq 1$. Then $Tf = \theta_f(T)f$ defines a homomorphism $\theta_f : \mathbf{T}_N \to \mathbf{C}$. The image is in fact contained in a number field, and the eigenvalues λ_n lie in its ring of integers; see Corollary 12.4.5 below. This was proved by Shimura [**Shi1**, Theorem 3.48] for $k \geq 2$; for $k \geq 1$, see [**Shi6**, Propositions 1.3 and 2.2] and references therein, and also [**Ser3**, §2.5].

3.6. Petersson inner product.

PRIMARY REFERENCES:
[**Shi1**, §3.4, 3.5], [**Lang2**, §III.4] and [**Miy2**, §2.1, 4.5].

Let Γ be an arbitrary congruence subgroup of $\mathrm{SL}_2(\mathbf{Z})$, and denote by $\overline{\Gamma}$ its projectivization, i.e., its image in $\mathrm{PSL}_2(\mathbf{Z}) = \mathrm{SL}_2(\mathbf{Z})/\{\pm 1\}$. On the space $\mathcal{S}_k(\Gamma)$ of cusp forms, define the Petersson inner product of two elements f and g by

$$\langle f, g\rangle = \frac{1}{[\overline{\Gamma(1)} : \overline{\Gamma}]} \int_D f(z)\overline{g(z)}\, y^k \frac{dxdy}{y^2}. \tag{3.6.1}$$

where D is a fundamental domain for Γ (see Remark 7.1.1). The convergence of the integral can be deduced from the following growth proprty (e.g. [**Shi1**, Lemma 3.61]) for cusp forms.

LEMMA 3.6.1. *If f is a cusp form in $\mathcal{S}_k(\Gamma)$ then $f(z)y^{k/2}$ is bounded on $\mathfrak{H}$ (here, y is the imaginary part of z).*

Moreover the integral is independent of the choice of the fundamental domain D, and of the choice of the congruence subgroup Γ with respect to which f, g are

modular. For the latter independence one uses that if Γ' is another congruence subgroup, say contained in Γ, then a fundamental domain for Γ' can be chosen which consists of $[\overline{\Gamma} : \overline{\Gamma'}]$ translates of a fundamental domain for Γ. The Petersson inner product is positive definite on the space of cusp forms.

REMARK 3.6.2. The notation $\langle\ ,\ \rangle$ is also used whenever the integral (3.6.1) converges. For instance, if f, g are weight k modular forms of which at least one is a cusp form then fg is a cusp form of weight $2k$, so that $f(z)g(z)y^k$ is bounded on $\mathfrak{H}$ by Lemma 3.6.1. Hence the integral (3.6.1) defining $\langle f, g\rangle$ is meaningful and finite in this case as well.

With respect to the Petersson product, the operation $[\alpha]_k$ of $\alpha \in \mathrm{GL}_2^+(\mathbf{Q})$ is unitary and its adjoint is given by $[\alpha^\iota]_k$ where α^ι is the main involution of α, i.e., $\alpha^\iota\alpha = (\det\alpha)I$. For example, on $\mathcal{S}_k(\Gamma_1(N))$ the adjoint of $\langle a\rangle_k$ for $(a, N) = 1$ is $\langle\bar{a}\rangle_k$, where $\bar{a}$ is an integer such that $a\bar{a} \equiv 1 \pmod{N}$.

On $\mathcal{S}_k(\Gamma_0(N))$ the Hecke operators T_n for $(n, N) = 1$ are self-adjoint with respect to the Petersson product; see [**Shi1**]. In fact, on $\mathcal{S}_k(N, \varepsilon)$ we have for all n prime to N

$$\langle T_n f, g\rangle = \varepsilon(n)\langle f, T_n g\rangle\,,$$

i.e., the adjoint of T_n with respect to $\langle\ ,\ \rangle$ is $T_n^* = \bar{\varepsilon}(n)T_n$. On $\mathcal{S}_k(\Gamma_1(N))$, the adjoint T_n^* of T_n for $(n, N) = 1$ is $T_n \circ \langle\bar{n}\rangle$. Thus the operators of the form T_n and $\langle n\rangle$ for n relatively prime to N form a mutually commutative set of normal operators on $\mathcal{S}_k(\Gamma_1(N))$. (Those operators T_n with $(n, N) \neq 1$ on $\mathcal{S}_k(N, \varepsilon)$ need not be normal.) Applying the spectral decomposition theorem for normal operators (e.g. [**Hers**, Theorem 6.10.4]), we deduce that there is an orthogonal decomposition

$$\mathcal{S}_k(\Gamma_1(N)) = \bigoplus_{\varepsilon} \mathcal{S}_k(N, \varepsilon)$$

(where ε runs over all Dirichlet characters mod N such that $\varepsilon(-1) = (-1)^k$), and that each $\mathcal{S}_k(N, \varepsilon)$ decomposes orthogonally into a direct sum of $\mathbf{T}^{(N)}$-eigenspaces [**Miy2**, Theorem 4.5.4].

4. W-operators

PRIMARY REFERENCES:
[**Shi1**, §3.5], [**Lang2**, §VII.6], [**AtLi**], [**LiOe**, §5] and [**Kna2**, §IX.4, IX.7].

We now discuss the W-operators, which form another useful class of operators on modular forms. On the space of forms on $\Gamma_0(N)$, these are involutions and they commute with the Hecke operators T_p for p not dividing N.

Let $\Gamma = \Gamma_1(N)$, and $w_N = \begin{pmatrix} 0 & -1 \\ N & 0 \end{pmatrix}$. On $\mathcal{M}_k(\Gamma)$, the linear operator $[w_N]_k = [\Gamma w_N \Gamma]_k$ satisfies $[w_N]_k^2 = (-N)^{k-2}$ and preserves the subspace $\mathcal{S}_k(\Gamma)$ of cusp forms. We let W_N be the operator on $\mathcal{M}_k(\Gamma)$ defined by $W_N(f) = N^{1-k/2} f|[w_N]_k$. Thus $W_N^2 = (-1)^k$, and W_N maps modular forms of type (k, N, ε) to those of type $(k, N, \bar{\varepsilon})$ since

$$\langle a\rangle_k[w_N]_k = [w_N]_k\langle\bar{a}\rangle_k$$

for every $a \in (\mathbf{Z}/N\mathbf{Z})^\times$, where $a\bar{a} \equiv 1 \pmod{N}$. Also, from the fact that

$$(\Gamma\alpha^\iota\Gamma)(\Gamma w_N\Gamma) = (\Gamma w_N\Gamma)(\Gamma\alpha\Gamma)$$

for every $\alpha \in \Delta$ with $(\det \alpha, N) = 1$, it follows that

$$f|T(n)_{k,\varepsilon}[w_N]_k = \varepsilon(n) f|[w_N]_k T(n)_{k,\bar{\varepsilon}}, \qquad f \in \mathcal{M}_k(N, \varepsilon)$$

for every n such that $(n, N) = 1$. For such n we have

$$W_N \langle n \rangle f = \langle n \rangle^{-1} W_N f \quad \text{and} \quad W_N T_n f = \langle n \rangle T_n W_N f$$

for f in $\mathcal{M}_k(\Gamma)$.

We also find that W_N and W_N^{-1} are adjoint with respect to the Petersson inner product on $\mathcal{S}_k(\Gamma)$. Moreover T_n is adjoint to $W_N^{-1} T_n W_N$ for all integers n, and if $(n, N) = 1$, then $\langle n \rangle$ is adjoint to $W_N^{-1} \langle n \rangle W_N$. If a cusp form f is a simultaneous eigenform away from N with eigenvalues λ_n, $(n, N) = 1$, then so is $W_N f$ with the corresponding eigenvalues $\overline{\lambda}_n$, $(n, N) = 1$. However, suppose for some prime p dividing N that f is an eigenvector under T_p. It need not be the case that $W_N f$ is an eigenvector under T_p. (See Remark 3.59 of [**Shi1**].) Indeed, if the condition "away from N" in the above statement is replaced by "for all $n \in \mathbf{N}$" the new statement is no longer true in general. The obstruction is due to the existence of the so-called "old" or "non-primitive" forms which come from lower levels (see §6.3).

More generally, we can associate an operator W_Q to each positive divisor Q of N such that Q and N/Q are relatively prime. Consider any matrix

$$w_Q = \begin{pmatrix} Qa & b \\ N & Qd \end{pmatrix}$$

of determinant Q with a, b and d integers and $d \equiv 1 \bmod N/Q$; such a matrix normalizes $\Gamma = \Gamma_1(N)$. The map $[w_Q]_k$ on $\mathcal{M}_k(\Gamma)$ is independent of the choice of defining matrix w_Q and is consistent with the old definition in the case $N = Q$. Moreover, the automorphism $\gamma \mapsto w_Q \gamma w_Q^{-1}$ induces the involution of $\Gamma_0(N)/\Gamma_1(N) \simeq (\mathbf{Z}/N\mathbf{Z})^\times$ $(\simeq (\mathbf{Z}/Q\mathbf{Z})^\times \times (\mathbf{Z}/(N/Q)\mathbf{Z})^\times)$ which is given by $d \mapsto d^{-1} \bmod Q$ and the identity mod N/Q on the respective factors. From this we deduce that if ε_Q and $\varepsilon_{N/Q}$ are Dirichlet characters mod Q and N/Q respectively, then $[w_Q]_k$ maps modular forms of type $(k, N, \varepsilon_Q \varepsilon_{N/Q})$ to those of type $(k, N, \bar{\varepsilon}_Q \varepsilon_{N/Q})$. We let W_Q denote the operator $f \mapsto Q^{1-k/2} f|[w_Q]_k$ on $\mathcal{M}_k(\Gamma)$. Note that it is not the case in general that $W_Q^2 = (-1)^k$ on $\mathcal{M}_k(\Gamma)$, but that W_Q preserves $\mathcal{M}_k(\Gamma_0(N))$ and satisfies $W_Q^2 = 1$ on this subspace. (Recall that this subspace is trivial unless k is even.)

For the remainder of this section, we restrict our attention to the $\Gamma_0(N)$ situation and consider the involution of $\mathcal{M}_k(\Gamma_0(N))$ defined by W_Q. We find that it commutes with all the Hecke operators T_n with $(n, N) = 1$. If Q and Q' are divisors of N as above with $(Q, Q') = 1$ then the operators W_Q and $W_{Q'}$ commute and $W_Q W_{Q'} = W_{QQ'}$. Hence, $W_N = \prod_{p|N} W_{Q(p)}$ where, for a prime $p|N$, $Q(p) = p^r$ denotes the highest power of p dividing N.

Returning to the case $Q = N$, we find that the involution W_N on $\mathcal{S}_k(\Gamma_0(N))$ is self-adjoint relative to the Petersson product and commutes with all T_n such that $(n, N) = 1$. The decomposition of $\mathcal{S}_k(\Gamma_0(N))$ into simultaneous eigenspaces away from N is therefore compatible with its decomposition into W_N-eigenspaces. More precisely, if $E^\pm$ denote the latter eigenspaces (under W_N) with eigenvalues ± 1 so that $\mathcal{S}_k(\Gamma_0(N)) = E^+ \oplus E^-$ then this decomposition is $\mathbf{T}^{(N)}$-equivariant, i.e., we have this decomposition as $\mathbf{T}^{(N)}$-modules. It is in general not equivariant under the

full Hecke algebra since T_n's with $(n, N) \neq 1$ do not commute with the involution W_N.

5. L-function and functional equation

PRIMARY REFERENCES:
[**Miy2**, §4.3, 4.7], [**Ogg**, §I, IV], [**Shi1**, §3.6] and [**Kna2**, §VIII.5, IX.4].

In this section we define Dirichlet series attached to modular forms and briefly discuss the main results of Hecke's theory [**Hec2**], [**Hec3**] of such series. They admit analytic continuations, satisfy functional equations and, in certain cases, have Euler products.

Let

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

be the q-expansion of a modular form on $\Gamma_1(N)$ of weight k. Its coefficients satisfy $a_n = O(n^c)$ for some constant $c \in \mathbf{R}$. For example, the Eisenstein series E_k $(k = 4, 6, \dots)$ have this property with $c = k - 1$ since $\sigma_{k-1}(n) \leq 2n^{k-1}$ for $k > 2$. For cusp forms f on $\Gamma_1(N)$ of weight $k \geq 1$ (the case $k = 0$ is trivial), c may be taken to be $k/2$ from the fact that $|f(x+iy)|y^{k/2}$ is bounded on $\mathfrak{H}$. In general, if f is in $\mathcal{M}_k(\Gamma_1(N))$, the value $c = k-1$ will suffice if k is at least 3. In the cases where $k = 2$ or $= 1$, we may take $c = 1 + \epsilon$ and $c = 1/2$, respectively. This follows from the fact that modular forms of weight $k \geq 1$ are spanned by the cusp forms and the "Eisenstein series". The definition of Eisenstein series in this context is that given in [**Hec1**], and includes those appearing in Examples 2.2.2–2.2.6. (In fact, it can be shown that the space spanned by Eisenstein series is the orthogonal complement of the space of cusp forms under the Petersson inner product of Remark 3.6.2; see e.g. [**Ogg**, §IV] and Theorem 4.7.2 or §7.2 of [**Miy2**].) That their coefficients have the growth property stated above follows from Satz 9 of [**Hec1**]; see also Theorem 4.7.3 of [**Miy2**] or Theorem 7 of [**Schn**, Ch. IX].

REMARK 5.0.1. The Ramanujan-Petersson conjecture asserts that for p not dividing N, the eigenvalues of T_p on $\mathcal{S}_k(\Gamma_1(N))$ have absolute value bounded by $2p^{(k-1)/2}$. This was proved by Deligne [**Del1**, §5], [**Del5**] (see [**DeSe**, 9.1, 9.2] for $k = 1$). As a consequence, one can even take $c = (k-1)/2 + \epsilon$ (for any $\epsilon > 0$) for f in $\mathcal{S}_k(\Gamma_1(N))$, and $c = \epsilon$ for f in $\mathcal{M}_1(\Gamma_1(N))$.

The L-function of f is defined initially as the Dirichlet series

$$L(s, f) = \sum_{n=1}^{\infty} a_n n^{-s};$$

it is sometimes written $L(f, s)$ as well. Since $a_n = O(n^c)$ this series converges absolutely and uniformly in the region $\Re(s) \geq c + 1 + \delta$ (for any $\delta > 0$) and thus defines a holomorphic function in some right half-plane, at least in $\Re(s) > c + 1$. The completed L-function defined by

$$\Lambda(s, f) = N^{s/2}(2\pi)^{-s}\Gamma(s)L(s, f)$$

is essentially the Mellin transform of f; the reader can verify that

$$\Lambda(s, f) = N^{s/2} \int_0^{\infty} (f(iy) - a_0)y^s \frac{dy}{y}$$

whenever the integral is convergent.

In the case when f is a cusp form on $\Gamma_0(N)$ and an eigenfunction of the involution W_N, the L-function of f can be extended to the whole s-plane as an entire function with functional equation

$$\Lambda(s,f) = \epsilon\, i^k \Lambda(k-s,f)\,, \tag{5.0.1}$$

where $\epsilon = \pm 1$ is the eigenvalue of W_N. More generally, if f is a modular form of type (k,N,ε) then $\Lambda(s,f)$ extends to a meromorphic function with the functional equation

$$\Lambda(s,f) = i^k \Lambda(k-s, W_N f)\,.$$

For details, see [**Miy2**, Theorem 4.3.5], [**Shi1**, Theorem 3.66] or [**Ogg**, §1]. Note that $W_N f$ is a modular form of the same weight and level, but with character $\overline{\varepsilon}$. The only possible poles of $\Lambda(s,f)$ are simple ones at $s = 0$, k, and $\Lambda(s,f)$ is entire if f is a cusp form.

REMARK 5.0.2. Later (see §6.3), we shall discuss the notion of newforms. If f is a newform of level N which is also a common eigenform under all the Hecke operators T_p (including $p|N$), then $W_N f = c\tilde{f}$ where $\tilde{f} = \sum \bar{a}_n q^n$ is the contragredient of f and c a scalar. In particular, the functional equation may be rewritten as

$$\Lambda(s,f) = c\, i^k \Lambda(k-s, \tilde{f})\,,$$

which is analogous to that for Artin L-functions.

REMARK 5.0.3. A "converse theorem" due to Weil [**Weil**] (see also [**Ogg**, §V], [**Miy2**, §4.3] and [**JaLa**]) provides sufficient conditions for a Dirichlet series to be the L-function of a modular form. We will not state the conditions here, but only stress that they include functional equations.

Let f be a normalised $\mathbf{T}_N$-eigenform of type (k,N,ε). Then its L-function has an Euler product (see e.g. [**Shi1**, Theorem 3.43], [**Miy2**, Theorem 4.5.16]): if f has q-expansion $\sum \lambda_n q^n$ with $\lambda_1 = 1$ we have formally

$$L(s,f) = \prod_p \left(1 - \lambda_p p^{-s} + \varepsilon(p) p^{k-1-2s}\right)^{-1}. \tag{5.0.2}$$

Conversely if f is a modular form of type (k,N,ε) whose q-expansion coefficients are given by such an Euler product, then f is a $\mathbf{T}_N$-eigenform with λ_n as the eigenvalue of the n-th Hecke operator T_n for all $n \in \mathbf{N}$.

EXAMPLE 5.0.4. Let Δ be as in Example 2.2.7. It is a cuspidal $\mathbf{T}_1$-eigenform of weight 12 (of level 1 with trivial character). Its L-function is

$$L(s,\Delta) = \prod_p \left(1 - \tau(p) p^{-s} + p^{11-2s}\right)^{-1}$$

and $\Lambda(s,\Delta) = (2\pi)^{-s}\Gamma(s)L(s,\Delta)$ is entire and satisfies the functional equation which is invariant under $s \mapsto 12 - s$. Note that $W_1 \Delta = \Delta$ since $w_1 \in \mathrm{SL}_2(\mathbf{Z})$, so that $\epsilon = 1$.

EXAMPLE 5.0.5. Let f be the weight 2 cusp form of conductor 11 as in Example 2.2.8. One verifies that $W_{11} f = -f$. Hence, the completed L-function $\Lambda(s,f) = (2\pi/\sqrt{11})^{-s}\Gamma(s)L(s,f)$ is entire and satisfies the functional equation $\Lambda(s,f) = \Lambda(2-s,f)$.

EXAMPLE 5.0.6. Take an Eisenstein series of weight two with prime conductor, say $H(z) = E_2(z) - pE_2(pz)$ of Example 2.2.6 where p is a prime. Then, its L-function is -24 times

$$\sum_{n=1}^{\infty} \Big(\sum_{d|n} \varepsilon_0(d)d\Big) n^{-s} = \zeta(s)L(s-1, \varepsilon_0)$$

where ε_0 is the trivial character mod p; recall that $\varepsilon_0(d) = 0$ if $p|d$. There are simple poles of $\Lambda(s, H) = (2\pi/\sqrt{p})^{-s}\Gamma(s)L(s, H)$ at $s = 0$ and $s = 2(= k)$, owing respectively to the presence of $\Gamma(s)$ and

$$L(s-1, \varepsilon_0) = (1 - 1/p^{s-1})\zeta(s-1)\,.$$

There are no other poles, e.g. the pole of $\zeta(s)$ at $s = 1$ is cancelled by the zero of $L(s-1, \varepsilon_0)$ there (at $s = 1$, the Euler factor $(1 - 1/p^{s-1})$ is zero while $\zeta(s-1)$ is finite).

Similarly, if we take the weight one Eisenstein series $E_{1,\varepsilon}$ with an odd character ε mod p (with p prime) as in Example 2.2.4 then $L(s, E_{1,\varepsilon})$ is essentially $\zeta(s)L(s, \varepsilon)$.

6. Newforms and multiplicity one

In this section we explain some of the relationships between cusp forms, especially Hecke eigenforms, of different levels. The main result is the multiplicity one theorem of Atkin and Lehner [**AtLe**]. They consider only modular forms on $\Gamma_0(N)$, but here we follow [**Lang2**, Chapter VIII] for exposition of the theorem in $\Gamma_1(N)$ case. We shall return to the notion of multiplicity one from the point of view of automorphic representations in §11.

6.1. Old and new subspaces.

PRIMARY REFERENCES:
[**Lang2**, §VIII.1], [**Miy2**, §4.6] and [**AtLe**]

We consider the action of $\mathbf{T}_N$ and $\mathbf{T}^{(N)}$ on $\mathcal{S}_k(\Gamma_1(N))$, the space of cusp forms of weight k and level N. We shall fix the weight $k(\geq 1)$ throughout the section, but consider different levels.

Let d, M be positive integers such that dM divides N and let $\iota_d = \left(\begin{smallmatrix} d & 0 \\ 0 & 1 \end{smallmatrix}\right)$. If $f(z)$ is a modular form on $\Gamma_1(M)$, then $f|[\iota_d]_k(z) = d^{k-1}f(dz)$ is a modular form on $\Gamma_1(N)$ since $\iota_d^{-1}\Gamma_1(M)\iota_d$ contains $\Gamma_1(N)$. Moreover if f is a cusp form then so is $f|[\iota_d]_k$; so $f \mapsto f|[\iota_d]_k$ defines an injective map

$$\iota^*_{d,M,N} : \mathcal{S}_k(\Gamma_1(M)) \to \mathcal{S}_k(\Gamma_1(N)) \tag{6.1.1}$$

which we will denote ι_d^* when M and N are fixed.

Let us examine the extent to which ι_d^* is compatible with the action of the Hecke operators. Using (3.4.2) we find that if p is a prime not dividing N, then

$$f|[\iota_d]_k|T(p)_k = f|T(p)_k|[\iota_d]_k$$

where $T(p)_k$ in the left side of the equation is relative to level N while that in the right side is relative to level M. A similar statement holds for $T(p, p)$ if p does not divide N. Thus ι_d^* is a homomorphism of $\mathbf{T}^{(N)}$-modules where we regard $\mathcal{S}_k(\Gamma_1(M))$ as a module for $\mathbf{T}^{(N)}$ using the obvious inclusion $\mathbf{T}^{(N)} \subset \mathbf{T}^{(M)}$. In particular, if f in $\mathcal{S}_k(\Gamma_1(M))$ is a $\mathbf{T}^{(M)}$-eigenform then $f|[\iota_d]_k$ in $\mathcal{S}_k(\Gamma_1(N))$, for d dividing N/M, is a $\mathbf{T}^{(N)}$-eigenform with the same eigenvalues away from N.

REMARK 6.1.1. The map (6.1.1) commutes with $T(n)$ if $(n,d)=1$, but generally fails to commute otherwise.

For a fixed N, the linear span of the images of the maps $\iota_{d,M,N}^*$ over all d, M with $dM|N$, $M \neq N$, is called the *old subspace* of $\mathcal{S}_k(\Gamma_1(N))$, and is denoted $\mathcal{S}_k(\Gamma_1(N))^{\mathrm{old}}$. We define the *new subspace* of $\mathcal{S}_k(\Gamma_1(N))$, denoted $\mathcal{S}_k(\Gamma_1(N))^{\mathrm{new}}$, as the orthogonal complement of $\mathcal{S}_k(\Gamma_1(N))^{\mathrm{old}}$ in $\mathcal{S}_k(\Gamma_1(N))$ with respect to the Petersson inner product. One checks that the space $\mathcal{S}_k(\Gamma_1(N))^{\mathrm{old}}$ is stable under the action of $\mathbf{T}^{(N)}$ on $\mathcal{S}_k(\Gamma_1(N))$, and it follows that so is $\mathcal{S}_k(\Gamma_1(N))^{\mathrm{new}}$. Moreover we can write

$$\mathcal{S}_k(\Gamma_1(N)) = \sum_{dM|N} \left(\mathcal{S}_k(\Gamma_1(M))^{\mathrm{new}}\right)|[\iota_d]_k \tag{6.1.2}$$

as a $\mathbf{T}^{(N)}$-module with the space $\mathcal{S}_k(\Gamma_1(N))^{\mathrm{old}}$ of oldforms given by

$$\sum_{dM|N, M\neq N} \mathcal{S}_k(\Gamma_1(M))|[\iota_d]_k = \sum_{dM|N, M\neq N} \left(\mathcal{S}_k(\Gamma_1(M))^{\mathrm{new}}\right)|[\iota_d]_k\,.$$

For each M, the $\mathbf{T}^{(M)}$-module $\mathcal{S}_k(\Gamma_1(M))^{\mathrm{new}}$ admits a basis consisting of $\mathbf{T}^{(M)}$-eigenforms. Thus $\mathcal{S}_k(\Gamma_1(N))$ has a basis consisting of $\mathbf{T}^{(N)}$-eigenforms $\{f\}$, where each f is of the following form: $f = g_i|[\iota_d]_k$ with $g_i \in \mathcal{S}_k(\Gamma_1(M))^{\mathrm{new}}$ for some positive integers d, M such that $dM|N$ and g_i is a $\mathbf{T}^{(M)}$-eigenform.

REMARK 6.1.2. One can check directly that the space $\mathcal{S}_k(\Gamma_1(N))^{\mathrm{old}}$ is stable under the action of the bigger ring $\mathbf{T}_N$. We shall see later that the same is true for $\mathcal{S}_k(\Gamma_1(N))^{\mathrm{new}}$; moreover this space is spanned by $\mathbf{T}_N$-eigenforms called "newforms" or "primitive" forms. We shall also see that the sum in (6.1.2) is actually a direct sum decomposition. See the discussion following Corollary 6.3.1, especially Remark 6.3.4.

The maps $\iota_{d,M,N}^*$ commute with the action of $(\mathbf{Z}/N\mathbf{Z})^\times$ where we define the action of $(\mathbf{Z}/N\mathbf{Z})^\times$ on $\mathcal{S}_k(\Gamma_1(M))$ via the natural projection $(\mathbf{Z}/N\mathbf{Z})^\times \to (\mathbf{Z}/M\mathbf{Z})^\times$. It follows that we have a $\mathbf{T}^{(N)}$-equivariant decomposition

$$\mathcal{S}_k(\Gamma_1(N))^{\mathrm{old}} = \bigoplus_{\varepsilon} \mathcal{S}_k(N,\varepsilon)^{\mathrm{old}}$$

over Dirichlet characters ε mod N where $\mathcal{S}_k(N,\varepsilon)^{\mathrm{old}} = \mathcal{S}_k(N,\varepsilon) \cap \mathcal{S}_k(\Gamma_1(N))^{\mathrm{old}}$. We have an analogous decomposition of the new subspace into eigenspaces $\mathcal{S}_k(N,\varepsilon)^{\mathrm{new}}$, and these satisfy

$$\mathcal{S}_k(N,\varepsilon) = \mathcal{S}_k(N,\varepsilon)^{\mathrm{old}} \oplus \mathcal{S}_k(N,\varepsilon)^{\mathrm{new}}.$$

We can replace Γ_1 by Γ_0 in the appropriate definitions above to obtain maps ι_d^* and old and new subspaces $\mathcal{S}_k(\Gamma_0(N))^{\mathrm{old}}$ and $\mathcal{S}_k(\Gamma_0(N))^{\mathrm{new}}$ of $\mathcal{S}_k(\Gamma_0(N))$. These spaces coincide with $\mathcal{S}_k(N,\varepsilon)^{\mathrm{old}}$ and $\mathcal{S}_k(N,\varepsilon)^{\mathrm{new}}$ where ε is the trivial character. In fact, for the space of cusp forms of type (k,N,ε) we may define the old and new subspaces intrinsically as follows. Given $M|N$, a Dirichlet character χ mod M gives rise to a Dirichlet character mod N via the natural projection $(\mathbf{Z}/N\mathbf{Z})^\times \to (\mathbf{Z}/M\mathbf{Z})^\times$; denote it by χ_N. Moreover given a Dirichlet character ε mod N, there is at most one character χ mod M such that $\chi_N = \varepsilon$. Indeed, if ε_0 denotes the primitive character associated to ε and C its conductor, then χ exists only if $C|M$, in which case $\chi = (\varepsilon_0)_M$. Since ι_d^* respects the action of $(\mathbf{Z}/N\mathbf{Z})^\times$ for $dM|N$, we see that $f \mapsto f|[\iota_d]_k$ defines a map $\mathcal{S}_k(M,\chi) \to \mathcal{S}_k(N,\chi_N)$ which we again denote ι_d^*.

Then $\mathcal{S}_k(N,\varepsilon)^{\text{old}}$ is simply the linear span of the images of $\iota_d^* : \mathcal{S}_k(M,\chi) \to \mathcal{S}_k(N,\varepsilon)$ over all integers M and d such that $M \neq N$, $C|M|N$, χ is the Dirichlet character mod M with $\chi_N = \varepsilon$ and d divides N/M. The orthogonal complement of $\mathcal{S}_k(N,\varepsilon)^{\text{old}}$ in $\mathcal{S}_k(N,\varepsilon)$ relative to the Petersson product is $\mathcal{S}_k(N,\varepsilon)^{\text{new}}$. In particular, if ε mod N is primitive then $\mathcal{S}_k(N,\varepsilon)^{\text{new}} = \mathcal{S}_k(N,\varepsilon)$.

REMARK 6.1.3. The maps ι_d^* are essentially pullback homomorphisms induced by the degeneracy maps $Y_0(N) \to Y_0(M)$ defined in §7.3 (with the roles of M and N reversed).

EXAMPLE 6.1.4. We describe the decomposition of $\mathcal{S}_2(\Gamma_1(33))$ into its old and new subspaces. By the dimension formulas in §12.1, we find that $\mathcal{S}_2(\Gamma_1(33))$ is 21-dimensional. We also find that $\mathcal{S}_2(\Gamma_1(3)) = 0$ and that $\mathcal{S}_2(\Gamma_1(11)) = \mathcal{S}_2(\Gamma_0(11))$ is one-dimensional and therefore generated by a normalized $\mathbf{T}_{11}$-eigenform f. Therefore $\mathcal{S}_2(\Gamma_1(33))^{\text{old}} = \mathcal{S}_2(\Gamma_0(33))^{\text{old}}$ is spanned by the linearly independent forms $f(z)$ and $f(3z)$. The space $\mathcal{S}_2(\Gamma_1(33))^{\text{new}}$ decomposes as

$$\bigoplus_{\varepsilon} \mathcal{S}_2(33,\varepsilon)^{\text{new}},$$

where ε runs over the 10 Dirichlet characters mod 33 which are "even" in the sense that $\varepsilon(-1) = 1$. For the trivial character ε, we have that $\mathcal{S}_2(33,\varepsilon)^{\text{new}} = \mathcal{S}_2(\Gamma_0(33))^{\text{new}}$ is one-dimensional generated by a $\mathbf{T}_{33}$-eigenform. Applying the dimension formulas to groups intermediate to $\Gamma_1(33)$ and $\Gamma_0(33)$, and using the second part of Proposition 12.3.11, we find that $\mathcal{S}_2(33,\varepsilon)^{\text{new}} = \mathcal{S}_2(33,\varepsilon)$ is two-dimensional for each non-trivial even ε. We shall see from the theory of new-forms that each is spanned by $\mathbf{T}_{33}$-eigenforms. Moreover, while $f(z)$ and $f(3z)$ are not $\mathbf{T}_{33}$-eigenforms, suitable linear combinations will be, so that in this example, $\mathcal{S}_2(\Gamma_1(N))$ is spanned by $\mathbf{T}_N$-eigenforms. This is not the case in general. For example, the reader may check that the subspace of $\mathcal{S}_2(\Gamma_1(297))$ spanned by $f(z)$, $f(3z)$, $f(9z)$ and $f(27z)$ is stable under T_3 but does not have a basis of eigenforms for T_3.

Let us also note how the W-operators behave with respect to the maps ι_d^*. We find that

$$g|[\iota_d]_k|[w_{dM}]_k = d^{k-2}g|[w_M]_k \qquad \text{for } g \in \mathcal{S}_k(\Gamma_1(M))\,,$$

or equivalently,

$$f|[w_{dM}]_k = f|[w_M]_k|[\iota_d]_k \qquad \text{for } f \in \mathcal{S}_k(\Gamma_1(M))$$

since $[w_N]_k^2 = (-N)^{k-2}$ on $\mathcal{S}_k(\Gamma_1(N))$. Thus $W_N = (-N)^{1-k/2}[w_N]_k$ preserves the spaces $\mathcal{S}_k(\Gamma_1(N))^{\text{new}}$ and $\mathcal{S}_k(\Gamma_1(N))^{\text{old}}$, and gives an isomorphism

$$\mathcal{S}_k(N,\varepsilon)^{\text{new}} \longrightarrow \mathcal{S}_k(N,\bar{\varepsilon})^{\text{new}}$$

and an analogous isomorphism for the old subspaces.

6.2. Multiplicity one theorem.

PRIMARY REFERENCES:

[**Lang2**, §VIII.3, VIII.4], [**Miy2**, §4.6] and [**AtLe**].

Let $\mathbf{T}^{(N)}$ be as before. In addition to $\mathbf{T}^{(N)}$-eigenforms in $\mathcal{S}_k(\Gamma_1(N))$, we shall consider forms which are simultaneous eigenvectors under T_p for *almost all* primes p. For this, we introduce an auxilliary positive integer D and consider the action of $\mathbf{T}^{(ND)}$. Then, for $f \in \mathcal{S}_k(\Gamma_1(N))$ the following are equivalent by arguments similar to those used for Proposition 3.5.1:

- f is of type (k, N, ε) for some ε mod N and is a $\mathbf{T}^{(ND)}$-eigenform;
- f is a $\mathbf{T}^{(ND)}$-eigenform;
- f is a common eigenform under T_n for all $(n, ND) = 1$.

A form $f \in \mathcal{S}_k(\Gamma_1(N))$ is a common eigenform under T_p for almost all (i.e., all but finitely many) primes p if and only if there exists D such that f is a $\mathbf{T}^{(ND)}$-eigenform.

Let f be such an eigenform of weight $k \geq 1$. We can associate to f a homomorphism $\theta_f : \mathbf{T}^{(ND)} \to \mathbf{C}$ defined by $Tf = \theta_f(T)f$; we call θ_f the *eigencharacter* of f. Since the $\mathbf{T}^{(N)}$-eigenspace to which f belongs contains a non-zero $\mathbf{T}_N$-eigenform whose eigenvalues away from N are the same as those of f, the image $\theta_f(\mathbf{T}^{(ND)})$ is a subring of the ring generated by the eigenvalues of the $\mathbf{T}_N$-eigenform and therefore (see Remark 3.5.3 and Corollary 12.4.5) contained in the ring of integers of an algebraic number field (of finite degree over $\mathbf{Q}$). Note that the values of ε associated to f already lie in $\theta_f(\mathbf{T}^{(ND)})$.

Now, the main result of Atkin-Lehner theory is the multiplicity one theorem, which essentially says that an eigencharacter occurring in the new subspace $\mathcal{S}_k(\Gamma_1(N))^{\text{new}}$ does so with multiplicity one. This is a consequence of the following key fact in the theory whose proof we omit. (See e.g. [**Lang2**, §VIII.4].)

PROPOSITION 6.2.1. *Let $f = \sum_1^\infty a_n q^n$ be a cusp form on $\Gamma_1(N)$ and suppose there is an integer $D \geq 1$ such that for all $(n, ND) = 1$ we have $a_n = 0$. Then there exists a cusp form g_p on $\Gamma_1(N/p)$ for each prime $p|N$ such that*

$$f = \sum_{p|N} \iota_p^* g_p \,,$$

i.e., $f \in \mathcal{S}_k(\Gamma_1(N))^{\text{old}}$.

This implies the following

COROLLARY 6.2.2. *Let $f = \sum a_n q^n$ be a cusp form on $\Gamma_1(N)$ which is a simultaneous eigenfunction under T_p for almost all primes $p \nmid N$. If $a_1 = 0$ then f is in the old subspace.*

Indeed, with an auxilliary integer D chosen in an obvious way so that $T_p f = \lambda_p f$ for all $p \nmid ND$, if $a_1 = 0$ then from $a_{np} + \varepsilon(p)p^{k-1}a_{n/p} = \lambda_p a_n$ we get that $a_{p^\nu} = 0$ for such p for all ν by induction. (Recall that f of the proposition is necessarily of type (N, ε) for some Dirichlet character ε mod N.) Hence, $a_n = 0$ for all $(n, ND) = 1$. By Proposition 6.2.1, f is then in the old subspace.

In view of this, any (non-zero) $\mathbf{T}^{(ND)}$-eigenform f in $\mathcal{S}_k(\Gamma_1(N))^{\text{new}}$ can be normalised to have the first coefficient $a_1 = 1$. The multiplicity one theorem is

THEOREM 6.2.3. *Let f, $g \in \mathcal{S}_k(\Gamma_1(N))$ be $\mathbf{T}^{(ND)}$-eigenforms with the same eigencharacters, i.e., $\theta_f(T_p) = \theta_g(T_p)$ for all $p \nmid ND$. If $f \in \mathcal{S}_k(\Gamma_1(N))^{\text{new}}$, f normalized, then g is a scalar multiple of f. (In particular, if g is in the old subspace, then $g = 0$.)*

Proof. If $g \neq 0$ is in the new subspace, then we may assume that it is normalised, so that $f - g$ is a $\mathbf{T}^{(ND)}$-eigenform in the new subspace with the first coefficient 0. So $f - g$ is also in the old subspace by Proposition 6.2.2, hence $g - f = 0$. If g is in the old subspace then it is a linear combination of functions $\iota_d^* g_i$ where $g_i \in \mathcal{S}_k(\Gamma_1(M))^{\text{new}}$, $M \neq N$, $dM|N$, and where each g_i is an eigenform under $\mathbf{T}^{(ND)} = \mathbf{T}^{(MD')}$ with $ND = MD'$ for some D'. Note that $\theta_{g_i}(T_p) = \theta_g(T_p)$ for p not dividing ND. Unless $g = 0$, there is some i such that $a_1(g_i) \neq 0$ by Proposition

6.2.2 (at level M instead of N), and so there is a constant c such that $a_1(f-cg_i)=0$. As $f-cg_i \in \mathcal{S}_k(\Gamma_1(N))$ is a $\mathbf{T}^{(ND)}$-eigenform, Proposition 6.2.2 implies that $f-cg_i$ is old at level N, which in turn means that f is old at level N (because g_i is already old at that level). But $f \neq 0$ is also new of level N by assumption, and we have a contradiction, unless $g=0$. Finally, if $g=g^{\text{new}}+g^{\text{old}}$ then each component has the same eigencharacter as g (or f), and so the above argument shows that $g^{\text{old}}=0$, while $g=g^{\text{new}}$ is a multiple of f.

6.3. Newforms.

PRIMARY REFERENCES:
[**Lang2**, §VIII.3], [**Miy2**, §4.6] and [**AtLe**].

We have already noted that for a $\mathbf{T}^{(ND)}$-eigenform g on $\Gamma_1(M)$ (with $M \neq N$, $M|N$), the two forms $g(=\iota_1^* g)$ and $\iota_d^* g$ (with $d>1$, $dM|N$) have the same $\mathbf{T}^{(ND)}$-eigencharacter. If in addition $g(\neq 0)$ is in the new subspace of level M, then these two forms are linearly independent. Thus, we have the following corollary to Theorem 6.2.3.

COROLLARY 6.3.1. *The subspace $\mathcal{S}_k(\Gamma_1(N))^{\text{new}}$ (respectively, $\mathcal{S}_k(\Gamma_1(N))^{\text{old}}$) of $\mathcal{S}_k(\Gamma_1(N))$ is the orthogonal sum of the $\mathbf{T}^{(ND)}$-eigenspaces in $\mathcal{S}_k(\Gamma_1(N))$ whose eigencharacters occur with multiplicity one (respectively, >1).*

The same is true of course if we consider eigenforms under $\mathbf{T}^{(ND)}$ in the old and new subspaces of $\mathcal{S}_k(N,\varepsilon)$, since an eigencharacter determines the Nebentypus.

The corollary implies that $\mathcal{S}_k(\Gamma_1(N))^{\text{new}}$ is stable under the action of the full Hecke ring $\mathbf{T}_N$. In fact, a $\mathbf{T}^{(ND)}$-eigenspace in $\mathcal{S}_k(\Gamma_1(N))^{\text{new}}$ or $\mathcal{S}_k(N,\varepsilon)^{\text{new}}$ is one-dimensional and is therefore stable under $\mathbf{T}_N$ since the Hecke operators all commute. Therefore a $\mathbf{T}^{(ND)}$-eigenform in $\mathcal{S}_k(\Gamma_1(N))^{\text{new}}$ is necessarily a $\mathbf{T}_N$-eigenform. Thus the following are equivalent in the new subspace of level N:

(i) f is a $\mathbf{T}_N$-eigenform;
(ii) f is a $\mathbf{T}^{(N)}$-eigenform;
(iii) f is a $\mathbf{T}^{(ND)}$-eigenform for some D.

Recall also that such a form f can be normalized so that $a_1=1$. A normalized eigenform in $\mathcal{S}_k(\Gamma_1(N))^{\text{new}}$ is called a *newform* (or a *primitive* cusp form) of level N.

REMARK 6.3.2. Conditions (ii) and (iii) are equivalent for f in $\mathcal{S}_k(\Gamma_1(N))$, but they do not imply (i).

REMARK 6.3.3. The multiplicity one theorem holds also for forms of different levels. Let $f_i \in \mathcal{S}_k(\Gamma_1(N_i))^{\text{new}}$ $(i=1,2)$ be two normalized Hecke eigenforms with eigenvalues a_p^i under T_p for primes p. Suppose $a_p^1=a_p^2$ for all but finitely many primes p. Then we must have $f_1=f_2$. This follows from the multiplicity one Theorem 6.2.3 once the equality $N_1=N_2$ of their levels is established by considering their functional equations (Remark 5.0.2); see e.g. [**Miy2**, §4.6]. Thus for a $\mathbf{T}^{(ND)}$-eigenspace of $\mathcal{S}_k(\Gamma_1(N))$, there is a unique pair (f,M) such that f is in the eigenspace and is a newform of level M.

REMARK 6.3.4. We have mentioned, in Remark 6.1.2, that the decomposition (6.1.2) is actually a direct sum. This can be seen as follows: First, note that $\mathcal{S}_k(\Gamma_1(N))$ is an orthogonal sum of $\mathbf{T}^{(N)}$-eigenspaces. Let g be a newform of level M, and suppose M divides N. Then for every positive integer d dividing N/M,

$\iota_d^* g$ is a $\mathbf{T}^{(N)}$-eigenform belonging to the same $\mathbf{T}^{(N)}$-eigensubspace of $\mathcal{S}_k(\Gamma_1(N))$ to which g belongs. In this eigenspace, we have a direct sum $\oplus_d \mathbf{C}\iota_d^* g$, d over divisors of N/M, because the set $\{g(z), g(2z), g(3z), \dots\}$ is linearly independent over $\mathbf{C}$ for such a form g. Now, for each $M|N$, let $\{g_1^M, \dots, g_{j_M}^M\}$ be the set of newforms of level M, and let θ_j^M $(1 \le j \le j_M)$ denote the corresponding eigencharacters. Note that the set is actually a basis for $\mathcal{S}_k(\Gamma_1(M))^{\text{new}}$. Now by multiplicity one (Theorem 6.2.3) and Remark 6.3.3 these θ_j^M, over all $M|N$ and $1 \le j \le j_M$, are distinct as eigencharacters of $\mathbf{T}^{(N)}$. Hence, the forms $g_j^M \in \mathcal{S}_k(\Gamma_1(N))$ belong to mutually distinct $\mathbf{T}^{(N)}$-eigenspaces. These arguments thus yield the following direct sum

$$\bigoplus_{M|N} \bigoplus_{j=1}^{j_M} \Big(\bigoplus_{d:\, dM|N} \mathbf{C}\iota_d^* g_j^M \Big)$$

in $\mathcal{S}_k(\Gamma_1(N))$. Interchanging the two inner sums gives precisely the sum appearing in (6.1.2), but with $\oplus$ in place of Σ.

Let $f = \sum \lambda_n q^n$ be the newform in a $\mathbf{T}^{(N)}$-eigenspace of $\mathcal{S}_k(\Gamma_1(N))$ (or equivalently a $\mathbf{T}^{(ND)}$-eigenspace for some D). Since f is a $\mathbf{T}_M$-eigenform for some M dividing N, its L-function $L(s, f)$ has an Euler product (5.0.2) where ε is a character mod M. The L-functions of the $\mathbf{T}_N$-eigenforms in the $\mathbf{T}^{(N)}$-eigenspace are obtained by simple modifications of the Euler factors of $L(s, f)$ at primes dividing N/M.

EXAMPLE 6.3.5. Returning to Example 6.1.4, we see that $\mathcal{S}_2(\Gamma_1(33))$ decomposes into 20 $\mathbf{T}^{(33)}$-eigenspaces, two for each even Dirichlet character mod 33. One of those is two-dimensional, but the rest are one-dimensional, generated by a newform of level 33. The two-dimensional $\mathbf{T}^{(33)}$-eigenspace is generated by $f(z)$ and $f(3z)$ where f (Example 2.2.8) is a newform of level 11 with trivial Nebentypus and $\lambda_3 = -1$. Letting α_3 and β_3 denote the roots of $X^2 + X + 3 = 0$, one finds that this $\mathbf{T}^{(33)}$-eigenspace is generated by the $\mathbf{T}_{33}$-eigenforms $f_1 = f - \alpha_3 f(3z)$ and $f_2 = f - \beta_3 f(3z)$. The L-function $L(s, f)$ (Example 5.0.5) has an Euler product for which the Euler factor at 3 is $L_3(s, f) = (1 + 3^{-s} + 3^{1-2s})^{-1} = [(1 - \alpha_3 3^{-s})(1 - \beta_3 3^{-s})]^{-1}$. The L-function $L(s, f_1)$ (resp. $L(s, f_2)$) are obtained from $L(s, f)$ by replacing $L_3(s, f)$ with $(1 - \beta_3 3^{-s})^{-1}$ (resp. $(1 - \alpha_3 3^{-s})^{-1}$).

Let us also consider $N = 297$. One finds that the 4-dimensional $\mathbf{T}^{(N)}$-eigenspace of $\mathcal{S}_2(\Gamma_1(N))$ generated by $f(z)$, $f(3z)$, $f(9z)$ and $f(27z)$ contains only three normalized $\mathbf{T}_N$-eigenforms and that their L-functions are obtained from $L(s, f)$ by replacing $L_3(s, f)$ by $(1 - \beta_3 3^{-s})^{-1}$, $(1 - \alpha_3 3^{-s})^{-1}$ and 1.

Let us now consider the $\Gamma_0(N)$ situation. The involution W_N commutes with the Hecke operators T_p for all $p \nmid N$ (p prime) so that a newform f of level N is also an eigenvector for W_N (with eigenvalue $\epsilon = \pm 1$); similarly it is an eigenvector for $W_{Q(p)}$ for all $p|N$, with corresponding eigenvalues $\epsilon(Q(p)) = \pm 1$. Consequently, $L(s, f)$ satisfies the functional equation (5.0.1) with $\epsilon = \prod_{p|N} \epsilon(Q(p))$. Moreover, $\lambda_p = 0$ if $p^2|N$, while $\lambda_p = -p^{k/2-1}\epsilon(p)$ if $p \| N$. This last assertion is obtained using the following fact (see [**AtLe**, Lemma 7]): For f in $\mathcal{S}_k(\Gamma_0(N))$, $T_p f$ is a cusp form on $\Gamma_0(N/p)$ if $p^2|N$, while $T_p f + p^{k/2-1} W_p f$ is on $\Gamma_0(N/p)$ if $p \| N$. Indeed, in either case, the given form of level N/p is in $\mathcal{S}_k(\Gamma_0(N))^{\text{old}}$ and so, having the same $\mathbf{T}^{(N)}$-eigencharacter as the newform f, must be 0. Since $T_p f = \lambda_p f$, while $W_p f = \epsilon(p) f$ in the second case, we obtain the desired values for λ_p when $p|N$.

Part II. Modular curves

7. Elementary theory

Recall that in §3.2 we defined the modular curve associated to a congruence subgroup Γ of $\mathrm{SL}_2(\mathbf{Z})$ as the quotient space $\Gamma\backslash\mathfrak{H}$ where the action of $\mathrm{SL}_2(\mathbf{Z})$ on the complex upper-half plane $\mathfrak{H}$ is given by linear fractional transformations. We have thus defined a modular curve simply as a topological space, but we shall interpret it in §7.2 as a moduli space for elliptic curves. This interpretation will yield in §8 a natural algebraic-geometric description of the curve as the set of complex points of a "moduli scheme."

7.1. Topological structure.

PRIMARY REFERENCES:
[**Shi1**, §1.3–1.5], [**Ser1**, §VII.1], [**Lang2**, §III.1, III.2] and [**Miy2**, Ch. 1].

We first describe the topological structure of the modular curve $Y = \mathrm{SL}_2(\mathbf{Z})\backslash\mathfrak{H}$ by giving a convenient set of representatives in $\mathfrak{H}$ for this quotient. (See [**Ser1**, §VII.1] and [**Shi1**, §I.4].) As the diagonal matrix -1 acts trivially, we have $Y = \mathrm{PSL}_2(\mathbf{Z})\backslash\mathfrak{H}$. The group $\mathrm{PSL}_2(\mathbf{Z}) = \mathrm{SL}_2(\mathbf{Z})/\{\pm 1\}$ is generated by the elements

$$S = \pm\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \qquad T = \pm\begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}$$

with relations $S^2 = T^3 = 1$. Every element z of $\mathfrak{H}$ can be written in the form $\gamma(z')$ for some $\gamma \in \mathrm{PSL}_2(\mathbf{Z})$ and some z' in the set

$$D = \{x + iy \in \mathfrak{H} \mid x^2 + y^2 \geq 1, |x| \leq 1/2\}.$$

Letting D' denote the interior of D together with the subset of the boundary satisfying $x \geq 0$, we find that for every z there is a unique z' in D' such that z is in $\mathrm{PSL}_2(\mathbf{Z})z'$. The element γ of $\mathrm{PSL}_2(\mathbf{Z})$ such that $z = \gamma(z')$ is not necessarily unique, but the only points of D' with nontrivial stabilizers are i and $\zeta = e^{\pi i/3}$. Their stabilizers are the groups $\langle S\rangle$ and $\langle T\rangle$ respectively. Observe that the points of D' are in one-to-one correspondence with the points of Y, but the two topological spaces are not homeomorphic. Rather the topological space Y can be constructed from D as the quotient space obtained by identifying z with $-\overline{z}$ for boundary points of D; thus Y is homeomorphic to $\mathbf{R}^2$.

REMARK 7.1.1. A "nice" set of representatives in $\mathfrak{H}$ (or for some authors, its closure, and for others, its interior) for the modular curve $\Gamma\backslash\mathfrak{H}$ is called a "fundamental domain" for Γ. We shall not give a precise definition here, but remark only that D is a fundamental domain for $\mathrm{SL}_2(\mathbf{Z})$ and that for any Γ there is a fundamental domain of the form $\cup\gamma D$ where γ runs over a suitable set of coset representatives for $\Gamma\backslash\mathrm{SL}_2(\mathbf{Z})$. (See [**Shi1**, §1.4] and [**Miy2**, §1.6].)

The spaces $\Gamma\backslash\mathfrak{H}$ are Hausdorff and inherit from $\mathfrak{H}$ the structure of a one-dimensional complex manifold [**Shi1**, §1.5]. If the image $\overline{\Gamma}$ of Γ in $\mathrm{PSL}_2(\mathbf{Z})$ has no elements of finite order, then $\overline{\Gamma}$ acts without fixed points on $\mathfrak{H}$. This is the case for example if $\Gamma = \Gamma_1(N)$ with $N > 3$, and then the local homeomorphism $\mathfrak{H} \to \Gamma\backslash\mathfrak{H}$ fully describes the complex structure on the quotient. Slightly more care is required if $\Gamma\backslash\mathfrak{H}$ has elliptic points. These are points for which a preimage in $\mathfrak{H}$ has a non-trivial stabilizer, necessarily of finite order, in $\overline{\Gamma}$. Note that there are only two elliptic points on $Y = \mathrm{SL}_2(\mathbf{Z})\backslash\mathfrak{H}$; they are $\mathrm{PSL}_2(\mathbf{Z})i$ and $\mathrm{PSL}_2(\mathbf{Z})\zeta$. The function

$f(z) = ((z-i)/(z+i))^2$ defines a homeomorphism from a neighborhood of $\mathrm{PSL}_2(\mathbf{Z})i$ in Y to a neighborhood of the origin in $\mathbf{C}$. The complex structure at $\mathrm{PSL}_2(\mathbf{Z})i$ is then given by f and a similar function works in a neighborhood of $\mathrm{PSL}_2(\mathbf{Z})\zeta$. The resulting complex manifold Y is biholomorphic to the complex plane. For an arbitrary congruence subgroup Γ, the natural projection $\Gamma\backslash\mathfrak{H} \to \mathrm{SL}_2(\mathbf{Z})\backslash\mathfrak{H}$ maps the finite set of elliptic points to $\mathrm{SL}_2(\mathbf{Z})i \cup \mathrm{SL}_2(\mathbf{Z})\zeta$. To complete the description of the complex structure of $\Gamma\backslash\mathfrak{H}$, one can use the fact that this projection is a homeomorphism in a neighborhood of each elliptic point.

Note that the curves $\Gamma\backslash\mathfrak{H}$ are not compact. We shall explain in §9.1 how they are compactified by the addition of "cusps" to obtain a Riemann surface.

7.2. Moduli spaces.

PRIMARY REFERENCES:
[**Huse**, §11.1, 11.2] and [**Sil1**, Appendix C §13]

We are especially interested in the curves associated to $\Gamma_0(N)$ and $\Gamma_1(N)$ for positive integers N, and we denote these curves $Y_0(N)$ and $Y_1(N)$ respectively. We always wish to bear in mind their interpretation as "moduli spaces."

We begin with $Y_0(N)$, whose points are naturally in bijection with isomorphism classes of pairs (E, C) where E is an elliptic curve over $\mathbf{C}$ and C is a cyclic subgroup of E of order N. (We consider the pairs (E, C) and (E', C') to be isomorphic if there is an isomorphism $\phi : E \to E'$ such that $\phi(C) = C'$.) To establish the bijection, simply associate to $\tau \in \mathfrak{H}$ the pair

$$\mathbf{E}_\tau = (\mathbf{C}/\Lambda_\tau, \frac{1}{N}\mathbf{Z}/\mathbf{Z})$$

where Λ_τ is the lattice $\mathbf{Z}+\mathbf{Z}\tau$. One checks that any pair (E, C) is isomorphic to $\mathbf{E}_\tau$ for some $\tau \in \mathfrak{H}$, and that $\mathbf{E}_\tau$ is isomorphic to $\mathbf{E}_{\tau'}$ if and only if $\tau' \in \Gamma_0(N)\tau$. Note that if $N = 1$ then $Y_0(N)$ is simply the set of isomorphism classes of elliptic curves. As an elliptic curve over $\mathbf{C}$ is determined up to isomorphism by its j-invariant, the map $\tau \mapsto j(E_\tau)$ defines a bijection $Y_0(1) \to \mathbf{C}$.

Similarly the points of $Y_1(N)$ are in bijection with isomorphism classes of pairs (E, P) where E is an elliptic curve and P is a point of E of order N. (For $\tau \in \mathfrak{H}$, use $E = \mathbf{C}/\Lambda_\tau$ and $P = 1/N \bmod \Lambda_\tau$.) The action of $\Gamma_0(N)$ on $\mathfrak{H}$ induces an action of $\Gamma_0(N)/\Gamma_1(N)$ on $Y_1(N)$. Using the isomorphism $\Gamma_0(N)/\Gamma_1(N) \cong (\mathbf{Z}/N\mathbf{Z})^\times$ defined by $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto d \bmod N$, we view $(\mathbf{Z}/N\mathbf{Z})^\times$ as acting on $Y_1(N)$ as well. The corresponding automorphism $\langle d \rangle$ of $Y_1(N)$ has the moduli-theoretic interpretation $(E, P) \mapsto (E, dP)$. Note that $\langle -1 \rangle$ is the identity, so the action of $(\mathbf{Z}/N\mathbf{Z})^\times$ factors through $(\mathbf{Z}/N\mathbf{Z})^\times/\{\pm 1\}$. Note also that $Y_0(N)$ is naturally the quotient of $Y_1(N)$ by the action of this group and the natural projection $Y_1(N) \to Y_0(N)$ has the simple moduli-theoretic interpretation $(E, P) \mapsto (E, \langle P \rangle)$ where $\langle P \rangle$ is the subgroup of E generated by P.

7.3. Modular correspondences revisited.

PRIMARY REFERENCES:
[**Ser1**, §VII.5], [**Sil2**, §I.9], [**Kna2**, §VIII.7] and [**Kobl**, §III.5].

We have already considered in §3.2 certain natural projection maps or "degeneracy maps" between modular curves and used these maps to define correspondences. Let us return to this matter from a more moduli-theoretic point of view.

If M is a multiple of N, then $\Gamma_0(M)$ is contained in $\Gamma_0(N)$ and there is a natural projection from $Y_0(M)$ to $Y_0(N)$. There are in fact a number of natural degeneracy maps from $Y_0(M)$ to $Y_0(N)$; for any divisor d of M/N, we have that $\iota_d\Gamma_0(M)\iota_d^{-1} \subset \Gamma_0(N)$ so that $\tau \mapsto d\tau$ induces a map from $Y_0(M)$ to $Y_0(N)$. Here ι_d denotes the matrix $\begin{pmatrix} d & 0 \\ 0 & 1 \end{pmatrix}$ as introduced in §3.2. We are especially interested in the case where $M = Np$ for a prime p, and we denote by α and β the degeneracy maps defined, respectively, by $\tau \mapsto \tau$ and $\tau \mapsto p\tau$. These have the moduli-theoretic interpretations $\alpha(E, C) = (E, C_N)$ where C_N is the subgroup of C of order N, and $\beta(E, C) = (E/C_p, C/C_p)$ where C_p is the subgroup of C of order p. The coverings

$$\begin{array}{ccc} & Y_0(Np) & \\ \beta \swarrow & & \searrow \alpha \\ Y_0(N) & & Y_0(N), \end{array}$$

possibly branched, give rise to a correspondence (see §3.2) $T_p = \alpha \circ {}^t\beta$ on $Y_0(N) \times Y_0(N)$. If p does not divide N, then $T_p(\Gamma_0(N)\tau)$ is the divisor $\sum \Gamma_0(N)\gamma(\tau)$ where γ runs through the set

$$\left\{ \begin{pmatrix} 1 & 0 \\ 0 & p \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & p \end{pmatrix}, \ldots, \begin{pmatrix} 1 & p-1 \\ 0 & p \end{pmatrix}, \begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix} \right\}. \tag{7.3.1}$$

Identifying points of $Y_0(N)$ with pairs (E, C), we find that T_p has the following natural characterization

$$T_p(E, C) = \sum_D (E/D, (C + D)/D)$$

where D runs over cyclic subgroups of E of order p. If p divides N, then T_p (in this case frequently denoted U_p in the literature) has similar descriptions, except that we omit the last element from the above set of matrices and require that $D \not\subset C$. It follows directly from this description that $T_pT_q = T_qT_p$ for all primes p and q.

REMARK 7.3.1. In §3.4 the symbol T_p is used to denote the endomorphism of $\mathcal{S}_2(\Gamma_0(N))$ induced by the double coset $T(p) = \Gamma_0(N)\gamma\Gamma_0(N)$ where $\gamma = \begin{pmatrix} 1 & 0 \\ 0 & p \end{pmatrix}$. We have also explained in §3.2 how $T(p)$ gives rise to a correspondence on $Y_0(N) \times Y_0(N)$. There τ_γ is defined as $B \circ {}^tA$ where the maps A and B to $Y_0(N)$ are not from $Y_0(Np)$, but from $G\backslash\mathfrak{H}$ where G is conjugate to $\Gamma_0(Np)$ by γ. However, the correspondences τ_γ and T_p on $Y_0(N)$ are the same, as can be seen by composing A and B with the isomorphism between $X_0(Np)$ and $G\backslash\mathfrak{H}$ to obtain the degeneracy maps β and α, respectively.

More generally, to any positive integer n we can associate a modular correspondence T_n on $Y_0(N) \times Y_0(N)$ by the formula

$$T_n(E, C) = \sum_D (E/D, (C + D)/D) \tag{7.3.2}$$

where D runs over subgroups of E of order n satisfying $C \cap D = 0$. Then T_n coincides with the correspondence which arises from $T(n)$ by the construction of §3.2. In particular, having set $T_1 = 1$ and defined T_p for all primes p, the T_n are

characterized by the equations (see Proposition 3.3.1)

$$(7.3.3)\qquad \begin{array}{lcll} T_{p^r} & = & T_{p^{r-1}}T_p - pT_{p^{r-2}} & \text{if } r \geq 2 \text{ and } p \text{ is a prime not dividing } N; \\ T_{p^r} & = & T_p^r & \text{if } p \text{ is a prime dividing } N; \\ T_{mn} & = & T_m T_n & \text{if } (m,n) = 1. \end{array}$$

One can similarly define and describe modular correspondences on $Y_1(N) \times Y_1(N)$. In particular, suppose that p is a prime not dividing N and consider the curve $Y = \Gamma_1(N,p)\backslash\mathfrak{H}$, where $\Gamma_1(N,p)$ denotes $\Gamma_1(N) \cap \Gamma_0(p)$. The modular correspondence T_p is defined by $\alpha \circ {}^t\beta$ where α and β are the degeneracy maps from Y to $Y_1(N)$ defined, respectively, by $\tau \mapsto \tau$ and $\tau \mapsto p\tau$. The effect of T_p on a point $\Gamma_1(N)\tau$ of $Y_1(N)$ is given by the formal sum $\sum \Gamma_1(N)\gamma(\tau)$ where γ runs through the set in (7.3.1), except that the last matrix requires a slight modification. We now have the moduli-theoretic interpretation:

$$T_p(E,P) = \sum_D (E/D, P \bmod D)$$

where D runs over subgroups of E of order p. One can again define T_n for integers $n \geq 1$ with a moduli-theoretic interpretation analogous to the one in (7.3.2), and again these coincide with the correspondences which arise from $T(n)$ via the construction of §3.2. They satisfy the equations listed in (7.3.3) except that now $T_{p^r} = T_{p^{r-1}}T_p - \langle p\rangle pT_{p^{r-2}}$ for primes p not dividing N. Note also that the correspondence T_n commutes with the action of $(\mathbf{Z}/N\mathbf{Z})^\times$ and the natural projection $Y_1(N) \to Y_0(N)$.

REMARK 7.3.2. One can also give a simple moduli-theoretic interpretation of the involution of $Y_1(N)$ induced by the matrix w_N defined in §4. The pair (E,P) is sent to $(E/\langle P\rangle, P' \bmod \langle P\rangle)$ where P' is a point of $E[N]$ satisfying $\langle P, P'\rangle = e^{2\pi i/N}$ where $\langle\ ,\ \rangle$ is the Weil pairing on $E[N]$. Denoting the involution again by w_N we have the identities $w_N T_n w_N = {}^tT_n$ for all $n \geq 1$ and $w_N\langle d\rangle w_N = \langle d\rangle^{-1}$ for all integers d relatively prime to N. Similarly, for Q dividing N and satisfying $(Q, N/Q) = 1$, the involution w_Q of $Y_0(N)$ defined in §4 has the interpretation $(E,C) \mapsto (E/C[Q], (E[Q]+C)/C[Q])$.

8. Canonical models

The aim of this section is to explain how the interpretation of the modular curves as moduli spaces can be used to define canonical models for these curves. We will also discuss the Eichler-Shimura congruence relation in this context. The main reference will be [**DeRa**], but see also [**Shi1**], [**KaMa**] and [**MaWi**]. We assume some background in algebraic geometry, as can be found for example in [**Hart**].

By a model for a modular curve Y over a subring R of $\mathbf{C}$, we mean a pair $(\mathcal{Y}, \phi)$ where $\mathcal{Y}$ is a scheme over $\operatorname{Spec} R$ with one-dimensional fibers, and ϕ is an analytic isomorphism $Y \cong \mathcal{Y}(\mathbf{C})$.

EXAMPLE 8.0.1. Consider the scheme $\mathcal{Y} = \operatorname{Spec}(\mathbf{Z}[j])$ over $\operatorname{Spec}\mathbf{Z}$ and the isomorphism $\phi : Y_1(1) \to \mathcal{Y}(\mathbf{C})$ which sends $\mathrm{SL}_2(\mathbf{Z})\tau$ to the element of $\mathcal{Y}(\mathbf{C})$ defined by $j \mapsto j(\tau)$. The pair $(\mathcal{Y}, \phi)$ is a model for $Y_1(1)$ over $\mathbf{Z}$.

Shimura's theory of canonical models [**Shi1**, §6.7] provides a compatible system of models over number fields for modular curves associated to congruence

subgroups. In particular, he shows that $Y_0(N)$ and $Y_1(N)$ can be viewed as quasi-projective varieties defined over $\mathbf{Q}$ so that the projections, degeneracy maps and correspondences discussed above can be defined over $\mathbf{Q}$. We will adopt the point of view of Deligne-Rapoport [**DeRa**] in order to define models over rings of the form $\mathbf{Z}[1/N]$. These models are an important tool in the study of the arithmetic of modular curves and our discussion will barely scratch the surface. In addition to [**DeRa**], the reader can consult [**Igu1**], [**Igu2**], [**Igu3**], [**Drin**], [**KaMa**] and [**MaWi**].

8.1. Families of elliptic curves.

PRIMARY REFERENCES:
[**KaMa**, Chapter2], [**Gross**, §1] and [**Sil2**, Chapters III,IV].

Recall first that the points of $Y_1(N)$ correspond to pairs (E, P) where E is an elliptic curve over $\mathbf{C}$ and P is a point of E of order N. We will now rephrase the definition of a pair (E, P) so that it makes sense with $\mathbf{C}$ replaced by a scheme S over $\mathbf{Z}[1/N]$. By a family of elliptic curves over S, often simply called "an elliptic curve over S," we mean a smooth, proper group scheme over S whose geometric fibers are elliptic curves.

EXAMPLE 8.1.1. Let $S = \operatorname{Spec} \mathbf{Z}[1/11]$ and let $\mathcal{E}$ be the closed subscheme of $\mathbf{P}^2_S$ defined projectively by

$$Y^2Z + YZ^2 = X^3 - X^2Z - 10XZ^2 - 20Z^3.$$

Then $\mathcal{E}$ can be given the structure of an elliptic curve over S with zero section $S \to \mathcal{E}$ defined by "the point at ∞," $X \mapsto 0$, $Z \mapsto 0$.

EXAMPLE 8.1.2. Let $S = \operatorname{Spec}(\mathbf{Z}[j, j^{-1}(j-1728)^{-1}])$. Let $\mathcal{E}$ be the "generic" elliptic curve over S, i.e., the closed subscheme of $\mathbf{P}^2_S$ defined by

$$Y^2Z + XYZ = X^3 - 36(j-1728)^{-1}XZ^2 - (j-1728)^{-1}Z^3.$$

At a geometric point $\operatorname{Spec} k \to S$ defined by $j \mapsto j_0$ (where $j_0 \in k$ with $j_0 \neq 0, 1728$), the fiber of $\mathcal{E}$ has Weierstrass equation obtained by replacing j by j_0 in the equation above (see [**Sil1**, §III.1]).

EXAMPLE 8.1.3. The Tate curve [**Del4**, §8] over $S = \operatorname{Spec}(\mathbf{Z}((q)))$ is defined by $Y^2Z + XYZ = X^3 + a_4XZ^2 + a_6Z^3$ in $\mathbf{P}^2_S$ where

$$a_4 = -5\sum_{n\geq 1} n^3q^n/(1-q^n); \qquad a_6 = -\frac{1}{12}\sum_{n\geq 1}(7n^5+5n^3)q^n/(1-q^n).$$

8.2. Moduli problems.

PRIMARY REFERENCES:
[**DeRa**, Chapters III,IV], [**KaMa**, Chapters 3,4], [**Shi1**, Chapter 6] and [**MaWi**, §2.3].

Now define a contravariant functor $\mathcal{F}_1(N)$ from $\mathbf{Z}[1/N]$-schemes to sets as follows: For a scheme S over $\mathbf{Z}[1/N]$, $\mathcal{F}_1(N)(S)$ is the set of isomorphism classes of pairs $(\mathcal{E}, \mathcal{P})$ where $\mathcal{E}$ is an elliptic curve over S and $\mathcal{P}$ is an element of $\mathcal{E}(S)$ of exact order N. A section $\mathcal{P} : S \to \mathcal{E}$ is said to have exact order N if for all geometric points $s : \operatorname{Spec} k \to S$, $\mathcal{P} \circ s$ has order N in $\mathcal{E}(k)$. If $f : S \to T$ is a morphism of schemes, we define $\mathcal{F}_1(N)(f) : \mathcal{F}_1(N)(T) \to \mathcal{F}_1(N)(S)$ by "base-change"

$(\mathcal{E}, \mathcal{P}) \mapsto (\mathcal{E}_S, \mathcal{P}_S)$ where $\mathcal{E}_S$ and $\mathcal{P}_S$ are defined so the squares in the following diagram are cartesian:

$$\begin{array}{ccc} S & \to & T \\ \downarrow \mathcal{P}_S & & \downarrow \mathcal{P} \\ \mathcal{E}_S & \to & \mathcal{E} \\ \downarrow & & \downarrow \\ S & \to & T. \end{array}$$

It follows formally from standard properties of base-change that $(\mathcal{E}_S, \mathcal{P}_S)$ defines an element of $\mathcal{F}_1(N)(S)$ and that $\mathcal{F}_1(N)$ is a functor.

In this language, the substantive statement is the following:

THEOREM 8.2.1. *If $N > 3$, then there is a scheme $\mathcal{Y}_1(N)$ which represents the functor $\mathcal{F}_1(N)$. Moreover $\mathcal{Y}_1(N)$ is smooth of relative dimension one over $\mathbf{Z}[1/N]$ with irreducible geometric fibers.*

The proof is essentially due to Igusa [**Igu1**], but the statement in this form is most easily deduced from (2.7.3), (3.7.1) and (4.7.1) of Katz-Mazur [**KaMa**]. See also [**DeRa**] for a sketch of Igusa's method and statements similar to the one above. The meaning of "$\mathcal{Y}_1(N)$ represents $\mathcal{F}_1(N)$" is that for any scheme S over $\mathbf{Z}[1/N]$, there is a bijection, functorial in S, between the set of maps $S \to \mathcal{Y}_1(N)$ and the set of isomorphism classes of pairs $(\mathcal{E}, \mathcal{P})$ over S. It follows formally that the scheme $\mathcal{Y}_1(N)$ with this property is unique up to canonical isomorphism. Note also that corresponding to the identity map in the case $S = \mathcal{Y}_1(N)$ is a pair $(\mathcal{E}_{\text{univ}}, \mathcal{P}_{\text{univ}})$ which can be considered the "universal elliptic curve with a point of order N." Indeed any pair $(\mathcal{E}, \mathcal{P})$ over a $\mathbf{Z}[1/N]$-scheme T is obtained from $(\mathcal{E}_{\text{univ}}, \mathcal{P}_{\text{univ}})$ by base-change for a unique morphism $T \to \mathcal{Y}_1(N)$. Considering the case $S = \mathbf{C}$, we find a natural bijection ϕ between $Y_1(N)$ and $\mathcal{Y}_1(N)(\mathbf{C})$. This bijection is an analytic isomorphism, so $(\mathcal{Y}_1(N), \phi)$ is indeed a model for $Y_1(N)$.

VARIANT 8.2.2. It will be convenient at times to use models defined using a different set of conventions. Giving a section $\mathcal{P} : S \to \mathcal{E}$ of exact order N amounts to giving a closed immersion $(\mathbf{Z}/N\mathbf{Z})_S \hookrightarrow \mathcal{E}$ of group schemes over S, where $(\mathbf{Z}/N\mathbf{Z})_S$ denotes the constant group scheme $\mathbf{Z}/N\mathbf{Z}$ over S ([**KaMa**, (1.4.4)]). Some authors, Gross [**Gross**] and Katz [**Katz1**], [**Katz2**] for example, use a model for $Y_1(N)$ which instead parametrizes pairs $(\mathcal{E}, i)$ where i is a closed immersion $(\boldsymbol{\mu}_N)_S \hookrightarrow \mathcal{E}$. The resulting moduli problem is represented over $\mathbf{Z}$ by a smooth affine scheme we denote $\mathcal{Y}_\mu(N)$. We thus obtain a model for $Y_1(N)$ over $\mathbf{Z}$, $(\mathcal{Y}_\mu(N), \phi_\mu)$, with ϕ_μ defined by $\tau \mapsto (E/\Lambda_\tau, i_\tau)$ for $\tau \in \mathfrak{H}$. Here i_τ denotes the embedding defined by $i_\tau(\zeta_N) = 1/N \bmod \Lambda_\tau$ where $\zeta_N = e^{2\pi i/N}$. Note that the models are isomorphic when tensored with $\mathbf{Z}[1/N, \zeta_N]$. We caution that there is an isomorphism of schemes $\mathcal{Y}_1(N) \cong \mathcal{Y}_\mu(N)_{\mathbf{Z}[1/N]}$ over $\mathbf{Z}[1/N]$, but such an isomorphism does not respect the maps ϕ and ϕ_μ and thus is not an isomorphism of models over $\mathbf{Z}[1/N]$.

Consider now the situation for $Y_0(N)$, which is complicated slightly by torsion in $\Gamma_0(N)$. Let us continue to assume that $N > 3$ and return later to the case $N \leq 3$. We can define a functor $\mathcal{F}_0(N)$ on $\mathbf{Z}[1/N]$-schemes where $\mathcal{F}_0(N)(S)$ is the set of isomorphism classes of pairs $(\mathcal{E}, \mathcal{C})$ where $\mathcal{E}$ is an elliptic curve over S and $\mathcal{C}$ is a finite flat subgroup scheme of $\mathcal{E}$ whose geometric fibers are cyclic groups of order N. It is tempting to ask for a model for $Y_0(N)$ which represents $\mathcal{F}_0(N)$. The fact that a pair $(\mathcal{E}, \mathcal{C})$ has non-trivial automorphisms, multiplication by -1 for example, makes the issue of representability a subtler one. We can nonetheless

proceed as follows. For an integer d relatively prime to N, the pair $(\mathcal{E}_{\text{univ}}, d\mathcal{P}_{\text{univ}})$ over $\mathcal{Y}_1(N)$ defines a morphism $\mathcal{Y}_1(N) \to \mathcal{Y}_1(N)$ which we call $\langle d \rangle$. Then $d \mapsto \langle d \rangle$ defines a homomorphism $G \to \text{Aut}(\mathcal{Y}_1(N))$ where $G = (\mathbf{Z}/N\mathbf{Z})^\times$, or in other words, an action of G on $\mathcal{Y}_1(N)$. Equivalently, we can view $\langle d \rangle$ as the natural transformation $\mathcal{F}_1(N) \to \mathcal{F}_1(N)$ defined by $(\mathcal{E}, \mathcal{P}) \mapsto (\mathcal{E}, d\mathcal{P})$. Note that $\langle d \rangle$ is a model for the automorphism of $Y_1(N)$, which we also denoted $\langle d \rangle$, in the sense that $\langle d \rangle \circ \phi(z) = \langle d \rangle(z)$ for all $z \in Y_1(N)$. We can then consider the quotient scheme $\mathcal{Y}_0(N) = G \backslash \mathcal{Y}_1(N)$. We mention some of its properties, proved in [**DeRa**, Ch. VI] and [**KaMa**, Ch. 8]. It is a smooth scheme over $\mathbf{Z}[1/N]$, and the natural projection $\mathcal{Y}_1(N) \to \mathcal{Y}_0(N)$ is finite and flat, but not necessarily etale. There are also maps $\phi_S : \mathcal{F}_0(N)(S) \to \mathcal{Y}_0(N)(S)$, functorial in $\mathbf{Z}[1/N]$-schemes S, and bijective if $S = \text{Spec}\, k$ for an algebraically closed field k. Applying this for $k = \mathbf{C}$, we find that $\mathcal{Y}_0(N)$ is a model for $Y_0(N)$. As ϕ_S is not necessarily a bijection, $\mathcal{Y}_0(N)$ need not represent $\mathcal{F}_0(N)$. In any case $\mathcal{Y}_0(N)$ has an interpretation as a "coarse moduli scheme" ([**DeRa**, §I.8], [**KaMa**, §8.1]), but we will not define the term here. We mention only that for a field k, $\mathcal{Y}_0(N)(k)$ can be identified with the set of equivalence classes of pairs $(\mathcal{E}, \mathcal{C})$ over k, where two pairs are deemed equivalent if they are isomorphic over the algebraic closure of k.

In the case $N \leq 3$, a similar construction yields a model over $\mathbf{Z}[1/N]$ for $Y_0(N) = Y_1(N)$. Again the scheme $\mathcal{Y}_0(N) = \mathcal{Y}_1(N)$ can be interpreted as a coarse moduli scheme and it has the properties listed above for $\mathcal{Y}_0(N)$ in the case $N > 3$. For $N = 1$, we recover the model in Example 8.0.1. We find also that the map $\mathcal{F}_0(1)(k) \to \mathcal{Y}_0(1)(k)$ for a field k is described by sending an elliptic curve to its j-invariant. Note that this is not a bijection in the case $k = \mathbf{Q}$; quadratic twists of an elliptic curve have the same j-invariant, but are not necessarily isomorphic over $\mathbf{Q}$.

8.3. Models for modular correspondences.

PRIMARY REFERENCES:

[**DeRa**, §V.1], [**MaWi**, §2.5] and [**KaMa**, Chapters 5,6].

We now turn to the problem of defining models for the degeneracy maps and modular correspondences considered in §7.3. For this we will need a model over $\mathbf{Z}[1/N]$ for the curve $Y = \Gamma \backslash \mathfrak{H}$, where $\Gamma = \Gamma_1(N, p) = \Gamma_1(N) \cap \Gamma_0(p)$. We are now working in a situation of bad reduction at the prime p, in the sense that the "best" model turns out to be regular if $N > 3$, but the fiber over p is not smooth. The moduli-theoretic construction and analysis of this model is due to Deligne-Rapoport [**DeRa**], but for a more general construction of such models using Drinfeld's notion of "elliptic modules," see [**KaMa**].

First note that we can interpret Y as the space parametrizing triples (E, P, C) where E is an elliptic curve over $\mathbf{C}$, P is a point of order N and C is a cyclic subgroup of order p. To be more explicit, and consistent with our description of $Y_1(Np)$, we associate to $\tau \in \mathfrak{H}$ the triple (E, P, C) where $E = \mathbf{C}/\Lambda_\tau$, $P = dN^{-1} \bmod \Lambda_\tau$ with $dp \equiv 1 \bmod N$, and C is generated by p^{-1}.

Mimicking the above definition for $\mathcal{F}_1(N)$, we define a corresponding functor $\mathcal{F}$ on $\mathbf{Z}[1/N]$-schemes which assigns to such a scheme S the set of isomorphism classes of triples $(\mathcal{E}, \mathcal{P}, \mathcal{C})$ over S, where $\mathcal{E}$ is an elliptic curve over S, $\mathcal{P}$ is a point of order N and $\mathcal{C}$ is a finite flat subgroup scheme of $\mathcal{E}$ with geometric fibers of rank p. Note that if $S = \overline{\mathbf{F}}_p$, then the group scheme $\mathcal{C}$ cannot adequately be described as "the cyclic group of order p." Indeed $\mathcal{C}$ may be isomorphic to $\boldsymbol{\mu}_p$ or $\mathbf{Z}/p\mathbf{Z}$ if $\mathcal{E}$ is

ordinary, and it must be isomorphic to $\boldsymbol{\alpha}_p$ if $\mathcal{E}_x$ is supersingular. (See [**Sil1**, §V.3] for the definitions of ordinary and supersingular and [**Shatz**, §2] for definitions of $\boldsymbol{\mu}_p$ and $\boldsymbol{\alpha}_p$.)

The following is a consequence of the work of Deligne and Rapoport. (See [**DeRa**, V.1.20] and [**KaMa**, §5.1].)

THEOREM 8.3.1. *If $N > 3$, then $\mathcal{F}$ is representable by a scheme $\mathcal{Y}$ which provides a model for Y over $\mathbf{Z}[1/N]$. Moreover $\mathcal{Y}$ is regular and $\mathcal{Y}_{\mathbf{Z}[1/Np]}$ is smooth over $\mathbf{Z}[1/Np]$ with irreducible geometric fibers.*

We will return in §8.4 to consider the behavior of $\mathcal{Y}$ at p if p does not divide N, for it is closely related to the Eichler-Shimura relation and plays an important role in the work of Ribet [**Rib4**]. First let us give another description of the functor $\mathcal{F}$ and define models for the degeneracy maps and modular correspondences.

By an isogeny $\pi : (\mathcal{E}, \mathcal{P}) \to (\mathcal{E}', \mathcal{P}')$, we mean a finite flat homomorphism $\pi : \mathcal{E} \to \mathcal{E}'$ such that $\pi \circ \mathcal{P} = \mathcal{P}'$. If $\mathcal{E}$ is an elliptic curve over a scheme S and $\mathcal{C}$ is a finite flat subgroup scheme of $\mathcal{E}$, then there is an elliptic curve $\mathcal{E}' = \mathcal{E}/\mathcal{C}$ over S and an isogeny $\pi : \mathcal{E} \to \mathcal{E}'$ with kernel $\mathcal{C}$. Moreover $\mathcal{F}$ is naturally isomorphic to the functor which assigns to a $\mathbf{Z}[1/N]$-scheme S the set of isomorphism classes of isogenies $(\mathcal{E}, \mathcal{P}) \to (\mathcal{E}', \mathcal{P}')$ over S of degree p.

The definition of models $\alpha', \beta' : \mathcal{Y} \to \mathcal{Y}_1(N)$ for the degeneracy maps $\alpha, \beta : Y \to Y_1(N)$ is now as formal as that of $\langle d \rangle$ in §8.2. We define a model α' for α using the natural transformation $\mathcal{F} \to \mathcal{F}_1(N)$ defined by sending an isogeny $(\mathcal{E}, \mathcal{P}) \to (\mathcal{E}', \mathcal{P}')$ to the pair $(\mathcal{E}, p\mathcal{P})$. (Note that it is necessary to use $p\mathcal{P}$ in order to have $\alpha' \circ \phi(z) = \alpha(z)$ for $z \in Y$, where ϕ is the isomorphism $Y \to \mathcal{Y}(\mathbf{C})$ corresponding to our parametrization by Y of triples (E, P, C) over $\mathbf{C}$.) We define β' by sending the isogeny to its target $(\mathcal{E}', \mathcal{P}')$. We can now describe a "model" over $\mathbf{Z}[1/N]$ for the correspondence T_p as the mapping

$$\mathcal{T} = (\beta', \alpha') : \mathcal{Y} \to \mathcal{Y}_1(N) \times \mathcal{Y}_1(N).$$

To see how this gives rise to T_p, consider $\mathcal{T}$ as a morphism of schemes over $\mathcal{Y}_1(N)$ via the projection $\pi_1 : \mathcal{Y}_1(N) \times \mathcal{Y}_1(N) \to \mathcal{Y}_1(N)$. Identifying $Y_1(N)$ with $\mathcal{Y}_1(N)(\mathbf{C})$, we find that the geometric fiber

$$\mathcal{T}_x : \mathcal{Y}_x \to \operatorname{Spec} \mathbf{C} \times \mathcal{Y}_1(N) \cong \mathcal{Y}_1(N)$$

of $\mathcal{T}$ over a point $x \in Y_1(N)$ defines the divisor $T_p(x)$. Indeed on points, $\mathcal{T}_x$ is simply the restriction to $\beta^{-1}(x)$ of $\alpha : Y \to Y_1(N)$.

REMARK 8.3.2. If X and Y are varieties over a field k, then a correspondence on $X \times Y$, or from X to Y, is usually defined as a divisor on $X \times Y$ (see [**Shi1**, §7.2]). We have defined $\mathcal{T}$ as a morphism rather than as a divisor in order to avoid complications which arise when considering relative divisors over more general base schemes than $S = \operatorname{Spec} k$.

8.4. Bad reduction.

PRIMARY REFERENCES:
[**DeRa**, §V.1] and [**KaMa**, Chapter 13].

We now return to the analysis by Deligne and Rapoport of the "bad reduction" $\mathcal{Y}_{\mathbf{F}_p} = \mathcal{Y} \times \mathbf{F}_p$ of the model in Theorem 8.3.1. We will define two natural maps $\mathcal{Y}_1(N)_{\mathbf{F}_p} \to \mathcal{Y}_{\mathbf{F}_p}$. Consider first the elliptic curve $\mathcal{E}_0 = (\mathcal{E}_{\text{univ}})_{\mathbf{F}_p}$ over $\mathcal{Y}_1(N)_{\mathbf{F}_p}$ and

the commutative diagram

$$\begin{array}{ccc} \mathcal{E}_0 & \to & \mathcal{E}_0 \\ \downarrow & & \downarrow \\ \mathcal{Y}_1(N)_{\mathbf{F}_p} & \xrightarrow{\Phi} & \mathcal{Y}_1(N)_{\mathbf{F}_p} \end{array}$$

where the horizontal maps are the absolute Frobenius endomorphisms. The diagram gives rise to Frob : $\mathcal{E}_0 \to \mathcal{E}_0^{(p)}$ where $\mathcal{E}_0^{(p)}$ is the elliptic curve over $\mathcal{Y}_1(N)_{\mathbf{F}_p}$ defined as the base-change of $\mathcal{E}_0$ relative to Φ. The map Frob is called the relative Frobenius of $\mathcal{E}_0$; it is an isogeny of degree p of elliptic curves over $\mathcal{Y}_1(N)_{\mathbf{F}_p}$. Writing $\mathcal{P}_0$ for $(\mathcal{P}_{\text{univ}})_{\mathbf{F}_p}$ and $\mathcal{P}_0^{(p)}$ for Frob $\circ\, \mathcal{P}_0$, we get an isogeny

$$\text{Frob} : (\mathcal{E}_0, \mathcal{P}_0) \to (\mathcal{E}_0^{(p)}, \mathcal{P}_0^{(p)})$$

which defines an element of $\mathcal{F}(\mathcal{Y}_1(N)_{\mathbf{F}_p})$ and thus a map

$$i_F : \mathcal{Y}_1(N)_{\mathbf{F}_p} \to \mathcal{Y}_{\mathbf{F}_p}.$$

For a more concrete description, recall that a point $x : \operatorname{Spec} \overline{\mathbf{F}}_p \to \mathcal{Y}_1(N)$ corresponds to an elliptic curve E_x over $k = \overline{\mathbf{F}}_p$ together with a point P_x of order N. Its image $i_F \circ x$ is the point $\operatorname{Spec} k \to \mathcal{Y}$ which corresponds to the triple (E_x, P_x, C) where C is the kernel of the Frobenius $E_x \to E_x^{(p)}$. Note that $E_x^{(p)}$ is the elliptic curve obtained by composing x with the absolute Frobenius automorphism of $\operatorname{Spec} k$, or equivalently, by applying the Frobenius of k to the coefficients of an equation defining E_x over k.

To define a second natural map $\mathcal{Y}_1(N)_{\mathbf{F}_p} \to \mathcal{Y}_{\mathbf{F}_p}$, we use the dual isogeny Ver : $\mathcal{E}_0^{(p)} \to \mathcal{E}_0$. This isogeny, often called the Verschiebung, is characterized by the fact that Ver $\circ$ Frob is multiplication by p on $\mathcal{E}_0$. Thus

$$\text{Ver} : (\mathcal{E}_0^{(p)}, d\mathcal{P}_0^{(p)}) \to (\mathcal{E}_0, \mathcal{P}_0)$$

defines a map

$$i_V : \mathcal{Y}_1(N)_{\mathbf{F}_p} \to \mathcal{Y}_{\mathbf{F}_p},$$

where $dp \equiv 1 \bmod N$. For a point x as above, the image $i_V \circ x$ corresponds $(E_x^{(p)}, dP_x^{(p)}, D)$ where D is the kernel of the isogeny $E_x^{(p)} \to E_x$ dual to the Frobenius.

From their effect on pairs $(\mathcal{E}, \mathcal{P})$, we can read off the composites of $\alpha'_{\mathbf{F}_p}$ and $\beta'_{\mathbf{F}_p}$ with i_F and i_V. We find that

$$\begin{array}{rcl} \alpha'_{\mathbf{F}_p} \circ i_F & = & \langle p \rangle_{\mathbf{F}_p} \\ \alpha'_{\mathbf{F}_p} \circ i_V & = & \Phi \\ \beta'_{\mathbf{F}_p} \circ i_F & = & \Phi \\ \beta'_{\mathbf{F}_p} \circ i_V & = & \text{id}. \end{array} \tag{8.4.1}$$

In particular, it follows that i_F and i_V are closed immersions. Using these immersions, we can give a complete description of $\mathcal{Y}_{\mathbf{F}_p}$ in terms of $\mathcal{Y}_1(N)_{\mathbf{F}_p}$. First consider the non-empty finite set of points on $\mathcal{Y}_1(N)_{\mathbf{F}_p}$ over which the geometric fiber of $\mathcal{E}_0$ is a supersingular elliptic curve. These form a closed subscheme of $\mathcal{Y}_1(N)_{\mathbf{F}_p}$ whose complement we denote $\mathcal{Y}_1(N)^{\text{ord}}$. Define $\mathcal{Y}^{\text{ord}}$ similarly and consider the restrictions $i_F^{\text{ord}}, i_V^{\text{ord}} : \mathcal{Y}_1(N)^{\text{ord}} \to \mathcal{Y}^{\text{ord}}$. A point $\operatorname{Spec} k \to \mathcal{Y}^{\text{ord}}$ corresponding to a triple (E, P, C) over k is in the image of i_F if and only if C is connected, and in the image of i_V if and only if C is etale. It follows that $i_F^{\text{ord}} \coprod i_V^{\text{ord}}$ is an

isomorphism. On the other hand, the finite set of points corresponding to supersingular elliptic curves lie in the intersection of the images of i_F and i_V, and this forms the singular locus of $\mathcal{Y}_{\mathbf{F}_p}$. We conclude that $\mathcal{Y}_{\mathbf{F}_p}$ consists of two irreducible components, each isomorphic to $\mathcal{Y}_1(N)_{\mathbf{F}_p}$, one via i_F, the other via i_V. A more careful analysis shows that the components cross transversally at the supersingular points, identifying $i_F \circ x$ with $i_V \circ \Phi \circ x$ for the points $x : \operatorname{Spec} k \to \mathcal{Y}_1(N)$ with E_x supersingular.

REMARK 8.4.1. We may also define models for the maps defined in Remark 7.3.2 by the matrices w_N of §4 (see [**MaWi**, §2.5]). For a positive integer N and an elliptic curve $\mathcal{E}$ over S we can regard the Weil pairing as a morphism $\mathcal{E}[N] \times_S \mathcal{E}[N] \to \boldsymbol{\mu}_{N,S}$. If S is a $\mathbf{Z}[1/N]$-scheme and $\mathcal{P}$ is a point in $\mathcal{E}(S)$ of exact order N, then pairing with $\mathcal{P}$ defines a surjective morphism of group schemes $\mathcal{E}[N] \to \boldsymbol{\mu}_{N,S}$ whose kernel is $\mathcal{C} = \langle \mathcal{P} \rangle$, the subgroup scheme generated by $\mathcal{P}$. Now let $S = \operatorname{Spec} \mathbf{Z}[1/N, e^{2\pi i/N}]$. Then $e^{2\pi i/N}$ defines a point of $\boldsymbol{\mu}_N(S)$ and thus gives rise to a point $\mathcal{P}'$ of $(\mathcal{E}[N]/\mathcal{C})(S) \subset \mathcal{E}'(S)$ of exact order N where $\mathcal{E}'$ is the elliptic curve $\mathcal{E}/\mathcal{C}$. We can then naturally define a model for w_N on $\mathcal{Y}_1(N)_S$ by $(\mathcal{E}, \mathcal{P}) \mapsto (\mathcal{E}', \mathcal{P}')$.

For a prime p not dividing N, we can define a model w for w_p on $\mathcal{Y}$ by sending an isogeny $(\mathcal{E}, \mathcal{P}) \to (\mathcal{E}', \mathcal{P}')$ to its dual isogeny $(\mathcal{E}', \mathcal{P}') \to (\mathcal{E}, p\mathcal{P})$. Note the relation $w^2 = \langle p \rangle$ on $\mathcal{Y}$ and the relations $w_{\mathbf{F}_p} i_F = \langle p \rangle_{\mathbf{F}_p} i_V$ and $w_{\mathbf{F}_p} i_V = i_F$ in characteristic p. In particular, w interchanges the two irreducible components of $\mathcal{Y}_{\mathbf{F}_p}$.

Similarly, for suitable divisors Q of N we can define models for the involutions w_Q on the coarse moduli schemes $\mathcal{Y}_0(N)$. Furthermore, we can define a model for w_p on $\mathcal{Y}_0'(Np)$ which interchanges i_F and i_V.

8.5. The Eichler-Shimura relation.

PRIMARY REFERENCES:
[**Shi1**, Chapter 7], [**Del1**, §4] and [**DeRa**, §VI.6]

Note that our computation of the four composites in (8.4.1) describes the composite of the normalization $i = i_F \coprod i_V$ with the modular correspondence $\mathcal{T}_{\mathbf{F}_p}$ in characteristic p:

$$\begin{array}{c} \mathcal{Y}_1(N)_{\mathbf{F}_p} \coprod \mathcal{Y}_1(N)_{\mathbf{F}_p} \\ \downarrow \\ \mathcal{Y}_{\mathbf{F}_p} \\ \downarrow \\ \mathcal{Y}_1(N)_{\mathbf{F}_p} \times \mathcal{Y}_1(N)_{\mathbf{F}_p}. \end{array}$$

We have

$$\mathcal{T}_{\mathbf{F}_p} \circ i = (\Phi, \langle p \rangle_{\mathbf{F}_p}) \coprod (\mathrm{id}, \Phi). \tag{8.5.1}$$

This formula can be viewed as a form of the Eichler-Shimura congruence relation [**Eich**], [**Shi1**, Theorem 7.9] (see also [**Del1**, §4]), which essentially says that the correspondence T_p in characteristic p is generically the sum of the correspondences defined by the maps $(\Phi, \langle p \rangle_{\mathbf{F}_p})$ and (id, Φ). For a precise statement, again consider $\mathcal{T}$ as a morphism of schemes over $\mathcal{Y}_1(N)$ via π_1. Taking fibers over an ordinary point $x : \operatorname{Spec} \overline{\mathbf{F}}_p \to \mathcal{Y}_1(N)_{\mathbf{F}_p} \hookrightarrow \mathcal{Y}_1(N)$, we have

$$\begin{array}{ccc} [\mathcal{Y}_1(N) \coprod \mathcal{Y}_1(N)]_x & \xrightarrow{\sim} & \mathcal{Y}_x \\ & \xrightarrow{\mathcal{T}_x} & \operatorname{Spec} k \times \mathcal{Y}_1(N). \end{array}$$

The composite agrees with the one obtained from the fiber over x of the morphism $(\Phi, \langle p \rangle_{\mathbf{F}_p}) \coprod (\mathrm{id}, \Phi)$. Note that the divisor image of the resulting morphism is simply

$$\Phi_*(x) + \langle p \rangle_{\mathbf{F}_p,*} \Phi^*(x). \tag{8.5.2}$$

The situation for $\mathcal{Y}_0(N)$ for $N \geq 1$ is much the same, except that instead of $\mathcal{Y}$ we use a model $\mathcal{Y}_0'(Np)$ for $Y_0(Np)$ over $\mathbf{Z}[1/N]$ defined as a coarse moduli scheme. Again $\mathcal{Y}_0'(Np)_{\mathbf{Z}[1/Np]} = \mathcal{Y}_0(Np)$ is smooth over $\mathbf{Z}[1/Np]$. If p does not divide N, then $\mathcal{Y}_0'(Np)$ may not be regular, but $\mathcal{Y}_0'(Np)_{\mathbf{F}_p}$ can still be described as the union of two copies of $\mathcal{Y}_0(N)$ crossing transversally at supersingular points. The Eichler-Shimura relation takes the same form as in (8.5.1) or (8.5.2), except that $\langle p \rangle$ is replaced by the identity. For $N = 1$, the image of $\mathcal{T}_{\mathbf{F}_p}$ in

$$\mathcal{Y}_0(1)_{\mathbf{F}_p} \times \mathcal{Y}_0(1)_{\mathbf{F}_p} \cong \mathrm{Spec}\,(\mathbf{F}_p[j_1, j_2])$$

is defined by

$$(j_1 - j_2^p)(j_1^p - j_2),$$

a form of the Eichler-Shimura relation already known to Kronecker.

9. Compactification

We will now explain how to adjoin cusps to compactify the modular curve $\Gamma\backslash\mathfrak{H}$ and obtain a Riemann surface [**Shi1**, Chapter 1]. Following Deligne and Rapoport [**DeRa**] we give the moduli-theoretic interpretation for this compactification in the case $\Gamma = \Gamma_0(N)$ or $\Gamma_1(N)$ and discuss the properties of the resulting canonical models.

9.1. The cusps.

PRIMARY REFERENCES:
[**Shi1**, §1.3–1.6] and [**Miy2**, §1.7,1.8,4.2].

Let $\mathfrak{H}^* = \mathfrak{H} \cup \mathbf{Q} \cup \{\infty\}$ and let Γ be a congruence subgroup of $\mathrm{SL}_2(\mathbf{Z})$. Using the natural action of $\mathrm{GL}_2(\mathbf{Q})$ on $\mathbf{P}^1(\mathbf{Q}) = \mathbf{Q} \cup \{\infty\}$ defined by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \frac{m}{n} = \frac{am + bn}{cm + dn},$$

we extend the action of Γ on $\mathfrak{H}$ to one on $\mathfrak{H}^*$. We now consider the quotient $\Gamma\backslash\mathfrak{H}^*$. We write $X_0(N)$ for $\Gamma_0(N)\backslash\mathfrak{H}^*$ and $X_1(N)$ for $\Gamma_1(N)\backslash\mathfrak{H}^*$. Before defining a topology on $\mathfrak{H}^*$ and making the quotient a Riemann surface, note that $\mathrm{SL}_2(\mathbf{Z})$ acts transitively on $\mathbf{P}^1(\mathbf{Q})$ and in general $\Gamma\backslash\mathbf{P}^1(\mathbf{Q})$ is finite. The elements of this finite set, which is the complement of $\Gamma\backslash\mathfrak{H}$ in $\Gamma\backslash\mathfrak{H}^*$, are called the *cusps* of $\Gamma\backslash\mathfrak{H}^*$.

EXAMPLE 9.1.1. There is a unique cusp $\mathrm{SL}_2(\mathbf{Z}) \cdot \infty = \mathbf{Q} \cup \{\infty\}$ on $X_0(1)$.

EXAMPLE 9.1.2. Let B denote the subgroup $\{\pm\begin{pmatrix} x & y \\ 0 & x^{-1} \end{pmatrix}\}$ of $\mathrm{PSL}_2(\mathbf{Z}/N\mathbf{Z})$ and let U denote the subgroup $\{\pm\begin{pmatrix} 1 & y \\ 0 & 1 \end{pmatrix}\}$. Then the set of cusps of $X_0(N)$ is in bijection with the double coset space

$$B\backslash\mathrm{PSL}_2(\mathbf{Z}/N\mathbf{Z})/U.$$

The bijection is defined by

$$\Gamma_0(N) \cdot \frac{a}{c} = \Gamma_0(N) \cdot \gamma(\infty) \mapsto B\overline{\gamma}U$$

where $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is in $\mathrm{SL}_2(\mathbf{Z})$ and $\overline{\gamma}$ is the image of γ in $\mathrm{PSL}_2(\mathbf{Z}/N\mathbf{Z})$. In particular, $X_0(p)$ has two cusps $\Gamma_0(p) \cdot 0$ and $\Gamma_0(p) \cdot \infty$.

EXAMPLE 9.1.3. Similarly there is a bijection between the set of cusps of $X_1(N)$ and $U\backslash \mathrm{PSL}_2(\mathbf{Z}/N\mathbf{Z})/U$. Note also that this is in bijection with the set

$$\{(c,d) \mid c \in \mathbf{Z}/N\mathbf{Z}, d \in (\mathbf{Z}/(c,N)\mathbf{Z})^\times\}/\{\pm 1\}. \tag{9.1.1}$$

Explicitly, the cusp $\Gamma_1(N)\cdot\frac{a}{c}$ corresponds to the pair (c,d) where d is chosen so that $ad \equiv 1 \bmod (c,N)$. The determination of the number of cusps on $X_1(N)$ is then straightforward (see [**Miy2**, Theorem 4.2.9]); one finds that $X_1(1)$ has one cusp, $X_1(2)$ has 2, $X_1(4)$ has 3 and that if $N \neq 1$, 2 or 4, then the number of cusps on $X_1(N)$ is

$$\frac{1}{2}\sum \phi(d)\phi(N/d) = \frac{N}{2}\prod_{p|N}[1-p^{-2}+v_p(N)(1-p^{-1})^2],$$

where the first sum is over positive divisors d of N.

Let us now define a topology on $\mathfrak{H}^*$. As a base of open neighborhoods of $\mathfrak{H}^*$ we use the open subsets of $\mathfrak{H}$ (with its usual topology) and the sets

$$\gamma(\{x+iy \mid y > C\} \cup \{\infty\})$$

for $\gamma \in \mathrm{SL}_2(\mathbf{Z})$ and $0 \leq C \in \mathbf{R}$. Thus $\mathfrak{H}$ is an open subspace of $\mathfrak{H}^*$ and $\mathbf{Q} \cup \{\infty\}$ is discrete. The quotient space $\Gamma\backslash\mathfrak{H}^*$ is compact and connected. It is the union of the open subspace $\Gamma\backslash\mathfrak{H}$ and the finite set of cusps. We have already defined a complex structure on $\Gamma\backslash\mathfrak{H}$ and we shall now define a complex structure in a neighborhood of each cusp. For $\gamma \in \mathrm{SL}_2(\mathbf{Z})$, let h be the positive integer such that $\pm\begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}$ generates the stabilizer of ∞ in the image of $\gamma^{-1}\Gamma\gamma$ in $\mathrm{PSL}_2(\mathbf{Z})$. The map $\gamma(\tau) \mapsto e^{2\pi i\tau/h}$, $\gamma(\infty) \mapsto 0$ defines a homeomorphism from a neighborhood of $\Gamma\gamma(\infty)$ in $\Gamma\backslash\mathfrak{H}^*$ to the unit disk in $\mathbf{C}$. The homeomorphism depends only on the cusp and not on the choice of γ and the resulting complex structure is compatible with the one we have already defined on $\Gamma\backslash\mathfrak{H}$. We may now regard $\Gamma\backslash\mathfrak{H}^*$ as a Riemann surface.

EXAMPLE 9.1.4. Recall that $Y_0(1)$ is biholomorphic to $\mathbf{C}$ via $\tau \mapsto j(E_\tau)$. Mapping $\mathrm{SL}_2(\mathbf{Z})\cdot\infty$ to ∞ extends this to an isomorphism of $X_0(1)$ with the Riemann sphere $\mathbf{P}^1(\mathbf{C})$.

We find also that if Γ and Γ' are congruence subgroups of $\mathrm{SL}_2(\mathbf{Z})$ and γ is an element of $\mathrm{GL}_2^+(\mathbf{Q})$ such that $\Gamma \subset \gamma^{-1}\Gamma'\gamma$, then $\tau \mapsto \gamma(\tau)$ for τ in $\mathfrak{H}^*$ induces a holomorphic map $\Gamma\backslash\mathfrak{H}^* \to \Gamma'\backslash\mathfrak{H}^*$. The only possible ramification occurs over the cusps and elliptic points of $\Gamma'\backslash\mathfrak{H}^*$.

Now let $\Gamma' = \mathrm{SL}_2(\mathbf{Z})$ and let γ be the identity. The resulting map to $\mathrm{SL}_2(\mathbf{Z})\backslash\mathfrak{H}^*$ may be ramified only over the points $\mathrm{SL}_2(\mathbf{Z})i$, $\mathrm{SL}_2(\mathbf{Z})e^{\pi i/3}$ and $\mathrm{SL}_2(\mathbf{Z})\cdot\infty$, which correspond via j (see Example 9.1.4) to the points 1728, 0 and ∞ in $\mathbf{P}^1(\mathbf{C})$. Applying the Hurwitz formula to this covering of the Riemann sphere yields the formula [**Shi1**, Prop. 1.40]

$$g = 1 + \frac{\mu}{12} - \frac{\nu_2}{4} - \frac{\nu_3}{3} - \frac{\nu_\infty}{2} \tag{9.1.2}$$

for the genus g of $\Gamma\backslash\mathfrak{H}^*$. Here μ is the index of the image of Γ in $\mathrm{PSL}_2(\mathbf{Z})$, ν_∞ is the number of cusps and ν_2 (respectively ν_3) is the number of elliptic points over $j = 1728$ (respectively $j = 0$).

EXAMPLE 9.1.5. As $X_0(11)$ has no elliptic points and two cusps, and $\Gamma_0(11)$ has index 12 in $\mathrm{SL}_2(\mathbf{Z})$, we see that $X_0(11)$ has genus one. We find also that $X_1(11)$ has genus one and the cover $X_1(11) \to X_0(11)$ is cyclic of degree 5, with Galois group $(\mathbf{Z}/11\mathbf{Z})^\times/\{\pm 1\}$.

EXAMPLE 9.1.6. By the formula of Example 9.1.3 together with the fact that $X_1(N)$ has no elliptic points for $N > 3$, we see that the genus of $X_1(N)$ for $N > 4$ is given by

$$g = 1 + \frac{N^2}{24}\prod_{p|N}(1-p^{-2}) - \frac{N}{4}\prod_{p|N}[1-p^{-2}+v_p(N)(1-p^{-1})^2].$$

For $N \leq 4$ the genus of $X_1(N)$ is 0; in fact, this is the case for $N \leq 10$. To compute the genus of $X_0(N)$, see [**Shi1**, Prop. 1.43].

Letting $\Gamma = \Gamma' = \Gamma_1(N)$ and taking γ in $\Gamma_0(N)$, we obtain the action of $(\mathbf{Z}/N\mathbf{Z})^* \cong \Gamma_0(N)/\Gamma_1(N)$ on $X_1(N)$ extending the one on $Y_1(N)$. The quotient of $X_1(N)$ by this action is naturally identified with $X_0(N)$. Note also that the degeneracy maps $Y_0(M) \to Y_0(N)$ defined by $\gamma = \iota_d$ (see §7.3) for divisors d of M/N extend to maps $X_0(M) \to X_0(N)$. In particular for $M = Np$ we denote the extensions of α and β by the same symbols and $\alpha \circ {}^t\beta$ gives rise to a correspondence on $X_0(N)$ which we again denote T_p. Similarly using $\Gamma' = \Gamma_1(N)$ and $\Gamma = \Gamma_1(N) \cap \Gamma_0(Np)$ we define the modular correspondence T_p on $X_1(N)$.

9.2. Generalized elliptic curves.

PRIMARY REFERENCES:
[**DeRa**, Chapter II] and [**Del4**].

Our next task is to explain the modular interpretation of the cusps as generalized elliptic curves. This interpretation was introduced by Deligne and Rapoport [**DeRa**] in their construction of smooth, proper models over $\mathbf{Z}[1/N]$ for $X_0(N)$ and $X_1(N)$.

To motivate the definition of a generalized elliptic curve, let us first recall that we identify the point $\mathrm{SL}_2(\mathbf{Z})\tau$ of $Y_0(1)$ with the elliptic curve $E_\tau = \mathbf{C}/(\mathbf{Z} \oplus \mathbf{Z}\tau)$ (up to isomorphism). Observe that as a complex Lie group, E_τ is isomorphic to $\mathbf{C}^\times/\{e^{2\pi i n\tau}\}$ via the exponential map $z \mapsto e^{2\pi i z}$. Moreover if $\tau = x+iy$ with $y > 0$, then an equation for the curve E_τ over $\mathbf{C}$ is obtained by substituting $q = e^{2\pi i \tau}$ in the power series that appear in the definition of the Tate curve (Example 8.1.3). This provides the following intuitive description of the behavior of E_τ as $\mathrm{SL}_2(\mathbf{Z})\tau$ tends to the cusp $\mathrm{SL}_2(\mathbf{Z}) \cdot \infty$; the real number y tends to ∞, q tends to 0 and the equation for E_τ degenerates to

$$Y^2 + XY = X^3. \tag{9.2.1}$$

So the modular interpretation of the cusp of $X_0(1)$ should be provided by the "degenerate elliptic curve" C, the projective variety over $\mathbf{C}$ defined by (9.2.1). Note that the only singularity of C is the ordinary double point $X = Y = 0$. Writing C^{reg} for the smooth locus of C, we can define an "addition" morphism $+ : C^{\mathrm{reg}} \times C \to C$ by substituting $q = 0$ in the group law for the Tate curve. Moreover the isomorphism $\mathbf{G_m} = \mathrm{Spec}\,\mathbf{C}[Z, Z^{-1}] \to C^{\mathrm{reg}}$ defined by $X \mapsto Z(Z-1)^{-2}$,

$Y \mapsto Z(Z-1)^{-3}$ extends to a normalization $\mathbf{P}^1 \to C$ and the diagram

$$\begin{array}{ccc} \mathbf{G_m} \times \mathbf{P}^1 & \to & \mathbf{P}^1 \\ \downarrow & & \downarrow \\ C^{\mathrm{reg}} \times C & \to & C \end{array}$$

commutes, where the upper arrow is given by the natural action of $\mathbf{G_m}$ on $\mathbf{P}^1$.

The pair $(C, +)$ we have just described is called the Néron 1-gon over $\mathbf{C}$. To interpret the cusps of other modular curves, we shall need to consider Néron polygons [**DeRa**, §II.1]. We define the Néron N-gon $(C_N, +)$ over an algebraically closed field k as follows. The scheme C_N over k has N irreducible components, each isomorphic to the projective line $\mathbf{P}^1$, and N ordinary double points. To complete the characterization of C_N up to isomorphism, we index the components with elements of $\mathbf{Z}/N\mathbf{Z}$ and require the normalization $r : \coprod_{i \in \mathbf{Z}/N\mathbf{Z}} \mathbf{P}^1 \to C_N$ to send $(\infty)_i$ and $(0)_{i+1}$ to the same point for each i. Thus r restricts to an isomorphism $\coprod_{i \in \mathbf{Z}/N\mathbf{Z}} \mathbf{G_m} \to C_N^{\mathrm{reg}}$, and the dual graph of C_N is an N-gon. We let $+$ be the morphism $C_N^{\mathrm{reg}} \times C_N \to C_N$ characterized by the commutativity of the diagram

$$\begin{array}{ccc} \coprod_{i \in \mathbf{Z}/N\mathbf{Z}} \mathbf{G_m} \times \coprod_{i \in \mathbf{Z}/N\mathbf{Z}} \mathbf{P}^1 & \to & \coprod_{i \in \mathbf{Z}/N\mathbf{Z}} \mathbf{P}^1 \\ \downarrow & & \downarrow \\ C_N^{\mathrm{reg}} \times C_N & \overset{+}{\to} & C_N, \end{array}$$

where the vertical arrows are given by r and the top arrow is defined by

$$((x)_i, (y)_j) \mapsto (xy)_{i+j}$$

on closed points. In particular, $+$ extends the addition on the group scheme $C_N^{\mathrm{reg}} \cong \mathbf{G_m} \times \mathbf{Z}/N\mathbf{Z}$ to an action of C_N^{reg} on C_N, and the induced action on the dual graph is via rotations.

We are now ready to generalize the notion of an elliptic curve so as to include schemes whose geometric fibers are elliptic curves or Néron polygons. A *generalized elliptic curve* ([**DeRa**, II.1.4]) over S is a pair $(E, +)$ where E is a scheme of curves over S and $+$ is an S-morphism $E^{\mathrm{reg}} \times_S E \to E$. We require that $+$ makes E^{reg} a commutative group scheme over S acting on E and that the geometric fibers of $(E, +)$ are elliptic curves or Néron polygons. Two elementary observations are that a generalized elliptic curve over S is smooth if and only if it is an elliptic curve and that a generalized elliptic curve over an algebraically closed field is either an elliptic curve or a Néron polygon. The key example is the following (see [**DeRa**, Chapter VII] and [**Del4**, §7]:

EXAMPLE 9.2.1. Define the Tate curve E_q as in Example 8.1.3, but working over $S = \mathrm{Spec}\,(\mathbf{Z}[[q]])$ rather than $\mathrm{Spec}\,(\mathbf{Z}(\!(q)\!))$. Then E_q^{reg} is the complement of the closed subscheme defined by $X = Y = q = 0$. The group law on the elliptic curve $E_q \times_S \mathrm{Spec}\,(\mathbf{Z}(\!(q)\!))$ extends to a morphism $+ : E_q^{\mathrm{reg}} \times_S E_q \to E_q$ making E_q^{reg} a commutative group scheme acting on E_q. Suppose that $s : \mathbf{Z}[[q]] \to k$ defines a geometric point of S. If $s(q) = 0$, then $(E_{q,k}, +_k)$ is isomorphic to the Néron 1-gon over k. On the other hand if $s(q) \neq 0$, then $E_{q,k}$ is an elliptic curve. In particular if $k = \mathbf{C}$ and $s(q) = e^{2\pi i \tau}$ with $\tau = x + iy$ and $y > 0$, then this elliptic curve is isomorphic to $\mathbf{C}^\times / \{e^{2\pi i n \tau}\} \cong E_\tau$.

9.3. Canonical models revisited.

PRIMARY REFERENCES:
[**DeRa**] and [**KaMa**, Chapter 8].

We can now regard the Riemann surface $X_1(N)$ as a moduli space. Its points are naturally in bijection with the isomorphism classes of pairs (E, P) where E is a generalized elliptic curve over $\mathbf{C}$ and P is a point of E^{reg} of order N such that the subgroup generated by P meets every component. Indeed we shall now complement the bijection defined in §7.2 by a natural one-to-one correspondence between the set of cusps of $X_1(N)$ and the set of isomorphism classes of pairs (E, P) where E is a Néron polygon. To a pair of integers (c, d) we associate the pair (E, P) where E is the $N/(c, N)$-gon over $\mathbf{C}$ and P is the point $(e^{2\pi i d/N})_{c/(c,N)}$. The group generated by P meets every component, and if d is relatively prime to (c, N), then P has order N. The image of the pair (c, d) in (9.1.1) determines (E, P) up to isomorphism, and the resulting map from the set of cusps is the desired bijection.

To define canonical models for the curves $X_1(N)$, we proceed as we did for $Y_1(N)$ in §8.2, but using *generalized* elliptic curves. More precisely, for a $\mathbf{Z}[1/N]$-scheme S we define $\mathcal{G}_1(N)(S)$ to be the set of isomorphism classes of pairs $(\mathcal{E}, \mathcal{P})$ where $\mathcal{E}$ is a generalized elliptic curve and $\mathcal{P}$ is a section $S \to \mathcal{E}^{\mathrm{reg}}$ of exact order N. We further require that for all geometric points $s : \operatorname{Spec} k \to S$, the image of the resulting immersion $(\mathbf{Z}/N\mathbf{Z})_k \hookrightarrow \mathcal{E}_k^{\mathrm{reg}}$ meets every component [**DeRa**, IV.4.14]. By the results of Deligne and Rapoport [**DeRa**, Chapter IV] (see [**Gross**, Proposition 2.1]), if $N > 4$, then $\mathcal{G}_1(N)$ is representable by a smooth curve $\mathcal{X}_1(N)$ over $\operatorname{Spec} \mathbf{Z}[1/N]$

The bijection $X_1(N) \to \mathcal{X}_1(N)(\mathbf{C})$ we defined above is holomorphic and we now have a smooth, proper model for $X_1(N)$ over $\mathbf{Z}[1/N]$. One can define an analogous functor $\mathcal{G}_0(N)$ and a smooth, proper model for $X_0(N)$ over $\mathbf{Z}[1/N]$ is provided by a scheme $\mathcal{X}_0(N)$ which can be interpreted as a coarse moduli scheme. (This is also the case for $X_1(N) = X_0(N)$ for $N \leq 4$.) We also have a natural action of $(\mathbf{Z}/N\mathbf{Z})^\times$ on $\mathcal{X}_1(N)$ and $\mathcal{X}_0(N)$ can be identified with the quotient scheme. There is a tautological natural transformation $\mathcal{F}_1(N) \to \mathcal{G}_1(N)$ which identifies $\mathcal{Y}_1(N)$ with an open subscheme of $\mathcal{X}_1(N)$, and similarly $\mathcal{Y}_0(N)$ can be identified with an open subscheme of $\mathcal{X}_0(N)$.

EXAMPLE 9.3.1. The isomorphism $\mathcal{Y}_0(1) \cong \mathbf{A}^1_{\mathbf{Z}} = \operatorname{Spec}(\mathbf{Z}[j])$ in §8 extends to an isomorphism $\mathcal{X}_0(1) \cong \mathbf{P}^1_{\mathbf{Z}}$. The resulting bijection $\mathcal{G}_0(N)(k) \to \mathbf{P}^1(k)$ for an algebraically closed field k sends an elliptic curve to its j-invariant and the 1-gon to ∞.

EXAMPLE 9.3.2. The curve $\mathcal{X}_0(11)_{\mathbf{Q}}$ is of genus one (see Example 9.1.5) and has a rational point, the cusp at ∞ for instance. Therefore $\mathcal{X}_0(11)_{\mathbf{Q}}$ can be given the structure of an elliptic curve over $\mathbf{Q}$. It is known to be isomorphic over $\mathbf{Q}$ to the curve $\mathcal{E}_{\mathbf{Q}}$ where $\mathcal{E}$ is the elliptic curve of Example 8.1.1 (see the tables of [**Ant4**] for example).

By [**DeRa**, §VII.2] (see also [**MaWi**, §2.10]), a formal neighborhood of the complement of $\mathcal{F}_0(N)$ in $\mathcal{G}_0(N)$, or $\mathcal{F}_1(N)$ in $\mathcal{G}_1(N)$, can be described using Tate curves.

EXAMPLE 9.3.3. The Tate curve E_q of Example 9.2.1 is a generalized elliptic curve over $\mathbf{Z}[[q]]$. It therefore defines an element of $\mathcal{G}_0(1)(\mathbf{Z}[[q]])$ and thus gives rise to a morphism

$$\phi : \operatorname{Spec} \mathbf{Z}[[q]] \to \operatorname{Spec} \mathbf{Z}[j^{-1}] \hookrightarrow \mathcal{X}_0(1).$$

The morphism can be made explicit by writing j^{-1} as a formal power series in q. The complement of $\mathcal{Y}_0(1)$ in $\mathcal{X}_0(1)$ is defined by $j^{-1} = 0$ and is the image

of the immersion $\operatorname{Spec} \mathbf{Z} \to \mathcal{X}_0(1)$ defined by the reduction mod q of the Tate curve (which can be considered a Néron 1-gon over $\mathbf{Z}$). Moreover ϕ induces an isomorphism between the formal scheme $\operatorname{Spf} \mathbf{Z}[[q]]$ and the completion of $\mathcal{X}_0(1)$ along the complement of $\mathcal{Y}_0(1)$.

EXAMPLE 9.3.4. The two cusps, $\Gamma_0(p)\cdot\infty$ and $\Gamma_0(p)\cdot 0$, of $X_0(p) \cong \mathcal{X}_0(p)(\mathbf{C}) \cong \mathcal{G}_0(p)(\mathbf{C})$ are given by the pairs (C, D) and (C', D') respectively, where C is the 1-gon over $\mathbf{C}$ and D is the image of $\boldsymbol{\mu}_p \hookrightarrow \mathbf{G_m} \cong C^{\mathrm{reg}}$, and C' is the p-gon over $\mathbf{C}$ and D' is any subgroup of C'^{reg} of order p which meets every component. Let $\mathcal{E}$ be the Tate curve $E_{q,S}$ over $S = \mathbf{Z}[1/p][[q]]$. Then $\mathcal{E}^{\mathrm{reg}}$ has a subgroup scheme $\mathcal{D}$ canonically isomorphic to $\boldsymbol{\mu}_p$. The pair $(\mathcal{E}, \mathcal{D})$ defines a morphism $\phi : \operatorname{Spec} S \to \mathcal{X}_0(p)$ and the cusp $\Gamma_0(p)\cdot\infty$ corresponds to the complex point of $\mathcal{X}_0(p)$ given by $q \mapsto 0$. The Tate curve over $\mathbf{Z}[1/p](\!(q^{1/p})\!)$ extends to a generalized elliptic curve $\mathcal{E}'$ over $S' = \mathbf{Z}[1/p][[q^{1/p}]]$ whose geometric fibers are elliptic curves or p-gons [**DeRa**, §VII.1]. (There is a map $\mathcal{E}' \to \mathcal{E}_{q,S'}$ which is an isomorphism over $\mathbf{Z}[1/p](\!(q^{1/p})\!)$ but contracts those components of the singular geometric fibers of $\mathcal{E}'$ which do not meet the unit section.) Furthermore, $\mathcal{E}'^{\mathrm{reg}}$ has a finite flat subgroup scheme $\mathcal{D}'$ of degree p which meets every component. The pair $(\mathcal{E}', \mathcal{D}')$ defines $\phi' : \operatorname{Spec} S' \to \mathcal{X}_0(p)$ and mapping q to 0 yields the complex point corresponding to $\Gamma_0(p)\cdot 0$. The completion of $\mathcal{X}_0(p)$ along the complement of $\mathcal{Y}_0(p)$ is isomorphic to $\operatorname{Spf} S \coprod \operatorname{Spf} S'$. A similar construction, but involving more cusps, describes a formal neighborhood of the complement of $\mathcal{Y}_0(N)$ in $\mathcal{X}_0(N)$ for composite N.

EXAMPLE 9.3.5. Note that the cusp $\Gamma_1(N)\cdot\infty$ of $X_1(N) \cong \mathcal{X}_1(N)(\mathbf{C})$ is usually not defined over $\mathbf{Q}$, as it corresponds to C_1 together with a generating section of $\boldsymbol{\mu}_N \subset \mathbf{G_m} \cong C_1^{\mathrm{reg}}$. Rather the closure in $\mathcal{X}_1(N)$ of the image of this complex point is isomorphic to $\operatorname{Spec} \mathbf{Z}[1/N, e^{2\pi i/N} + e^{-2\pi i/N}]$. Moreover a formal neighborhood of this closed subscheme can be described as in Example 9.3.4 but now using the Tate curve over $\mathbf{Z}[1/N, e^{2\pi i/N}][[q]]$ (see §12.3). In fact, one can give an explicit description of the completion of $\mathcal{X}_1(N)$ along the complement of $\mathcal{Y}_1(N)$ as a disjoint union of formal spectra of power series rings in one variable over etale extensions of $\mathbf{Z}[1/N]$.

VARIANT 9.3.6. Recall that the alternate convention of Variant 8.2.2 provides a model $\mathcal{Y}_\mu(N)$ for $Y_1(N)$ over $\mathbf{Z}$. Using immersions $(\boldsymbol{\mu}_N)_S \hookrightarrow \mathcal{E}^{\mathrm{reg}}$ instead of sections $S \to \mathcal{E}^{\mathrm{reg}}$, one obtains a model $\mathcal{X}_\mu(N)$ over $\mathbf{Z}$ for $X_1(N)$ (assuming $N > 4$). Then $\mathcal{X}_\mu(N)$ contains $\mathcal{Y}_\mu(N)$ as an open subscheme and the cusp $\Gamma_1(N)\cdot\infty$ is defined over $\mathbf{Q}$ with respect to this model. This convention will be more convenient for discussing the q-expansion principle in §12.3.

Although $\mathcal{X}_\mu(N)$ is not proper, we have the following [**Katz2**, §II.2.5].

THEOREM 9.3.7. *The scheme $\mathcal{X}_\mu(N)$ is smooth over $\mathbf{Z}$ with geometrically irreducible fibers, and $X_\mu(N)_{\mathbf{Z}[1/N]}$ is proper over $\mathbf{Z}[1/N]$.*

We can proceed as in §8.3–§8.5, but now using proper models for the compactified modular curves (see [**DeRa**, §V.1]). Suppose p is a prime not dividing N and assume that $N > 4$. Let $\Gamma = \Gamma_1(N) \cap \Gamma_0(p)$ and let $X = \Gamma\backslash\mathfrak{H}^*$. We first define a proper model $\mathcal{X}$ for X over $\mathbf{Z}[1/N]$ which parametrizes triples $(\mathcal{E}, \mathcal{P}, \mathcal{D})$ where now $\mathcal{E}$ is a generalized elliptic curve over S and for each geometric point $\operatorname{Spec} k \to S$ the image of the fiber $(\mathbf{Z}/N\mathbf{Z})_k \times \mathcal{D}_k \hookrightarrow \mathcal{E}_k^{\mathrm{reg}}$ meets every component. We then define models as in §8.3 for the degeneracy maps $\alpha, \beta : X \to X_1(N)$ and consequently the

modular correspondence $T_p : X \to X_1(N) \times X_1(N)$. We remark only on a slight complication in the case of α. The pair $(\mathcal{E}, p\mathcal{P})$ does not necessarily satisfy the condition on the geometric fibers $(\mathbf{Z}/N\mathbf{Z})_k \hookrightarrow \mathcal{E}_k^{\mathrm{reg}}$. There is however a generalized elliptic curve $\overline{\mathcal{E}}$ over S and a morphism $\pi : \mathcal{E} \to \overline{\mathcal{E}}$ which induces on geometric fibers an isomorphism $G_k \to \overline{\mathcal{E}}^{\mathrm{reg}}$ where G_k is the open subgroup scheme of $\mathcal{E}_k^{\mathrm{reg}}$ consisting of the components which meet the image of $(\mathbf{Z}/N\mathbf{Z})_k$. (Thus those which do not are contracted to points on the singular locus. See [**DeRa**, §IV.1] and Example 9.3.4 above.) We define a model for α using $(\overline{\mathcal{E}}, p\pi \circ \mathcal{P})$ where $(\mathcal{E}, \mathcal{P}, \mathcal{D})$ is the universal triple over $\mathcal{X}$.

We can analyze $\mathcal{X}_{\mathbf{F}_p}$ just as we did $\mathcal{Y}_{\mathbf{F}_p}$ in §8.4. The definition of

$$i_F : \mathcal{X}_1(N)_{\mathbf{F}_p} \to \mathcal{X}_{\mathbf{F}_p}$$

is essentially as in §8.4, but slightly more care is required to define i_V. In particular, $\mathcal{E}_0^{(p)}$ must be replaced by a generalized elliptic curve $(\mathcal{E}_0^{(p)})'$ so that the number of components on each singular fiber is divisible by p. (See Example 9.3.4.) We then obtain (8.4.1), but now for endomorphisms of $\mathcal{X}_1(N)_{\mathbf{F}_p}$. We find that $\mathcal{X}$ is regular and that $\mathcal{X}_{\mathbf{F}_p}$ has two irreducible components, each isomorphic to $\mathcal{X}_1(N)_{\mathbf{F}_p}$, crossing transversally at points where the geometric fiber of $\mathcal{E}_0$ is a supersingular elliptic curve. Thus the map

$$i_F \coprod i_V : \mathcal{X}_1(N)_{\mathbf{F}_p} \coprod \mathcal{X}_1(N)_{\mathbf{F}_p} \to \mathcal{X}_{\mathbf{F}_p} \tag{9.3.1}$$

identifies $\mathcal{X}_1(N)_{\mathbf{F}_p} \coprod \mathcal{X}_1(N)_{\mathbf{F}_p}$ with the normalization of $\mathcal{X}_{\mathbf{F}_p}$, which we denote $(\mathcal{X}_{\mathbf{F}_p})^{\sim}$.

A consequence of this description is the formula [**DeRa**, VI (6.11.2)]

$$g' = 2g + s - 1 \tag{9.3.2}$$

where g' is the genus of X, g is the genus of $X_1(N)$ and s is the number of supersingular points (E, P) in $\mathcal{X}_1(N)(\overline{\mathbf{F}}_p)$. Combined with the Hurwitz formula (recall we assume $N > 4$ so there are no elliptic elements), this yields

$$s = \frac{1}{12}(p-1)[\mathrm{PSL}_2(\mathbf{Z}) : \Gamma_1(N)].$$

Note also that the Eichler-Shimura relation, (8.5.1) or (8.5.2), remains valid for the correspondence $\mathcal{X}_{\mathbf{F}_p} \to \mathcal{X}_1(N)_{\mathbf{F}_p} \times \mathcal{X}_1(N)_{\mathbf{F}_p}$.

One finds a similar description in terms of $\mathcal{X}_0(N)$ for a coarse moduli scheme $\mathcal{X}_0'(Np)$ which is a proper model for $X_0(Np)$ over $\mathbf{Z}[1/N]$ (see [**DeRa**, §VI.6]). In particular $\mathcal{X}_0'(Np)_{\mathbf{F}_p}$ can be described in terms of $\mathcal{X}_0(p)_{\mathbf{F}_p}$ and the Eichler-Shimura relation holds. (The only changes are that $\mathcal{X}_0'(Np)$ is not necessarily regular and the formula for the number of supersingular points is slightly more complicated.) In particular, $\mathcal{X}_0'(p)_{\mathbf{F}_p}$ has two irreducible components, each isomorphic to $\mathcal{X}_0(1)_{\mathbf{F}_p} \cong \mathbf{P}^1_{\mathbf{F}_p}$. A formal neighborhood of the complement of $\mathcal{Y}_0'(p)$ in $\mathcal{X}_0'(p)$ is described exactly as in Example 9.3.4. Note that there is a "cuspidal section" in $\mathcal{X}_0'(p)(\mathbf{Z})$ whose image in $\mathcal{X}_0'(p)(\mathbf{C}) \cong X_0(p)$ is $\Gamma_0(p) \cdot \infty$ (respectively, $\Gamma_0(p) \cdot 0$) and whose image in $\mathcal{X}_0'(p)(\mathbf{F}_p)$ factors through i_F (respectively, i_V).

Finally, we remark that the models for the various w-operators defined in Remark 8.4.1 can be extended to the proper models considered above.

10. Jacobians of modular curves

In this section we will examine the Jacobians of modular curves, their reduction modulo primes, and the endomorphisms induced by Hecke operators.

10.1. Abelian varieties and Jacobians.

PRIMARY REFERENCES:
[**Mum1**], [**Rosen**], [**Mil2**], [**Mil3**] and [**BLRa**, Chapters 8,9].

We now review some generalities concerning abelian varieties and Jacobians of algebraic curves.

We first recall that an abelian variety A over an algebraically closed field k is a proper group variety over k. It is necessarily smooth, projective and commutative [**Mil2**, §1,2]. One can consider more generally abelian schemes, or families of abelian varieties, over an arbitrary base scheme S. An abelian scheme over S is a smooth proper group scheme over S whose geometric fibers are abelian varieties [**Mil2**, §20].

If $k = \mathbf{C}$ and A is a g-dimensional abelian variety, then the complex manifold $A(\mathbf{C})$ is isomorphic to a complex torus V/L where V is a g-dimensional vector space and L is a discrete subgroup of rank $2g$ [**Rosen**, §1]. An arbitrary complex torus V/L can be identified with the set of complex points of an abelian variety over $\mathbf{C}$ if and only if V/L possesses a non-degenerate Riemann form [**Rosen**, §3], i.e., a positive definite Hermitian form on V whose imaginary part is integer valued on L. In this case, the same is true for the complex torus V^*/L^* where $V^* \subset \mathrm{Hom}_{\mathbf{R}}(V, \mathbf{C})$ is the space of conjugate linear functions on V (i.e., additive functions ϕ satisfying $\phi(zv) = \overline{z}\phi(v)$ for all $z \in \mathbf{C}$, $v \in V$), and $L^* = \{\phi \in V^* | \phi(L) \subset \mathbf{R} + i\mathbf{Z}\}$. If A and A^* are abelian varieties satisfying $A(\mathbf{C}) \cong V/L$ and $A^*(\mathbf{C}) \cong V^*/L^*$, then A^* is called the dual abelian variety of A [**Rosen**, §4]. Note that A is isomorphic to $(A^*)^*$.

Now let C be a Riemann surface and let W denote the complex vector space of holomorphic differentials on C. Consider the complex torus V/L where $V = \mathrm{Hom}(W, \mathbf{C})$ and L is the image of the map $H_1(C, \mathbf{Z}) \to \mathrm{Hom}(W, \mathbf{C})$ defined by integration. Note that the cotangent space of V/L at the origin may be naturally identified with W. The intersection pairing on $H_1(C, \mathbf{Z})$ can be used to define a nondegenerate Riemann form on V/L, and the resulting abelian variety J is called the Jacobian of C [**Mil3**, §2]. Moreover this Riemann form gives rise to a canonical isomorphism $J \cong J^*$.

Another interpretation of the Jacobian of C is provided by the Picard functor Pic^0 (see [**Mil3**, §1]). Let $\mathrm{Div}^0(C)$ denote the group of divisors on C of degree zero, and let $\mathrm{Pic}^0(C)$ denote $\mathrm{Div}^0(C)$ modulo the group of principal divisors. Integration then defines a natural map $\mathrm{Div}^0(C) \to V/L$ which, according to the Abel-Jacobi theorem, induces a natural isomorphism of groups $\mathrm{Pic}^0(C) \cong J(\mathbf{C})$. Now choose a base-point P in C and define a mapping $C \to \mathrm{Pic}^0(C)$ by sending Q to the divisor $Q - P$. The resulting map $C \to V/L$ is analytic and induces an isomorphism $H^0(J(\mathbf{C}), \Omega^1) \to H^0(C, \Omega^1) = W$ which is independent of the base-point. Moreover the isomorphism is compatible with the natural identification of W with the cotangent space of $J(\mathbf{C}) \cong V/L$ at the origin.

To describe the Jacobian of a curve over any field, or indeed an arbitrary base scheme S, we use the Picard functor [**Mil3**, §8], [**BLRa**, Chapter 8]. For a morphism of schemes $s : X \to S$, Grothendieck [**Gro1**] defines a relative Picard

functor $\mathrm{Pic}_{X/S}$ on S-schemes by "sheafifying" the functor which sends T to the group of isomorphism classes of invertible sheaves on $X_T = X \times_S T$. Under quite general hypotheses (see Chapters 8 and 9 of [**BLRa**]) this contravariant functor is represented by a group scheme over S, and we denote its identity component $\mathrm{Pic}^0_{X/S}$. The definition is functorial in X, so that a morphism $Y \to X$ of S-schemes gives rise to a natural transformation $\mathrm{Pic}_{X/S} \to \mathrm{Pic}_{Y/S}$ and consequently a morphism $\mathrm{Pic}^0_{X/S} \to \mathrm{Pic}^0_{Y/S}$. We remark also that formation of $\mathrm{Pic}^0_{X/S}$ commutes with base change, meaning that $\mathrm{Pic}^0_{(X_T)/T}$ is naturally isomorphic to $\mathrm{Pic}^0_{X/S} \times_S T$.

If $X \to S$ is a relative curve, meaning that it is smooth and proper and its geometric fibers are curves, then $\mathrm{Pic}^0_{X/S}$ is an abelian scheme which we denote $J_{X/S}$ and call the Jacobian of X (over S), [**BLRa**, §9.2]. If also $S = \mathrm{Spec}\, k$ for an algebraically closed field k, then $\mathrm{Pic}_{X/S}(S)$ may be identified with the group of invertible sheaves on X, or equivalently, with $\mathrm{Div}(X)$ modulo the group of principal divisors. Then $\mathrm{Pic}^0_{X/S}(S)$ may be identified with $\mathrm{Pic}^0(X)$, the group $\mathrm{Div}^0(X)$ modulo the group of principal divisors. Moreover if $k = \mathbf{C}$, then the isomorphism $V/L \cong J(\mathbf{C}) \cong J_{X/\mathbf{C}}(\mathbf{C})$ is analytic, so our two descriptions of the Jacobian in this case are equivalent.

The relative Picard functor also provides a general construction of the dual of an abelian scheme. If A is an abelian scheme over S, then $\mathrm{Pic}_{A/S}$ is representable by a scheme, and $\mathrm{Pic}^0_{A/S}$ is an abelian scheme, [**BLRa**, §8.4, Theorem 5], [**FaCh**, I.1]. We write A^* for $\mathrm{Pic}^0_{A/S}$ and call it the dual abelian scheme of A. Again there is a natural isomorphism $A \cong (A^*)^*$. For a relative curve X over S there is a general construction of a "Θ-divisor" on $J_{X/S}$ which gives rise to an isomorphism $\phi_{X/S}$ of $J_{X/S}$ with $J^*_{X/S}$, [**BLRa**, §9.4]. The constructions of the dual abelian scheme, its biduality and the autoduality of the Jacobian are compatible with base-change. They are also compatible with the descriptions given above in the case $S = \mathrm{Spec}\, \mathbf{C}$.

A morphism $\pi : Y \to X$ of relative curves over S induces by Picard functoriality a homomorphism of abelian schemes $\pi^* : J_{X/S} \to J_{Y/S}$. We obtain also a homomorphism $\pi_* : J_{Y/S} \to J_{X/S}$ defined by the composite $\phi^{-1}_{Y/S} \circ (\pi^*)^* \phi_{X/S}$ where $(\pi^*)^* : J^*_{Y/S} \to J^*_{X/S}$ is again defined by Picard functoriality. We thus have two functors from the category of relative curves over S to the category of abelian schemes over S; the contravariant Picard functor Pic^0 defined by $\mathrm{Pic}^0(X) = J_{X/S}$ and $\mathrm{Pic}^0(\pi) = \pi^*$, and the covariant Albanese functor Alb defined by $\mathrm{Alb}(X) = J_{X/S}$ and $\mathrm{Alb}(\pi) = \pi_*$, [**Mil3**, §6]. If $S = \mathrm{Spec}\, k$ for an algebraically closed field k, then π^* on $J_{X/S}(S)$ is induced by the map $\mathrm{Div}(X) \to \mathrm{Div}(Y)$ defined by pull-back of divisors; a point $x \in X(S)$ is sent to $\Sigma_{y \in \pi^{-1}(x)} e_{y/x} y$ where $e_{y/x}$ is the ramification degree. On the other hand, π_* on $J_{Y/S}(S)$ is induced by the map $\mathrm{Div}(Y) \to \mathrm{Div}(X)$ which sends $y \in Y(S)$ to $\pi(y)$. Note that $\pi_* \circ \pi^*$ is simply multiplication by the degree of π.

There is in general a natural isomorphism of $s^* \Omega^1_{X/S}$ with the cotangent sheaf $i^* \Omega^1_{J_{X/S}/S}$ along the zero section $i : S \to J_{X/S}$. For $S = \mathrm{Spec}\, k$, this can be viewed as an isomorphism $H^0(X, \Omega^1_{X/S}) \cong \mathrm{Cot}_0(J_{X/S})$ (see [**Mil3**, Proposition 2.2]). Consider now the maps induced by π^* and π_* on the cotangent spaces at

zero of the Jacobians. We get a commutative diagram

$$\begin{array}{ccc} H^0(X, \Omega^1_{X/S}) & \to & H^0(Y, \Omega^1_{Y/S}) \\ \downarrow & & \downarrow \\ \mathrm{Cot}_0(J_{X/S}) & \to & \mathrm{Cot}_0(J_{Y/S}) \end{array}$$

where the upper arrow is obtained by Serre duality from the natural map

$$H^1(Y, \mathcal{O}_Y) \to H^1(X, \mathcal{O}_X)$$

and the lower arrow is induced by π^*. The description of the map induced by π_* on cotangent spaces is even simpler, for it is given by pull-back of differentials $W_X = H^0(X, \Omega^1_{X/S}) \to W_Y = H^0(Y, \Omega^1_{Y/S})$. For $k = \mathbf{C}$, the isomorphism $H^0(X, \Omega^1_{X/S}) \cong \mathrm{Cot}_0(J_{X/S})$ is simply the identification of W with the cotangent space at zero of V/L. We find also that the map π_* is given on complex tori by $V_Y/L_Y \to V_X/L_X$ where $V_Y \to V_X$ is dual to the natural pull-back $W_X \to W_Y$ and $L_Y \to L_X$ is defined by $H_1(Y(\mathbf{C}), \mathbf{Z}) \to H_1(X(\mathbf{C}), \mathbf{Z})$.

10.2. Models for Jacobians.

PRIMARY REFERENCES:

[**Shi1**, Chapter 7], [**MaWi**, §2.1,2.5], [**BLRa**, Chapter 1] and [**Ray3**].

Let us consider the Jacobian of the curve $X_0(N)$, denoted $J_0(N)$. The modular correspondence T_p, regarded as an endomorphism of $\mathrm{Div}(X_0(N))$ induces an endomorphism of $J_0(N)$, which we also denote T_p. We have that $T_p = \alpha_* \circ \beta^*$ where α and β are the degeneracy maps $X_0(Np) \to X_0(N)$ defined in §9.1 (see e.g. [**MaWi**, §2.5]). Since the curve $X_0(N)$ is defined over $\mathbf{Q}$, so is the abelian variety $J_0(N)$. Moreover, since α and β are defined over $\mathbf{Q}$, so is T_p. More generally we can define endomorphisms T_n of $J_0(N)$ and of $J_1(N)$, the Jacobian of $X_1(N)$. These are defined over $\mathbf{Q}$ as is the action of $(\mathbf{Z}/N\mathbf{Z})^\times$ on $J_1(N)$ defined by the operators $\langle d \rangle_*$.

REMARK 10.2.1. Some authors, for example Ribet [**Rib4**], use T_p to denote the endomorphism $\beta_* \circ \alpha^*$, which we shall call T_p^*. More generally, we can consider endomorphisms T_n^* of $J_0(N)$ and $J_1(N)$. For n relatively prime to N, these are related by $T_n^* = \langle n \rangle^* T_n$ on $J_1(N)$ and $T_n^* = T_n$ on $J_0(N)$ coincide.

REMARK 10.2.2. The Atkin-Lehner involutions w_Q give rise to involutions $w_{Q,*}$ of $J_0(N)$ defined over $\mathbf{Q}$. Similarly, for p not dividing N, the endomorphism $w_{p,*}$ of the Jacobian of $\Gamma_1(N,p)\backslash\mathfrak{H}^*$ is defined over $\mathbf{Q}$ and satisfies $w_{p,*}^2 = \langle p \rangle_*$. The involution $w_{N,*}$ of $J_1(N)$ is defined over $\mathbf{Q}(e^{2\pi i/N})$ and satisfies $w_{N,*} T_n w_{N,*} = T_n^*$ for all positive integers n. Thus $w_{N,*}$ intertwines the operators defined using our conventions and those mentioned in Remark 10.2.1. We find also that $w_{N,*} \langle d \rangle_* w_{N,*} = \langle d \rangle^*$ for all d relatively prime to N.

As we did for the modular curves, we would like to construct "good" integral models for their Jacobians and study their reduction modulo primes. We will then examine the effect of the Hecke operators. To begin, recall that we have defined a model for $X_0(N)$ over $\mathbf{Z}[1/N]$. It is obtained from the relative curve $\mathcal{X}_0(N)$ over $\mathbf{Z}[1/N]$ which is a coarse moduli scheme parametrizing pairs $(\mathcal{E}, \mathcal{C})$ where $\mathcal{E}$ is a generalized elliptic curve and $\mathcal{C}$ is a "cyclic subgroup scheme" of order N. Its Jacobian $J_{\mathcal{X}_0(N)/\mathbf{Z}[1/N]} = \mathrm{Pic}^0_{\mathcal{X}_0(N)/\mathbf{Z}[1/N]}$ is an abelian scheme which can be viewed as a model for $J_0(N)$ using the isomorphism of $J_{\mathcal{X}_0(N)/\mathbf{Z}[1/N]} \times \mathbf{C}$ with $J_0(N)$. Now consider a geometric fiber $J_{\mathcal{X}_0(N)/\mathbf{Z}[1/N]} \times k$ where k is an algebraic closure of $\mathbf{Q}$ or

of $\mathbf{F}_q$ where q is a prime not dividing N. The group of closed points of this abelian variety is naturally isomorphic to $\mathrm{Pic}^0(\mathcal{X}_0(N)_k)$.

To obtain a model for the endomorphism T_p using the Picard functor, let us first work over $\mathbf{Z}[1/Np]$. For then we can consider the two natural degeneracy maps $\alpha', \beta' : \mathcal{X}_0(Np) \to \mathcal{X}_0(N)_{\mathbf{Z}[1/Np]}$ of relative curves over $\mathbf{Z}[1/Np]$. The endomorphism $\mathcal{T}'_p = \alpha'_* \circ (\beta')^*$ of $J_{\mathcal{X}_0(N)/\mathbf{Z}[1/N]} \times \mathbf{Z}[1/Np]$ is in an obvious sense a model for T_p. For a prime q not dividing Np, $(\mathcal{T}'_p)_{\mathbf{F}_q}$ is given by the composite $(\alpha'_{\mathbf{F}_q})_* \circ (\beta'_{\mathbf{F}_q})^*$. We can also describe the effect of $\mathcal{T}'_p$ on $J_{\mathcal{X}_0(N)/\mathbf{Z}[1/N]}(k) \cong \mathrm{Pic}^0(\mathcal{X}_0(N)_k)$ for algebraically closed k of characteristic not dividing Np. It is gotten from the endomorphism of $\mathrm{Div}(\mathcal{X}_0(N)_k)$ which sends (E, C) to the divisor $\sum(E/D, (C+D)/D)$ where the sum is over cyclic subgroups D of E where D has order p and is not contained in C. The formula assumes that E is an elliptic curve, but it extends in a natural way to Néron polygons.

We would next like to extend $\mathcal{T}_p$ to an endomorphism of $J_{\mathcal{X}_0(N)/\mathbf{Z}[1/N]}$ and describe its reduction modulo p for a prime p not dividing N. More care is needed in this case since $X_0(Np)$ does not have a smooth and proper model over $\mathbf{Z}[1/N]$; the resulting description will be another manifestation of the Eichler-Shimura congruence relation.

We shall use the theory of Néron models for abelian varieties and begin by recalling some of the facts we need; see [**BLRa**, Chapter 1] or [**Artin**, §1]. Let R be a Dedekind domain and K its field of fractions. A smooth scheme $\mathcal{A}$ over R is said to have the Néron mapping property if for each smooth scheme $\mathcal{B}$ over R, the natural map $\mathrm{Hom}_R(\mathcal{B}, \mathcal{A}) \to \mathrm{Hom}_K(\mathcal{B}_K, \mathcal{A}_K)$ is a bijection. If A is an abelian scheme over K, then a smooth scheme $\mathcal{A}$ over R is called a Néron model for A if $\mathcal{A}$ has the Néron mapping property and there is an isomorphism $\phi : \mathcal{A}_K \to A$. The existence of such a model for A follows from the work of Néron. One checks formally that the pair $(\mathcal{A}, \phi)$ is unique up to canonical isomorphism, and also that $\mathcal{A}$ naturally inherits the structure of a commutative group scheme over R. If $\mathcal{A}$ is also proper over R, then $\mathcal{A}$ is an abelian scheme over R. Furthermore it is a consequence of a theorem of Weil that an abelian scheme over R has the Néron mapping property, so it is necessarily the Néron model of its generic fiber [**BLRa**, §1.2, Proposition 8].

EXAMPLE 10.2.3. Viewing $\mathcal{X}_0(11)_{\mathbf{Q}}$ as an elliptic curve over $\mathbf{Q}$ (see Example 9.3.2), we see that its Néron model over $\mathbf{Z}[1/11]$ is $\mathcal{X}_0(11)$, the elliptic curve of Example 8.1.1.

If A is an elliptic curve, then the possible types of reduction $\mathcal{A} \times_R R/\mathbf{m}$ of a Néron model $\mathcal{A}$ at a maximal ideal $\mathbf{m}$ of R are classified by Néron [**Neron**]; see also [**Sil2**, §IV.8] and [**BLRa**, §1.5].

EXAMPLE 10.2.4. For an example of a Néron model which is not an abelian scheme, let $R = \mathbf{C}[[q]]$ and consider $E_{q,R} = E_q \times_S \mathrm{Spec}\, R$ where E_q is the Tate curve over $S = \mathrm{Spec}\, \mathbf{Z}[[q]]$ (Example 9.2.1). Then the smooth commutative group scheme $\mathcal{A} = E_{q,R}^{\mathrm{reg}}$ turns out to be the Néron model for its generic fiber $\mathcal{A}_K = E_{q,K}$ where $K = \mathbf{C}((q))$. Recall from §9.2 that $\mathcal{A}_{R/qR}$ is isomorphic to $\mathbf{G}_{\mathbf{m}}$ over $R/qR \cong \mathbf{C}$.

Although the formation of a Néron model does not commute with arbitrary base change, it does commute with etale base change, as well as localization and completion at a maximal ideal of R [**BLRa**, §7.2]. In particular, if $\mathcal{A}$ is a Néron

model over $\mathbf{Z}$ for an abelian variety A over $\mathbf{Q}$, then $\mathcal{A}_{\mathbf{Z}[1/M]}$ is a Néron model over $\mathbf{Z}[1/M]$ for A, and $\mathcal{A}_{\mathbf{Z}_q}$ is a Néron model over $\mathbf{Z}_q$ for $A_{\mathbf{Q}_q}$.

EXAMPLE 10.2.5. For an example of the failure to commute with base change, let $R' = \operatorname{Spec} \mathbf{C}[[q^{1/2}]]$ and $K' = \operatorname{Spec} \mathbf{C}((q^{1/2}))$ and consider $E_{q,K'} = E_{q,K} \times_K K'$. By Hensel's lemma or the general theory of the Tate curve, we find that $E_{q,K'}(K')$ has a point of order 2 with coordinates (x, y) satisfying $x \equiv y \equiv 0 \bmod q^{1/2}R'$. As these points do not extend to elements of $\mathcal{A}(R')$ (with $\mathcal{A}$ as in Example 10.2.4), we see that $\mathcal{A}_{R'} = \mathcal{A} \times_R R'$ cannot be the Néron model of $E_{q,K'}$ over R'. Rather, in this case, the Néron model $\mathcal{A}'$ can be constructed by gluing two copies of $\mathcal{A}_{R'}$ along the automorphism of its generic fiber defined by translation by this point of order 2. The example also illustrates that the fibers of the Néron model need not be connected, for $\mathcal{A}' \times_{R'} (R'/q^{1/2}R')$ is isomorphic to the product of $\mathbf{G_m}$ with the constant group scheme of order 2. More generally, let $R' = k[[q^{1/p}]]$ and $K' = k((q^{1/p}))$ for a field k and a prime p and consider the generalized elliptic curve $\mathcal{E}'$ over R' defined as in Example 9.3.4. Then $\mathcal{A}' = (\mathcal{E}')^{\mathrm{reg}}$ is the Néron model of its generic fiber $E_{q,K'}$ and its reduction mod $q^{1/p}$ is isomorphic to $\mathbf{G_m} \times \mathbf{Z}/p\mathbf{Z}$.

EXAMPLE 10.2.6. Consider $J = J_{\mathcal{X}_0(27)_{\mathbf{Q}}/\mathbf{Q}}$ which is an elliptic curve over $\mathbf{Q}$ with conductor 27. Its minimal Weierstrass equation $Y^2 + Y = X^3 - 7$ produces a scheme $\overline{\mathcal{J}}$ over $\mathbf{Z}$ such that $\overline{\mathcal{J}}_{\mathbf{Z}[1/3]}$ an elliptic curve over $\mathbf{Z}[1/3]$, but $\overline{\mathcal{J}}_{\mathbf{F}_3}$ is not smooth. The smooth locus of $\overline{\mathcal{J}}$ is the identity component $\mathcal{J}^0$ of the Néron model $\mathcal{J}$, which is obtained by gluing three copies of $\mathcal{J}^0$ along translations of $\overline{\mathcal{J}}_{\mathbf{Z}[1/3]}$ by a point of order 3. We find that $\mathcal{J}_{\mathbf{F}_3}$ is isomorphic to the product of $\mathbf{G_a}$ with the constant group scheme of order 3. On the other hand, $J' = J_L$ extends to an elliptic curve $\mathcal{J}'$ over $\mathcal{O}_L$ where L is the Galois extension of $\mathbf{Q}$ gotten by adjoining the coordinates of all points of J of order 4. Therefore $\mathcal{J}'$ is the Néron model of J' over $\mathcal{O}_L$, so unlike Example 10.2.5, even formation of the identity component of the Néron model does not commute with base change.

Now let us return to the case of $J_0(N)$ or, to be more precise, its model $J_{\mathcal{X}_0(N)_{\mathbf{Q}}/\mathbf{Q}}$ over $\mathbf{Q}$. We let $\mathcal{J}_0(N)$ denote its Néron model over $\mathbf{Z}$. Then T_p extends uniquely to an endomorphism $\mathcal{T}_p$ of the Néron model. As $J_{\mathcal{X}_0(N)/\mathbf{Z}[1/N]}$ is an abelian scheme, it is naturally isomorphic to the Néron model $\mathcal{J}_0(N)_{\mathbf{Z}[1/N]}$ over $\mathbf{Z}[1/N]$. Moreover the endomorphism we have denoted $\mathcal{T}'_p$ is simply $\mathcal{T}_{p,\mathbf{Z}[1/Np]}$, so we have already described $\mathcal{T}_{p,\mathbf{F}_q}$ for primes q not dividing Np. Using for example the compatibility criterion in [**MaWi**, Section 2.1] we find that the description extends to $\mathcal{T}_{p,\mathbf{F}_p}$ on $\operatorname{Pic}^0(\mathcal{X}_0(N)_k)$ where k is an algebraic closure of $\mathbf{F}_p$ and p does not divide N. Namely it is given by the endomorphism of $\operatorname{Div}(\mathcal{X}_0(N)_k)$ which sends the pair (E, C) over k to

- $(E/D_0, (C + D_0)/D_0) + p \cdot (E/D_1, (C + D_1)/D_1)$ if E is ordinary, D_0 is the connected subgroup scheme of $E[p]$ and D_1 is the etale subgroup scheme of $E[p]$;
- $(p+1)\cdot(E/D, (C+D)/D)$ if E is supersingular and D is the unique subgroup scheme of order p.

(Again this assumes that E is an elliptic curve, but the description in the ordinary case can be extended to Néron polygons.) We thus find that

$$\mathcal{T}_{p,\mathbf{F}_p} = \Phi_* + \Phi^* \tag{10.2.1}$$

where Φ is the Frobenius endomorphism of the curve $\mathcal{X}_0(N)_{\mathbf{F}_p}$. This is the Eichler-Shimura congruence relation (see §8.5) on the Jacobian of $\mathcal{X}_0(N)_{\mathbf{F}_p}$; it can also be written as

$$\mathcal{T}_{p,\mathbf{F}_p} = \text{Frob} + \text{Ver} \tag{10.2.2}$$

where Frob is the Frobenius endomorphism of $\mathcal{J}_0(N)_{\mathbf{F}_p}$ and Ver is the Verschiebung endomorphism.

The situation is quite similar on the Jacobian $J_1(N)$ of $X_1(N)$. We can again consider the Néron model $\mathcal{J}_1(N)$ over $\mathbf{Z}$ of $J_{\mathcal{X}_1(N)_{\mathbf{Q}}/\mathbf{Q}}$ and define Hecke operators $\mathcal{T}_p$. Then $\mathcal{J}_1(N)_{\mathbf{Z}[1/N]}$ can be identified with the Jacobian of $\mathcal{X}_1(N)$ over $\mathbf{Z}[1/N]$ and $\mathcal{T}_{p,\mathbf{Z}[1/Np]}$ can be described as a composite $\alpha'_* \circ (\beta')^*$ where α' and β' are degeneracy maps from the curve $\mathcal{X}_{\mathbf{Z}[1/Np]}$. Recall from §9.3 that $\mathcal{X}$ is a model over $\mathbf{Z}[1/N]$ for the modular curve associated to the group $\Gamma_1(N,p) = \Gamma_1(N) \cap \Gamma_0(p)$. This gives a description of $\mathcal{T}_{p,\mathbf{Z}[1/N]}$ on divisors as $(E,P) \mapsto \sum_D (E/D, P \bmod D)$ which takes the form

$$\mathcal{T}_{p,\mathbf{F}_p} = \Phi_* + \langle p \rangle_{\mathbf{F}_p,*}\Phi^* = \text{Frob} + \langle p \rangle_{\mathbf{F}_p,*}\text{Ver} \tag{10.2.3}$$

in characteristic p if p does not divide N.

REMARK 10.2.7. As noted in Remark 10.2.1, some authors use T_p^* in terms of which (10.2.3) becomes

$$\mathcal{T}^*_{p,\mathbf{F}_p} = \Phi^* + \langle p \rangle^*_{\mathbf{F}_p}\Phi_* = \text{Ver} + \langle p \rangle^*_{\mathbf{F}_p}\text{Frob}.$$

VARIANT 10.2.8. Recall from Variant 9.3.6 the alternate model $\mathcal{X}_\mu(N)$ for $X_1(N)$. Then $J_{\mathcal{X}_\mu(N)_{\mathbf{Q}}/\mathbf{Q}}$ is a model over $\mathbf{Q}$ for $J_1(N)$ and we let $\mathcal{J}_\mu(N)$ denote its Néron model over $\mathbf{Z}$. Then T_p is defined over $\mathbf{Q}$ and we again write $\mathcal{T}_p$ for its extension to $\mathcal{J}_\mu(N)$. In this context the Eichler-Shimura relation is

$$\mathcal{T}_{p,\mathbf{F}_p} = \Phi^* + \langle p \rangle_{\mathbf{F}_p,*}\Phi_*; \qquad \mathcal{T}^*_{p,\mathbf{F}_p} = \Phi_* + \langle p \rangle^*_{\mathbf{F}_p}\Phi^*.$$

10.3. Bad reduction of Jacobians.

PRIMARY REFERENCES:

[**Rib4**, §2,3], [**BLRa**, Chapters 7,9] and [**DeRa**, Chapter V].

Now let us briefly discuss the structure of Jacobians of modular curves in some situations of bad reduction.

We first recall how some of the terminology used to describe the reduction of elliptic curves extends to the setting of abelian varieties [**BLRa**, §7.4]. If $\mathbf{m}$ is a maximal ideal of R, $\mathcal{A}$ is a Néron model over R and $\mathcal{A}_{R/\mathbf{m}}$ is an abelian scheme, then $A = \mathcal{A}_K$ is said to have *good reduction* at $\mathbf{m}$. If the identity component of $\mathcal{A}_{R/\mathbf{m}}$ is a torus, meaning that it is isomorphic over the algebraic closure of $R/\mathbf{m}$ to a product of copies of $\mathbf{G}_\mathbf{m}$, then A is said to have *multiplicative reduction* at $\mathbf{m}$. For example, the Tate curve over $k((q^{1/p}))$ (Example 10.2.5) has multiplicative reduction at the prime $q^{1/p}k[[q]]$ of $k[[q]]$. On the other hand, $J = J_{\mathcal{X}_0(27)_{\mathbf{Q}}/\mathbf{Q}}$ (Example 10.2.6) has neither good nor multiplicative reduction at $\mathbf{m} = 3\mathbf{Z}$; it is said to have *potentially good reduction* at $\mathbf{m}$ since $J' = J_L$ has good reduction at the primes lying over $\mathbf{m}$ in the integral closure of R in a finite Galois extension L of K.

Assume now that $N > 4$ and that p is a prime not dividing N. Recall from §9.3 that the model $\mathcal{X}$ for $X = \Gamma_1(N,p)\backslash\mathfrak{H}^*$ is not smooth over $\mathbf{Z}[1/N]$ but it is regular and the irreducible components of $\mathcal{X}_{\mathbf{F}_p}$ are smooth. The Néron model $\mathcal{J}$ over $\mathbf{Z}$ of $J_{\mathcal{X}_\mathbf{Q}/\mathbf{Q}}$ is naturally a model for the Jacobian of X and we can apply results of

Raynaud [**Ray2**] (see [**BLRa**, §9.5] and [**Rib4**, §2,3]) to describe $\mathcal{J}_{\mathbf{Z}_p}$ using the Picard functor. Raynaud proves that in such a situation the identity component of $\mathcal{J}_{\mathbf{Z}_p}$ is naturally isomorphic to $\mathrm{Pic}^0_{\mathcal{X}_{\mathbf{Z}_p}/\mathbf{Z}_p}$, the identity component of the algebraic space which represents $\mathrm{Pic}_{\mathcal{X}_{\mathbf{Z}_p}/\mathbf{Z}_p}$. Thus $\mathcal{J}^0_{\mathbf{F}_p}$ is isomorphic to $\mathrm{Pic}^0_{\mathcal{X}_{\mathbf{F}_p}/\mathbf{F}_p}$, which maps by Picard functoriality to the abelian scheme $\mathrm{Pic}^0_{(\mathcal{X}_{\mathbf{F}_p})^\sim/\mathbf{F}_p}$, where $\mathcal{X}^\sim_{\mathbf{F}_p}$ is the normalization of $\mathcal{X}_{\mathbf{F}_p}$. In fact there is an exact sequence of smooth group schemes over $\mathbf{F}_p$

$$1 \to T \to \mathcal{J}^0_{\mathbf{F}_p} \to B \to 1 \tag{10.3.1}$$

where $B = \mathrm{Pic}^0_{(\mathcal{X}_{\mathbf{F}_p})^\sim/\mathbf{F}_p}$ and T is a torus which can be described explicitly in terms of the singularities of $\mathcal{X}_{\mathbf{F}_p}$. Using the description of $(\mathcal{X}_{\mathbf{F}_p})^\sim$ provided by (9.3.1), we see that B is isomorphic to two copies of $\mathcal{J}_1(N)_{\mathbf{F}_p}$. The dimension of the torus T is $s-1$ where s is the number of supersingular points on $\mathcal{X}_1(N)_{\mathbf{F}_p}$ (see 9.3.2). We mention also that the component group $\mathcal{J}_{\mathbf{F}_p}/\mathcal{J}^0_{\mathbf{F}_p}$ can be computed as in [**Rib4**, §2] using a Picard-Lefschetz formula [**Gro2**, Exp.IX, §12] (see also [**Edi1**] and [**BLRa**, §9.6]).

Suppose that G is a smooth commutative group scheme over a field k and that there is an exact sequence $1 \to T \to G \to B \to 1$ where B is an abelian scheme and T is a torus. Then G is uniquely such an extension and is called a *semiabelian scheme* ([**BLRa**, §7.4]). Thus (10.3.1) shows that $\mathcal{J}^0_{\mathbf{F}_p}$ is a semiabelian scheme and we say that $\mathcal{J}_{\mathbf{Q}}$ has *semiabelian* (or *semistable*) *reduction* at p. The situation for $\mathcal{J}_0(Np)_{\mathbf{F}_p}$ with p not dividing N is quite similar to that for $\mathcal{J}_{\mathbf{F}_p}$, but slightly complicated by the fact that $\mathcal{J}_0(Np)$ may not be regular. One still finds that $\mathcal{J}_0(Np)_{\mathbf{Q}}$ has semiabelian reduction at p, now with B isomorphic to two copies of $\mathcal{J}_0(N)_{\mathbf{F}_p}$ and T described by the supersingular points of $\mathcal{X}_0(N)_{\mathbf{F}_p}$. For more details in this case we refer to [**Rib4**, §3] and D. Prasad's article in this volume.

It is a general fact that an abelian scheme A over K, the field of fractions of a Dedekind ring R, has "potentially semiabelian reduction" at all primes in R [**BLRa**, §7.4]. This means that there is a finite Galois extension K' of K such that $A_{K'}$ has semiabelian reduction at all primes in the integral closure of R in K'. Consider for example $\mathcal{J}_1(Np)_{\mathbf{Q}}$ with p not dividing N. This does not usually have semiabelian reduction at p, but it follows from the properties of the model for $X_1(Np)$ constructed by Deligne and Rapoport that $\mathcal{J}_1(Np)_{\mathbf{Q}(\zeta_p)}$ has semiabelian reduction at the prime over p, where ζ_p is a primitive pth root of unity. In fact, the results of [**DeRa**, §V.3] provide the following natural description of $J = J_{\mathcal{X}_1(Np)_{\mathbf{Q}}/\mathbf{Q}} = \mathcal{J}_1(Np)_{\mathbf{Q}}$ (which is valid without the hypothesis $N > 4$).

THEOREM 10.3.1. *Let π be the natural projection $\mathcal{X}_1(Np)_{\mathbf{Q}} \to \mathcal{X}_{\mathbf{Q}}$ and consider the filtration*

$$0 \subset A_1 \subset A_2 \subset J$$

where A_2 is the image of π^* and A_1 is the image of π^* composed with

$$\gamma = ((\alpha'_{\mathbf{Q}})^*, (\beta'_{\mathbf{Q}})^*) : \mathcal{J}_1(N)^2_{\mathbf{Q}} \to J.$$

Then A_1 has good reduction at p, A_2/A_1 has multiplicative reduction at p and J/A_2 acquires good reduction at the prime over p in $\mathbf{Q}(\zeta_p)$.

REMARK 10.3.2. That π^* and γ have finite kernels follows from the fact that the dual maps induce injections on cotangent spaces. A deeper result of Ribet [**Rib3**, Corollary 4.2] based on work of Ihara [**Ihara**] is that γ is actually injective.

In the case $N = 1$, the kernel of π^* is the Shimura subgroup of level p (see [**Maz1**, §II.11], [**LiOe**]).

EXAMPLE 10.3.3. Let $N = 11$ and $p = 3$. By Example 9.1.6, $X_1(33)$ has genus 21 and $X_1(11)$ has genus one. Appealing to the Hurwitz formula, we find also that $\Gamma_1(11,3)\backslash\mathfrak{H}^*$ has genus 11. Thus the dimensions of J, A_2 and A_1 are respectively 21, 11 and 2. Note that we can also interchange the roles of the primes 3 and 11 and define abelian subvarieties $A_1' \subset A_2' \subset J$. We find then that $A_1' = 0$ and A_2' coincides with the image of $\mathcal{J}_0(33)_{\mathbf{Q}} \to \mathcal{J}_1(33)_{\mathbf{Q}}$. Therefore A_2' is 3-dimensional and there is a filtration

$$0 \subset A_1 \subset A_2' \subset A_2 \subset J. \tag{10.3.2}$$

We conclude that

- A_1 is two-dimensional and has multiplicative reduction at 11 and good reduction at 3. In fact it is isogenous to two copies of $\mathcal{J}_0(11)_{\mathbf{Q}}$, an elliptic curve of conductor 11.
- A_2'/A_1 is one-dimensional and has multiplicative reduction at 3 and 11. It is an elliptic curve of conductor 33.
- A_2/A_2' is 8-dimensional, has multiplicative reduction at 3 and acquires good reduction at the prime over 11 in $\mathbf{Q}(\zeta_{11})$.
- J/A_2 is 10-dimensional and acquires everywhere good reduction over $\mathbf{Q}(\zeta_{33})$.

The description of suitable models for the curves $X_0(M)$ and $X_1(M)$, and consequently of the behavior of Néron models for $J_0(M)$ and $J_1(M)$, naturally becomes more complicated at a prime p when higher powers of p divide M. We will not pursue this here, but we refer the reader to [**KaMa**, Chapter 14] and [**MaWi**, Chapter 3] for more on the matter. We shall discuss in §12.5 the related problem of describing the natural Galois action on the Tate modules of these Jacobians.

Finally let us recover the Eichler-Shimura congruence relation from the description of $\mathcal{J}^0_{\mathbf{F}_p}$ given by (10.3.1). We first observe that the endomorphism $\mathcal{T}_p$ of $\mathcal{J}_1(N)$ is the composite of the homomorphisms of Néron models $\mathcal{J}_1(N) \to \mathcal{J}$ extending $(\beta')^*$ and $\mathcal{J} \to \mathcal{J}_1(N)$ extending α'_*. Thus $\mathcal{T}_{p,\mathbf{F}_p}$ can be computed as the composite

$$\mathcal{J}_1(N)_{\mathbf{F}_p} \to \mathcal{J}_{\mathbf{F}_p} \to \mathcal{J}_1(N)_{\mathbf{F}_p}.$$

Since $\mathcal{J}_1(N)_{\mathbf{F}_p}$ is an abelian scheme, the first map factors through the connected component of the identity, $\mathcal{J}^0_{\mathbf{F}_p}$. For the same reason, the second map restricted to $\mathcal{J}^0_{\mathbf{F}_p}$ factors through the projection in (10.3.1) to the abelian scheme $B = \mathrm{Pic}^0_{(\mathcal{X}_{\mathbf{F}_p})^\sim/\mathbf{F}_p}$. Using $i_F \coprod i_V$ to identify B with $\mathcal{J}_1(N)^2_{\mathbf{F}_p}$, we are reduced to computing the composite

$$\mathcal{J}_1(N)_{\mathbf{F}_p} \to \mathcal{J}_1(N)^2_{\mathbf{F}_p} \to \mathcal{J}_1(N)_{\mathbf{F}_p}.$$

The endomorphisms of $\mathcal{J}_1(N)_{\mathbf{F}_p}$ which come into play arise from the endomorphisms of $\mathcal{X}_1(N)_{\mathbf{F}_p}$ considered in the analysis of $\mathcal{X}_{\mathbf{F}_p}$ in §9.3. Indeed the first map arises by Picard functoriality from $\beta'_{\mathbf{F}_p}$ and sends a point x in $\mathcal{J}_1(N)_{\mathbf{F}_p}(S)$ to $(\Phi^*(x), x)$ in $\mathcal{J}_1(N)_{\mathbf{F}_p}(S)^2$. We have used here the extension of (8.4.1) to $\mathcal{X}_{\mathbf{F}_p}$ and the evident compatibility of Raynaud's description with Picard functoriality. Similarly using the compatibility with Albanese functoriality (see [**Ray3**]), we deduce that the second map, which arises from α', sends a point (y, z) to $\langle p\rangle_{\mathbf{F}_p,*}(y) + \Phi_*(z)$. Computing the composite we recover (10.2.3). The situation is similar for $\mathcal{J}_0(N)_{\mathbf{F}_p}$, but $\langle p\rangle$ is replaced by the identity.

Part III. Modular forms revisited

11. Automorphic representations

Let $\mathbf{A}$ be the ring of adeles of $\mathbf{Q}$. We will write $\mathbf{A}_\mathbf{f}$ for the ring of finite adeles. For each positive integer N, let U_N denote the open compact subgroup of $\mathbf{A}_\mathbf{f}^\times$ consisting of elements of $\hat{\mathbf{Z}}^\times = \prod_p \mathbf{Z}_p^\times$ which are congruent to 1 mod $N\hat{\mathbf{Z}}$.

Recall that a *Hecke character* is a continuous homomorphism $\mathbf{A}^\times \to \mathbf{C}^\times$ trivial on $\mathbf{Q}^\times$. To a Dirichlet character $\varepsilon : (\mathbf{Z}/N\mathbf{Z})^\times \to \mathbf{C}^\times$, we associate a Hecke character $\varepsilon_\mathbf{A}$ as follows. We write $\mathbf{A}^\times = \mathbf{Q}^\times \mathbf{R}_{>0}^\times \hat{\mathbf{Z}}^\times$, and let $\varepsilon_\mathbf{A}(\alpha x u) = \varepsilon(u^{-1} \bmod N)$ for $\alpha \in \mathbf{Q}^\times$, $x \in \mathbf{R}_{>0}^\times$ and $u \in \hat{\mathbf{Z}}^\times$. Recall that every Hecke character of finite order arises this way. More generally, every continuous quasi-character of $\mathbf{A}^\times/\mathbf{Q}^\times$ can be written as $\varepsilon_\mathbf{A}|\ |^s$ for some Dirichlet character ε and some $s \in \mathbf{C}$. We will usually omit the subscript $\mathbf{A}$. We will also use "character" to mean a continuous homomorphism to $\mathbf{C}^\times$ and call a character unitary if the values have norm 1.

In this section we discuss how modular forms can be regarded as functions on $\mathrm{GL}_2(\mathbf{A})$. These in turn give rise to (infinite-dimensional) automorphic representations of $\mathrm{GL}_2(\mathbf{A})$ which are, in some sense, generalizations of the Hecke characters of $\mathrm{GL}_1(\mathbf{A})$ we have just defined. We will also discuss how these automorphic representations are described in terms of local factors. The primary reference is Jacquet-Langlands [**JaLa**], but see also the expositions of their work by Godement [**Gode**] and Gelbart [**Gelb**].

Before proceeding, we give a word of motivation for this translation to the adelic language. Recall that it is the language in which class field theory most naturally describes abelian extensions of number fields. In the same spirit, Langlands' conjectures are expressed in this language, providing even deeper arithmetic information from the theory of modular forms.

11.1. The adelic setting.

PRIMARY REFERENCES:
[**Cas2**, §3], [**Gelb**, §3] and [**Cas1**, §1].

Write $G_\mathbf{Q}$, $G_\mathbf{A}$, G_∞ and $G_\mathbf{f}$, respectively, for $\mathrm{GL}_2(\mathbf{Q})$, $\mathrm{GL}_2(\mathbf{A})$, $\mathrm{GL}_2(\mathbf{R})$ and $\mathrm{GL}_2(\mathbf{A}_\mathbf{f})$. Put $\mathfrak{H}^\pm = \mathbf{C} - \mathbf{R}$. We let $U_\infty = \mathrm{SO}_2(\mathbf{R})\mathbf{R}^\times$, the stabilizer of $i = \sqrt{-1} \in \mathbf{C}$ in G_∞. We identify G_∞/U_∞ with $\mathfrak{H}^\pm$ by $g \mapsto gi$, and define $j : G_\infty \times \mathfrak{H}^+ \to \mathbf{C}$ by $j(\gamma, z) = cz + d$ where $\gamma = \left(\begin{smallmatrix} * & * \\ c & d \end{smallmatrix}\right)$. Let $\mathcal{S}_k$ be the space of functions $\phi : G_\mathbf{Q}\backslash G_\mathbf{A} \longrightarrow \mathbf{C}$ such that

(1) $\phi(gu) = \phi(g)$ for all u in some open compact subgroup U of $G_\mathbf{f}$;
(2) $\phi(gu_\infty) = j(u_\infty, i)^{-k}(\det u_\infty)\phi(g)$ for all $u_\infty \in U_\infty$, $g \in G_\mathbf{A}$;
(3) for all $g \in G_\mathbf{f}$ the map

$$\begin{array}{ccc} \mathfrak{H}^\pm & \to & \mathbf{C} \\ hi & \mapsto & \phi(gh)j(h,i)^k(\det h)^{-1}, \end{array}$$

where $h \in G_\infty$, is holomorphic (the map is well-defined by (2));
(4) ϕ is slowly increasing, i.e., for every $c > 0$ and every compact subset $K \subset G_\mathbf{A}$, there exist constants A, B such that

$$|\phi\left(\left(\begin{smallmatrix} a & 0 \\ 0 & 1 \end{smallmatrix}\right)h\right)| \leq A|a|^B$$

for all $h \in K$ and $a \in \mathbf{A}^\times$ with $|a| > c$;

(5) ϕ is cuspidal, i.e., for all $g \in G_{\mathbf{A}}$

$$\int_{\mathbf{Q}\backslash\mathbf{A}} \phi\left(\begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} g\right) dx = 0,$$

where dx is a non-trivial Haar measure.

We regard $\mathcal{S}_k$ as a $G_{\mathbf{f}}$-module where the action is given by right translation.

For an open compact subgroup U of $G_{\mathbf{f}}$, we write $\mathcal{S}_k(U)$ for the space of U-invariant functions in $\mathcal{S}_k$, i.e., those $\phi \in \mathcal{S}_k$ such that $\phi(gu) = \phi(g)$ for all $u \in U$, $g \in G_{\mathbf{A}}$. Note that $\mathcal{S}_k = \bigcup_U \mathcal{S}_k(U)$ over all such U.

For $N > 0$, let $U_0(N)$ (respectively, $U_1(N)$, V_N) be the subgroup of $\mathrm{GL}_2(\hat{\mathbf{Z}})$ consisting of matrices congruent to $\begin{pmatrix} * & * \\ 0 & * \end{pmatrix}$ (respectively, $\begin{pmatrix} * & * \\ 0 & 1 \end{pmatrix}$, the identity) modulo $NM_2(\hat{\mathbf{Z}})$. For an element ϕ of $\mathcal{S}_k(U_1(N))$, we define a function $f_\phi : \mathfrak{H} \to \mathbf{C}$ by

$$f(hi) = \phi(h)j(h,i)^k(\det h)^{-1} \qquad \text{for } h \in \mathrm{GL}_2^+(\mathbf{R}).$$

(See e.g. [**Cas2**, Theorem 3] or [**Gelb**, Proposition 3.1].) Then f_ϕ is in $\mathcal{S}_k(\Gamma_1(N))$ and $\phi \mapsto f_\phi$ in fact defines an isomorphism

$$\mathcal{S}_k(U_1(N)) \cong \mathcal{S}_k(\Gamma_1(N)). \tag{11.1.1}$$

Moreover for a mod N Dirichlet character ε, we find that $\mathcal{S}_k(N,\varepsilon)$ corresponds to the subspace of $\mathcal{S}_k(U_1(N))$ consisting of ϕ such that $u\phi = \varepsilon_{\mathbf{A}}(\det u)\phi$ for all $u \in U_0(N)$. In particular, $\mathcal{S}_k(U_0(N))$ corresponds to $\mathcal{S}_k(\Gamma_0(N))$.

REMARK 11.1.1. One can formulate the definition of a modular curve adelically as well. For an open compact subgroup U of $G_{\mathbf{f}}$, define

$$X_U = G_{\mathbf{Q}}\backslash G_{\mathbf{A}}/UU_\infty.$$

(Note that X_U need not be connected.) One then has a system of canonical models defined over $\mathbf{Q}$ [**Shi1**, §6.7], [**Del2**, §1,2] admitting a natural moduli-theoretic interpretation in terms of elliptic curves with level structure [**Del2**, §4,5], [**Mil1**, §2].

We also find that the Hecke action on the spaces $\mathcal{S}_k(U)$ has a very simple description. If U, U' are open compact subgroups in $G_{\mathbf{f}}$, then for $g \in G_{\mathbf{f}}$ we define the operator $[UgU'] : \mathcal{S}_k(U') \to \mathcal{S}_k(U)$ by

$$([UgU']\phi)(g) = \sum_i (h_i\phi)(g) = \sum_i \phi(gh_i) \tag{11.1.2}$$

where $UgU' = \coprod h_iU'$. Note that if $Ug_1U' = Ug_2U'$ as double cosets, then the operators coincide as well. To recover the classical Hecke operators from this, let $\varpi_q \in \mathbf{A}_{\mathbf{f}}^\times$ be the element such that $(\varpi_q)_v = q$ if $v = q$ and $(\varpi_q)_v = 1$ if $v \neq q$. Define endomorphisms of $\mathcal{S}_k(U)$ by

$$T_q = [U\eta_qU], \qquad S_q = [U\varpi_qU] \tag{11.1.3}$$

where $\eta_q = \begin{pmatrix} \varpi_q & 0 \\ 0 & 1 \end{pmatrix} \in G_{\mathbf{f}}$. For $U = U_1(N)$ (or $U_0(N)$), these are compatible under (11.1.1) with the operators denoted T_q and S_q on $\mathcal{S}_k(\Gamma_1(N))$ (or $\mathcal{S}_k(\Gamma_0(N))$) in §3.4; see [**Cas1**, Theorem 1.1] and the example following it.

We find also that if U contains V_N, then all the operators T_q and S_q commute for q not dividing N, thus making $\mathcal{S}_k(U)$ a $\mathbf{T}^{(N)}$-module. For each eigencharacter θ of $\mathbf{T} = \mathbf{T}^{(1)}$, we can form the union $\mathcal{S}_{k,\theta}$ of the eigenspaces in $\mathcal{S}_k(U)$ for the

restriction of θ to $\mathbf{T}^{(N)}$, where the union is over pairs (U, N) such that $V_N \subset U$. Then $\mathcal{S}_{k,\theta}$ is stable under the action of $G_{\mathbf{f}}$, and we write $\mathcal{S}_{k,\theta}(U)$ for $\mathcal{S}_{k,\theta} \cap \mathcal{S}_k(U)$.

THEOREM 11.1.2. *Let θ be a homomorphism $\mathbf{T} \to \mathbf{C}$ such that $\mathcal{S}_{k,\theta}$ is nontrivial. Then $\mathcal{S}_{k,\theta}$ is an irreducible $G_{\mathbf{f}}$-module, and there is a unique integer $N = N_\theta$ such that $\mathcal{S}_{k,\theta}(U_1(N))$ is one-dimensional. Conversely, each irreducible constituent of $\mathcal{S}_k$ is of the form $\mathcal{S}_{k,\theta}$ for some θ.*

We shall explain in §11.5 how the theorem is deduced from representation-theoretic results, but first we review some of the theory of irreducible admissible representations. We begin by describing the possible local "factors" of such a representation in the next two subsections and then deal with the global theory in §11.4. In §11.5 we relate the representations $\mathcal{S}_{k,\theta}$ to weight k cuspidal automorphic representations of $G_{\mathbf{A}}$ and apply a representation-theoretic multiplicity one theorem to obtain Theorem 11.1.2.

REMARK 11.1.3. Note that θ is only determined up to equivalence by $\mathcal{S}_{k,\theta}$, where θ_1 and θ_2 are deemed equivalent if their restriction to $\mathbf{T}^{(M)}$ coincides for some M. However we shall see that for each multiple M of N_θ, $\mathbf{T}^{(M)}$ acts via an eigencharacter on $\mathcal{S}_{k,\theta}(V_M)$.

REMARK 11.1.4. Recalling the theory of newforms §6.3, we see that the space $\mathcal{S}_{k,\theta}(U_1(N))$ of Theorem 11.1.2 is spanned by ϕ where f_ϕ is the newform of level N with $\mathbf{T}^{(N)}$-eigencharacter determined by the equivalence class of θ. We therefore have natural bijections among the following three sets:

1. equivalence classes of eigencharacters θ such that $\mathcal{S}_{k,\theta}$ is nontrivial;
2. irreducible constituents of $\mathcal{S}_k$;
3. newforms of weight k.

In fact, the theory of newforms can be recovered from the analysis of the structure of $\mathcal{S}_k$ as a $G_{\mathbf{f}}$-module provided by the theory of Jacquet and Langlands; see [**Cas2**, §3].

11.2. Admissible representations; p-adic case.

PRIMARY REFERENCES:
[**JaLa**, §2–4], [**Gode**, §1] and [**Gelb**, §4B].

Let p be a finite prime. In this subsection only, G denotes the group $\mathrm{GL}_2(\mathbf{Q}_p)$, K its standard maximal compact open subgroup $\mathrm{GL}_2(\mathbf{Z}_p)$ and Z the set of scalar matrices in G.

Let $\pi : G \to \mathrm{GL}(V)$ be a representation of G on a complex vector space V. The representation π is said to be *admissible* if (i) every vector $v \in V$ is fixed by some open subgroup of G, and (ii) for every open compact subgroup U of G, the subspace of vectors in V fixed by U is finite-dimensional. (See [**JaLa**, §2] or [**Gelb**, Definition 4.9].)

If U is an open compact subgroup of G, we write V^U for the subspace of vectors in V fixed by U. For open compact subgroups U and U' and an element g of G, we define the double coset operator $[UgU'] : V^{U'} \to V^U$ by

$$[UgU']\phi = \sum_i h_i \phi \tag{11.2.1}$$

where $UgU' = \coprod h_i U'$. (See (11.1.2).)

REMARK 11.2.1. A finite-dimensional admissible representation is continuous and the only continuous irreducible finite-dimensional representations of G are of the form $g \mapsto \omega(\det(g))$ where ω is a character of $\mathbf{Q}_p^\times$, [**JaLa**, Proposition 2.7].

The classification of irreducible infinite-dimensional admissible representations of G is carried out in [**JaLa**, §2,3]. (See also [**Gode**, §1.1–1.11].) We begin with certain induced representations defined as follows ([**JaLa**, (3.1)], [**Gelb**, (4.9)], [**Gode**, §1.8].) Let μ_1, μ_2 be any two characters of $\mathbf{Q}_p^\times$, and consider the space of all locally constant functions ϕ on G satisfying

$$\phi\left(\begin{pmatrix} a_1 & * \\ 0 & a_2 \end{pmatrix} g\right) = \mu_1(a_1)\mu_2(a_2)\left|\frac{a_1}{a_2}\right|^{1/2} \phi(g), \quad \forall a_1, a_2 \in \mathbf{Q}_p^\times ; \tag{11.2.2}$$

here $|\ |$ denotes the usual p-adic metric. The group G acts on the space by right translation, and this representation is denoted $\rho(\mu_1, \mu_2)$. The representation is reducible if and only if $\mu = \mu_1\mu_2^{-1} = |\ |^{\pm 1}$. (See [**JaLa**, Lemma 3.2.3], also [**Gode**, §1, Theorem 6] or [**Gelb**, Theorem 4.8]). When it is irreducible it is called a *principal series* representation.

If $\mu(x) = |x|^{-1}$, then $\rho(\mu_1, \mu_2)$ has a one-dimensional subrepresentation. Indeed putting $\omega = \mu_1 |\ |^{1/2} = \mu_2 |\ |^{-1/2}$ we see that the function $g \mapsto \omega(\det(g))$ spans a subspace stable under G. If $\mu(x) = |x|$, then there is a one-dimensional quotient, and in either of these cases the infinite-dimensional subquotient of $\rho(\mu_1, \mu_2)$ is irreducible, [**JaLa**, Lemma 3.2.3]. This subquotient is called the *special* or *Steinberg representation*, and is sometimes denoted $\mathrm{sp}(\mu_1, \mu_2)$.

In all of the above cases we let $\pi(\mu_1, \mu_2)$ denote the unique infinite dimensional irreducible subquotient of $\rho(\mu_1, \mu_2)$. Then $\pi(\mu_1, \mu_2)$ and $\pi(\mu_1', \mu_2')$ are equivalent if and only if $\{\mu_1, \mu_2\} = \{\mu_1', \mu_2'\}$. (See [**Gode**, §1, Theorem 4.7], also [**Gelb**, Remark 4.19]).

The admissible representations of G which are not of the form $\pi(\mu_1, \mu_2)$ are called *supercuspidal*. (See [**JaLa**, Proposition 2.17], [**Gode**, §1, Theorems 3,4].) These are characterized by the property that for all $v \in V$ and all ψ in $\check{V}$ of V, the functions $g \mapsto \psi(\pi(g)v)$, called *matrix coefficients*, are compactly supported modulo the centre Z. Here $\check{V}$ denotes the *admissible dual* of V, the space of linear functionals $\psi : V \to \mathbf{C}$ invariant under some open compact subgroup.

We also note that any irreducible admissible representation of G defines (by Schur's lemma) a character of the centre Z of G, called the *central character* of π. We denote by ω_π the corresponding character of $\mathbf{Q}_p^\times \cong Z$. For example, if $\pi = \pi(\mu_1, \mu_2)$, then $\omega_\pi = \mu_1\mu_2$.

We sometimes further restrict our attention to *unitarizible* representations, i.e., the admissible representations on which there is a G-invariant positive-definite Hermitian form. The irreducible ones are precisely (see [**Gode**])

- Principal series $\pi(\mu_1, \mu_2)$ with μ_1 and μ_2 unitary (called *continuous series*).
- Principal series $\pi(\mu, \bar{\mu}^{-1})$ with $\mu\bar{\mu} = |x|^\sigma$ for some real σ with $0 < |\sigma| < 1$ (called *complementary series*).
- Special or supercuspidal representations with unitary central character.

EXAMPLE 11.2.2. The unitarizible special representations are those of the form $\mathrm{sp}(\chi |\ |^{1/2}, \chi |\ |^{-1/2})$ with χ a unitary character.

REMARK 11.2.3. Special and supercuspidal representations are said to belong to the *discrete series*. Unitarizible discrete series representations are *square integrable* in the sense that their matrix coefficients are square integrable modulo the centre [**JaLa**, Lemma 15.2], [**Roga**, Proposition 2.6].

REMARK 11.2.4. For unitarizible π, we can form the completion of V with respect to the norm defined by the positive-definite Hermitian form (see [**Gode**, §1.17]). This determines a unitary representation $\hat{\pi}$ of G on a Hilbert space $\hat{V}$ from which π can be recovered as the representation on K-finite vectors in $\hat{V}$. (A vector v is called *K-finite* if the span of $\hat{\pi}(K)v$ is finite-dimensional, or equivalently if v is fixed by some open compact subgroup of G.) Moreover π is irreducible if and only if $\hat{\pi}$ is topologically irreducible, [**Gode**, §1, Lemma 10]. We remark also that every topologically irreducible unitary representation of G arises in this way, i.e., as the completion of an irreducible unitarizible representation [**Cart**, Corollary 2.3].

The *conductor* $\mathfrak{c}_\pi$ of an infinite-dimensional irreducible admissible representation π is defined to be the largest ideal $\mathfrak{c}$ of $\mathbf{Z}_p$ such that $V^{U_1(\mathfrak{c})} \neq 0$, where

$$U_1(\mathfrak{c}) = \{\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in K \mid c, d-1 \in \mathfrak{c}\}. \tag{11.2.3}$$

For $\mathfrak{c} = \mathfrak{c}_\pi$ this space of fixed vectors is in fact one-dimensional. (See [**Cas2**, Theorem 1]; the reader can check that our definition of the conductor is equivalent to the one given by Casselman.) Note that $\mathfrak{c}_\pi$ is divisible by the conductor of ω_π, and that v is fixed by $U_1(\mathfrak{c})$ if and only if $\pi(g)v = \omega_\pi(d)v$ for all $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in U_0(\mathfrak{c})$, where $U_0(\mathfrak{c})$ is defined as the group of matrices in K with $c \in \mathfrak{c}$. Note also that $\mathfrak{c}_\pi$ is determined by the restriction of π to K. We have the following list of possible conductors (see the proof of Theorem 1 in [**Cas2**]; [**Gelb**, Remark 4.25]):

- If $\pi = \pi(\mu_1, \mu_2)$ is principal series then $\mathfrak{c}_\pi = \mathfrak{f}_1\mathfrak{f}_2$ where $\mathfrak{f}_i$ $(i = 1, 2)$ denotes the conductor of μ_i.
- If $\pi = \mathrm{sp}(\chi|\ |^{1/2}, \chi|\ |^{-1/2})$ is special, then $\mathfrak{c}_\pi = \mathfrak{f}^2 \cap p\mathbf{Z}_p$, where $\mathfrak{f}$ is the conductor of χ;
- If π is supercuspidal then $\mathfrak{c}_\pi = p^n\mathbf{Z}_p$ for some $n \geq 2$.

EXAMPLE 11.2.5. An infinite-dimensional irreducible admissible representation π of G is called *unramified* (or *class 1* or *spherical*) if $\mathfrak{c}_\pi = \mathbf{Z}_p$, or equivalently if the subspace V^K of V fixed by $\pi(K)$ is one-dimensional. (See [**Gelb**, §4.B.3].) The unramified representations play an important role in global theory; see section 11.4. Note that according to the list above, these are precisely the principal series representations $\pi(\mu_1, \mu_2)$ for unramified characters μ_1, μ_2 (with $\mu_1\mu_2^{-1} \neq |\ |^{\pm 1}$). We find that V^K is spanned by the function on G defined by

$$\phi_0(\begin{pmatrix} a & b \\ 0 & d \end{pmatrix}k) = \mu_1(a)\mu_2(d)\left|\frac{a}{d}\right|^{1/2}, \quad k \in K.$$

Applying the double coset operators

$$T_p = K\begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix}K; \qquad S_p = K\begin{pmatrix} p & 0 \\ 0 & p \end{pmatrix}K$$

to ϕ_0, we find

$$T_p\phi_0 = p^{1/2}(\mu_1(p) + \mu_2(p))\phi_0; \qquad pS_p\phi_0 = p\mu_1(p)\mu_2(p)\phi_0. \tag{11.2.4}$$

So the characters μ_1 and μ_2 are determined by the eigenvalues of T_p and S_p on a nonzero vector fixed by K; therefore so is the isomorphism class of the unramified representation $\pi(\mu_1, \mu_2)$.

11.3. Admissible representations; real case.

PRIMARY REFERENCES:

[**JaLa**, §5], [**Gode**, §2], [**Gelb**, §4A], [**Wal2**, Chapters 3,5] and [**Wal1**, §2, 8].

In this subsection, we put $G = G_\infty = \mathrm{GL}_2(\mathbf{R})$, denote by K its maximal compact subgroup $O_2(\mathbf{R})$, and let $\mathfrak{g}$ be the complexification $\mathfrak{gl}_2(\mathbf{C})$ of the Lie algebra of G.

Let π be a unitary representation of G on a Hilbert space V such that the map $G \times V \to V$ is continuous. Let V_0 be the subspace of K-finite vectors in V (see Remark 11.2.4). Though V_0 is stable under K it is not necessarily stable under G, unlike the p-adic case. We assume that $\mathrm{Hom}_K(W, V_0)$ is finite-dimensional for every irreducible $\rho : K \to \mathrm{GL}(W)$.

REMARK 11.3.1. By a theorem of Harish-Chandra [**HaCh**] (see [**Wal2**, Chapter 3] and [**Gode**, §2.1]), this holds if π is topologically irreducible.

REMARK 11.3.2. Under our assumption, the vectors in V_0 are smooth in the sense of [**Wal2**, 1.6.6]; see [**Gode**, §2.1], [**Wal1**, Theorem 2.8].

To such a π one can associate, essentially by differentiation, a representation of the Lie algebra $\mathfrak{g}$. For X in the Lie algebra of G and $v \in V_0$, the derivative

$$\frac{d}{dt}\pi(\exp tX)v|_{t=0} = \lim_{t\to 0} t^{-1}(\pi(\exp tX)v - v) \tag{11.3.1}$$

exists and defines an element of V_0. (See [**Wal2**, 1.6.3], [**Gelb**, (4.5)].) Extending linearly to $\mathfrak{g}$ we obtain the desired homomorphism of complex Lie algebras

$$d\pi : \mathfrak{g} \to \mathfrak{gl}(V_0).$$

We denote by π_0 the pair of representations $d\pi$ and $\pi|_K$ on V_0; this pair satisfies certain continuity and compatibility conditions making V_0 a $(\mathfrak{g}, K)$-module. (See [**Wal2**, §3.3], for example, for the definition of a $(\mathfrak{g}, K)$-module.)

A $(\mathfrak{g}, K)$-module M is *admissible* if $\mathrm{Hom}_K(W, M)$ is finite-dimensional for every irreducible $\rho : K \to \mathrm{GL}(W)$; thus V_0 is automatically admissible. There are natural notions of homomorphisms and irreducibility for $(\mathfrak{g}, K)$-modules. We say that an admissible $(\mathfrak{g}, K)$-module is *unitarizible* if it is isomorphic to V_0 for some unitary π as above. The association of π_0 to π is evidently functorial, but we have moreover the following theorem of Harish-Chandra (see [**Wal1**, §2], [**Wal2**, Theorem 3.4.11]).

THEOREM 11.3.3. *Let $\pi : G \to \mathrm{GL}(V)$ and $\pi' : G \to \mathrm{GL}(V')$ be unitary representations as above. Then V is topologically irreducible if and only if V_0 is an irreducible $(\mathfrak{g}, K)$-module; in that case, V is isomorphic to V' as topological G-modules if and only if V_0 is isomorphic to V_0' as $(\mathfrak{g}, K)$-modules.*

Thus according to the theorem and Remark 11.3.1, the classification of irreducible unitary representations of G is equivalent to that of irreducible, unitarizible, admissible $(\mathfrak{g}, K)$-modules (cf. Remark 11.2.1). We have thus shifted our attention to $(\mathfrak{g}, K)$-modules from representations of G.

REMARK 11.3.4. Here we have strayed somewhat from the formulation of Jacquet-Langlands [**JaLa**, §5], where the focus is instead shifted to the classification of irreducible, admissible representations of a certain algebra $\mathcal{H}_\mathbf{R}$ called the Hecke algebra of G. See also [**Gode**, §2, (9)] and [**Gelb**, Definition 4.1] for a variant; there $\mathcal{H}_\mathbf{R}$ is defined as the algebra

$$\mathcal{H}_\mathbf{R} = \mathcal{U}(\mathfrak{g}) \bigoplus (\epsilon_-) * \mathcal{U}(\mathfrak{g}),$$

under convolution product of distributions, where $\mathcal{U}(\mathfrak{g})$ denotes the universal enveloping algebra of $\mathfrak{g}$ and ϵ_- is the Dirac measure at the point $\left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ of G. (See also [**Flath**, §3] and [**Cas3**, §1.2.1].)

Recall that a character $\varepsilon : \mathbf{R}^\times \to \mathbf{C}^\times$ has the form $\varepsilon(t) = |t|^s \mathrm{sgn}(t)^m$ for some $s \in \mathbf{C}$ and $m \in \mathbf{Z}/2\mathbf{Z}$. We say that ε is the *central character* of a $(\mathfrak{g}, K)$-module if $\{\pm 1\} = K \cap \mathbf{R}^\times$ acts via sgn^m and the center of $\mathfrak{g}$ acts via multiplication by s (where we have identified $\mathbf{R}^\times$ with the center of G and $\mathbf{C}$ with the center of $\mathfrak{g}$ in the obvious way). An analogue of Schur's lemma [**Wal2**, Lemma 3.3.2] shows that every irreducible $(\mathfrak{g}, K)$-module has a central character. Note also that if ε is the central character of π, then it is also the central character of the associated $(\mathfrak{g}, K)$-module.

Now we recall the classification of irreducible admissible $(\mathfrak{g}, K)$-modules. Analogous results in the context of admissible representations of the Hecke algebra (see Remark 11.3.4) are given in [**JaLa**, Theorem 5.11], [**Gode**, §2, Theorem 2] and [**Gelb**, Theorems 4.4, 4.5]. The version given here can be deduced from the Langlands classification [**Wal2**, Theorem 5.4.4], [**Wal1**, §8.4].

Let μ_1, μ_2 be two characters of $\mathbf{R}^\times$. As in [**JaLa**, §5] (also, [**Gode**, §2, (14)], [**Gelb**, (4.2)]), consider the space $\mathcal{B} = \mathcal{B}_{\mu_1,\mu_2}$ of all functions ϕ on G satisfying

$$\phi(\left(\begin{smallmatrix} t_1 & * \\ 0 & t_2 \end{smallmatrix}\right) g) = \mu_1(t_1)\mu_2(t_2)\left|\frac{t_1}{t_2}\right|^{1/2} \phi(g)$$

for all $g \in G$, t_1, $t_2 \in \mathbf{R}^\times$ and which are right K-finite (i.e., the functions $g \mapsto \phi(gk)$, $k \in K$, generate a finite dimensional space). The action of K is by right translation and that of $\mathfrak{g}$ is defined as in (11.3.1). Note that the central character of $\mathcal{B}$ is $\mu_1\mu_2$. Let μ be the character $\mu_1\mu_2^{-1}$ and let $\eta(t) = \mathrm{sgn}(t)$.

- The $(\mathfrak{g}, K)$-module $\mathcal{B}$ is irreducible unless $\mu(t) = t^n\eta(t)$ for some nonzero integer n.
- If $\mu(t) = t^n\eta(t)$ for some integer $n > 0$, then $\mathcal{B}$ contains exactly one proper $(\mathfrak{g}, K)$-submodule $\mathcal{B}^s$. It is infinite dimensional; the quotient $\mathcal{B}^f = \mathcal{B}/\mathcal{B}^s$ has dimension n.
- If $\mu(t) = t^n\eta(t)$ for some integer $n < 0$, then $\mathcal{B}$ contains exactly one proper $(\mathfrak{g}, K)$-submodule $\mathcal{B}^f$. It is $|n|$-dimensional; the quotient $\mathcal{B}^s = \mathcal{B}/\mathcal{B}^f$ is infinite dimensional.

Let us denote by $\pi(\mu_1, \mu_2)$ the $(\mathfrak{g}, K)$-module $\mathcal{B}_{\mu_1,\mu_2}$ if it is irreducible, but the finite-dimensional $\mathcal{B}^f_{\mu_1,\mu_2}$ otherwise. In the latter case, the infinite-dimensional $\mathcal{B}^s_{\mu_1,\mu_2}$ is denoted $\sigma(\mu_1, \mu_2)$; it is defined only if $\mu(t) = t^n\eta(t)$ for some nonzero integer n.

The $(\mathfrak{g}, K)$-modules $\pi(\mu_1, \mu_2)$ make up the *principal series* for G, the terminology often being reserved for the case where $\mu(t)$ is not of the form $t^n\eta(t)$ with $n \in \mathbf{Z}$. The $\sigma(\mu_1, \mu_2)$ are called *discrete series*; $\pi(\mu_1, \mu_2)$ is called a *limit of discrete series* if $\mu = \eta$.

Every irreducible admissible $(\mathfrak{g}, K)$-module is isomorphic to either $\pi(\mu_1, \mu_2)$ or $\sigma(\mu_1, \mu_2)$ for some characters μ_1 and μ_2. Moreover, the only equivalences among them are the following

- $\pi(\mu_1, \mu_2) \simeq \pi(\mu_1', \mu_2')$ if $\{\mu_1, \mu_2\} = \{\mu_1', \mu_2'\}$;
- $\sigma(\mu_1, \mu_2) \simeq \sigma(\mu_1', \mu_2')$ if $\{\mu_1, \mu_2\} = \{\mu_1', \mu_2'\}$ or $\{\mu_1'\eta, \mu_2'\eta\}$.

Next we list those which are unitarizible ([**Gelb**, Remark 4.7], [**Wal1**, §8.7], [**Kna1**]):

- Principal series $\pi(\mu_1, \mu_2)$ with μ_1 and μ_2 unitary (*continuous series*).
- Principal series $\pi(\mu, \bar{\mu}^{-1})$ with $\mu\bar{\mu} = |x|^\sigma$ for some real σ with $0 < |\sigma| < 1$ (*complementary series*).
- Discrete series $\sigma(\mu_1, \mu_2)$ with unitary central character.

REMARK 11.3.5. Each $(\mathfrak{g}, K)$-module in the above list arises from an irreducible unitary representation π of G. Moreover π is unique up to isomorphism, and its central character is $\mu_1\mu_2$. Among these, it is precisely the unitarizible discrete series which arise from square-integrable π (see e.g. [**Kna1**, §1], and also Remark 11.2.3).

Finally, we distinguish the $(\mathfrak{g}, K)$-modules which arise in the consideration of cusps forms of weight k for $k \geq 2$; see [**Roga**, Proposition 2.5]. Let

$$\sigma_k = \sigma(|\ |^{(k-1)/2}, |\ |^{(1-k)/2}\eta^k). \tag{11.3.2}$$

These are precisely the unitarizible discrete series with central character 1 or η. There is a nonzero subspace of σ_k consisting of vectors v satisfying

$$\begin{aligned} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} v &= e^{ik\theta} v \quad \text{for all } \theta \in \mathbf{R}; \\ \begin{pmatrix} 1 & -i \\ -i & -1 \end{pmatrix} v &= 0 \end{aligned} \tag{11.3.3}$$

(where the first matrix is in K and the second in $\mathfrak{g}$). The subspace is unique and its existence characterizes σ_k among the irreducible admissible $(\mathfrak{g}, K)$-modules. We fix such a nonzero vector v_k, called a *lowest weight vector*.

11.4. The global theory.

PRIMARY REFERENCES:

[**JaLa**, §9], [**Gode**, §3.2], [**Gelb**, §4C] and [**Flath**].

Using the notion of a restricted tensor product enables us to describe global admissible representations in terms of a factorization into local components. The procedure is analogous to the description of a Hecke character in terms of its local components.

Suppose that we are given, for each finite prime p, an irreducible admissible representation $\pi_p : G_p \to \mathrm{GL}(V_p)$ where $G_p = \mathrm{GL}_2(\mathbf{Q}_p)$. (Here V_p could be one or infinite-dimensional.) Suppose also that π_p is unramified for all p not in a finite set S. For each $p \notin S$, choose a non-zero vector e_p in the one-dimensional subspace of K_p-fixed vectors in V_p, where $K_p = \mathrm{GL}_2(\mathbf{Z}_p)$. Let W be the linear span of elements of the form $\bigotimes_p v_p$ such that $v_p = e_p$ for all but finitely many p. Then we define the action of $G_{\mathbf{f}}$ componentwise on such elements and extend the action linearly to W. This yields an irreducible representation $G \to \mathrm{GL}(W)$ which is called the *restricted tensor product* of the π_p and is denoted $\bigotimes \pi_p$ (see [**JaLa**, §9], [**Flath**, §2]). Up to isomorphism, $\bigotimes \pi_p$ is independent of the choice of $\{e_p\}$. Moreover the $G_{\mathbf{f}}$-module W is admissible in the sense that (i) every vector in W is fixed by some open subgroup of $G_{\mathbf{f}}$, and (ii) for every open compact subgroup U of $G_{\mathbf{f}}$, the subspace of vectors in W fixed by U is finite-dimensional.

Suppose that we are also given an irreducible admissible $(\mathfrak{g}, K_\infty)$-module V_∞, where $\mathfrak{g} = \mathfrak{gl}_2(\mathbf{C})$ and $K_\infty = O_2(\mathbf{R})$. We can then consider $V = V_\infty \otimes W$; it is a $(\mathfrak{g}, K_\infty) \times G_{\mathbf{f}}$-module, by which we simply mean that it is compatibly a $(\mathfrak{g}, K_\infty)$-module and a $G_{\mathbf{f}}$-module. It is irreducible in the sense that it has no proper

$(\mathfrak{g}, K_\infty) \times G_{\mathbf{f}}$-submodules. Moreover V is admissible in the sense that (i) every vector in V is fixed by some open subgroup of $G_{\mathbf{f}}$, and (ii) for every open compact subgroup U of $G_{\mathbf{f}}$, the subspace of vectors in V fixed by U is an admissible $(\mathfrak{g}, K_\infty)$-module.

REMARK 11.4.1. If we suppose further that each π_p is unitarizible, then we could require that e_p be a unit vector for $p \notin S$. The resulting $G_{\mathbf{f}}$-module is then endowed with an invariant positive-definite Hermitian form. We can then form the Hilbert space completion $\hat{W}$ of W and obtain a unitary representation of $G_{\mathbf{f}}$. If V_∞ is also unitarizible, then the construction yields an admissible unitary representation of $G_{\mathbf{A}}$ from which V is recovered as a dense $(\mathfrak{g}, K_\infty) \times G_{\mathbf{f}}$-submodule satisfying certain finiteness conditions (see [**Gode**, §3.3]).

Conversely every irreducible, admissible $(\mathfrak{g}, K_\infty) \times G_{\mathbf{f}}$-module can be written as a restricted tensor product; moreover the local factors V_p and V_∞ are unique up to isomorphism. (This follows from results of [**JaLa**, §9]; see also [**Flath**, Theorem 3], [**Gode**, §3.2] and [**GGPS**].)

In particular, suppose that we are given an irreducible, admissible representation $\pi : G_{\mathbf{f}} \to \mathrm{GL}(W)$. Then π is isomorphic to $\bigotimes \pi_p$ for a collection of local representations $\pi_p : G_p \to \mathrm{GL}(V_p)$. Note also that π has a central character which factors into the product of the central characters of the local representations.

Suppose now that each π_p is infinite-dimensional. We then define the *conductor* of π to be $N_\pi = \prod_p p^{n_p}$ where for each p, the conductor of π_p is $p^{n_p}\mathbf{Z}_p$. Observe that then N_π is the least positive integer N such that there is a nonzero vector in W fixed by $U_1(N)$; moreover the space of such vectors is one-dimensional.

Given a character $\varepsilon : \mathbf{A}_{\mathbf{f}}^\times \to \mathbf{C}^\times$ and a $G_{\mathbf{f}}$-module W, we let $W(\varepsilon \circ \det)$ denote its twist by $\varepsilon \circ \det$, i.e., the $G_{\mathbf{f}}$-module $W \otimes M$ where M is the one-dimensional $G_{\mathbf{f}}$-module gotten from the representation $\varepsilon \circ \det$. Then W is admissible (respectively, irreducible) if and only if $W(\varepsilon \circ \det)$ is admissible (respectively, irreducible). If W is the restricted tensor product formed from local representations $\pi_p : G_p \to \mathrm{GL}(V_p)$, then $V(\varepsilon \circ \det)$ is formed from the representations $\pi_p \otimes (\varepsilon_p \circ \det)$.

11.5. Cuspidal representations.

PRIMARY REFERENCES:
[**JaLa**, §10], [**Gode**, §3], [**Gelb**, §5,6] and [**Cas2**].

Let X denote the space $\mathbf{R}_{>0}^\times G_{\mathbf{Q}} \backslash G_{\mathbf{A}}$ where we regard $\mathbf{R}_{>0}^\times$ as contained in the scalar matrices of G_∞. We let dx be a $G_{\mathbf{A}}$-invariant measure on X and we consider the Hilbert space $L^2(X)$. Recall that this is the space of (equivalence classes of) measurable functions $\phi : X \to \mathbf{C}$ such that

$$\int_X |\phi(x)|^2 \, dx < \infty;$$

it is endowed with the inner product

$$\langle \phi_1, \phi_2 \rangle = \int_X \phi_1(x) \bar{\phi}_2(x) \, dx.$$

Then $G_{\mathbf{A}}$ acts on $L^2(X)$ unitarily by right translation.

We shall consider here only the subspace $L_0^2(X)$ consisting of ϕ satisfying a certain cuspidality condition. (We are thus ignoring the contribution of Eisenstein

series; see [**Gelb**, §8] for example.) We let $L_0^2(X)$ denote the set of ϕ such that the function

$$g \mapsto \int_{\mathbf{Q}\backslash\mathbf{A}} \phi(\begin{pmatrix} 1 & y \\ 0 & 1 \end{pmatrix} g)\, dy$$

vanishes almost everywhere on $G_{\mathbf{A}}$. Then $L_0^2(X)$ is a closed subspace stable under the action of $G_{\mathbf{A}}$. It decomposes into a Hilbert space direct sum

$$L_0^2(X) = \hat{\bigoplus} R_\alpha, \tag{11.5.1}$$

where the sum is over (a countable set of) closed irreducible subspaces stable under the action of $G_{\mathbf{A}}$. The isomorphism classes of unitary representations $G_{\mathbf{A}} \to \mathrm{GL}(R_\alpha)$ which arise in this way are called *cuspidal automorphic representations*. The existence of such a decomposition can be established by a method found in [**GGPS**] (see also [**Gode**, §3.1] and [**GeJa**, §2]), along with the fact that each isomorphism class occurs with only finite multiplicity. Using the existence and uniqueness of Whittaker models, Jacquet-Langlands [**JaLa**, §10,11] prove that the multiplicities are one.

The theory developed in [**JaLa**] in fact yields a *strong multiplicity one* theorem which we state below. To do so, we first switch to the context of admissible $(\mathfrak{g}, K_\infty) \times G_{\mathbf{f}}$-modules by further restricting our attention to the space of *cuspidal automorphic forms*, denoted $\mathcal{A}_0$. This is the space of smooth, K-finite, $\mathfrak{z}$-finite, slowly increasing functions in $L_0^2(X)$. Here "smooth" is as a function of G_∞, K is the maximal compact $K_\infty \mathrm{GL}_2(\hat{\mathbf{Z}})$ and $\mathfrak{z}$ is the center of $\mathcal{U}(\mathfrak{g})$ (Remark 11.3.4); we have already discussed the notions of "finite" (Remark 11.4.1) and "slowly increasing" (§11.1). Note that our $\mathcal{A}_0$ is the algebraic direct sum over finite order Hecke characters ε of the spaces denoted $\mathcal{A}_0(\varepsilon)$ in [**Gelb**, Definition 3.3].

The space $\mathcal{A}_0$ is a dense subspace of L_0^2 and an admissible $(\mathfrak{g}, K_\infty) \times G_{\mathbf{f}}$-module. It decomposes into an algebraic direct sum

$$\mathcal{A}_0 \simeq \bigoplus V_\alpha \tag{11.5.2}$$

where for each α, V_α is an irreducible admissible $(\mathfrak{g}, K_\infty) \times G_{\mathbf{f}}$-module dense in the space R_α occurring in (11.5.1); see [**Gelb**, Theorem 5.1]. Now factor each V_α as explained in §11.4 and denote the corresponding admissible G_p-modules (respectively $(\mathfrak{g}, K_\infty)$-module) by $V_{\alpha,p}$ (respectively, $V_{\alpha,\infty}$). We can now state the strong multiplicity one theorem as follows. (See [**Gelb**, §6], [**Cas2**, Theorem 2] and [**PSh2**].)

THEOREM 11.5.1. *Suppose V_α and V_β are constituents of $\mathcal{A}_0$ such that $V_{\alpha,p} \cong V_{\beta,p}$ as G_p-modules for all but finitely many primes p. Then $V_\alpha = V_\beta$.*

Note that the theorem asserts not only that V_α and V_β are isomorphic, but that they coincide as subspaces of $\mathcal{A}_0$.

The theorem also incorporates results about non-holomorphic automorphic forms called Maass forms [**Maass**], but we will content ourselves with a discussion of the transition back to the setting of §11.1 and the theory of newforms. (See [**Cas2**, §3] and [**Gelb**, §5].)

Recall that for each $k \geq 2$, we distinguished a unitarizible discrete series $(\mathfrak{g}, K_\infty)$-module denoted σ_k. Let

$$\mathcal{A}_{0,k} = \mathrm{Hom}_{(\mathfrak{g},K_\infty)}(\sigma_k, \mathcal{A}_0).$$

Then $\mathcal{A}_{0,k}$ is an admissible $G_{\mathbf{f}}$-module and decomposes as the direct sum of

$$V_{\alpha,\mathbf{f}} = \mathrm{Hom}_{(\mathfrak{g},K_\infty)}(\sigma_k, V_\alpha),$$

over α such that $V_{\alpha,\infty}$ is isomorphic to σ_k. By Schur's Lemma, $V_{\alpha,\mathbf{f}}$ is an irreducible admissible $G_{\mathbf{f}}$-module isomorphic to the restricted tensor product of the $V_{\alpha,p}$. The constituents in the decomposition of $\mathcal{A}_{0,k}$ correspond to the irreducible constituents of $\mathcal{S}_k$, but the local factors $V_{\alpha,p}$ are unitarizible whereas the factors of $\mathcal{S}_k$ are not. To complete the transition, we twist $\mathcal{A}_{0,k}$ by a Hecke character and define an isomorphism

$$\mathcal{A}_{0,k}(|\det|^{1-k/2}) \xrightarrow{\sim} \mathcal{S}_k$$

of $G_{\mathbf{f}}$-modules as follows. Recall that at the end of §11.3 we characterized σ_k by the existence of a certain lowest weight vector v_k. Now if τ is a homomorphism $\sigma_k \to \mathcal{A}_0$ of $(\mathfrak{g}, K_\infty)$-modules, then we define the function ϕ_τ on $G_{\mathbf{A}}$ by

$$\phi_\tau(g) = |\det(g)|^{1-k/2}(\tau(v_k))(g).$$

Then ϕ_τ is in $\mathcal{S}_k$; the left invariance under $G_{\mathbf{Q}}$ and right invariance under an open compact subgroup of $G_{\mathbf{f}}$ are clear from the definitions; the cuspidality and slowly increasing properties follow from those of $\tau(v_k)$; the transformation property with respect to U_∞ and the holomorphicity follow from the properties of v_k (see [**Roga**, Proposition 2.13] and [**Gelb**, Proposition 2.1]). One can show that, conversely, each function in $\mathcal{S}_k$ arises in this way.

Thus the $G_{\mathbf{f}}$-module $\mathcal{S}_k$ decomposes as a direct sum of admissible irreducible

$$W_\alpha = V_{\alpha,\mathbf{f}}(|\ |^{1-k/2}),$$

where α runs over the constituents of $\mathcal{A}_0$ with $V_{\alpha,\infty} \cong \sigma_k$. Moreover, writing $\pi_\alpha : G_{\mathbf{f}} \to \mathrm{GL}(W_\alpha)$ as

$$\bigotimes \pi_{\alpha,p},$$

we find that

- For each pair (α, p), $\pi_{\alpha,p}$ is an infinite-dimensional, admissible, irreducible representation of G_p. (See [**JaLa**, Proposition 9.3], also [**Cas1**, §1].)
- For each α, $\pi_{\alpha,p}$ is unramified for all but finitely many p.
- If $\pi_{\alpha,p} \simeq \pi_{\beta,p}$ for all but finitely many p, then the constituents W_α and W_β coincide (by Theorem 11.5.1).

We can now deduce Theorem 11.1.2 from the results we have collected (see [**Cas2**, §3]). Indeed each $\mathcal{S}_{k,\theta}$ is a sum of constituents W_α. If W_α is such a constituent, we see that for all but finitely many p, $\pi_{\alpha,p}$ is the unramified principal series $\pi(\mu_1, \mu_2)$ characterized by

$$p^{1/2}(\mu_1(p) + \mu_2(p)) = \theta(T_p); \qquad p\mu_1(p)\mu_2(p) = \theta(pS_p) \tag{11.5.3}$$

as can be seen from (11.2.4). Therefore W_α is unique, and conversely each W_α is an $\mathcal{S}_{k,\theta}$ with the equivalence class of θ determined by the above formulas. Furthermore, the conductor of π_α is the unique positive integer N such that $S_{k,\theta}(U_1(N))$ is one-dimensional. We thus deduce the existence of a unique newform corresponding to the (equivalence class) of the eigencharacter θ.

Note also that the Dirichlet character ε of the newform is determined by the central character of the corresponding π_α; to be precise, the central character is $\varepsilon_{\mathbf{A}}|\ |^{2-k}$.

REMARK 11.5.2. Note that for all p, $\pi_{\alpha,p} \otimes |\det|^{-1+k/2}$ is unitarizible, and for all but finitely many p, it is an unramified principal series. That it is in fact a continuous series representation for all but finitely many p is equivalent to the Ramanujan-Petersson conjecture that

$$\theta(T_p) \leq 2p^{(k-1)/2}$$

for all but finitely many p (see [**Sata**, §4], [**Gelb**, Proposition 5.17], [**Roga**, Theorem 2.14] and Remark 5.0.1). By proving and applying the Weil conjectures, Deligne [**Del1**], [**Del5**] shows that this holds for all primes p not dividing the conductor of π_α (see Remark 12.5.10 below).

Whereas the unramified local representations $\pi_{\alpha,p}$ are completely determined by the eigenvalues of T_p and S_p on the newform corresponding to α, this is not the case in general. Let us consider some examples where we can at least determine the type of ramified local representations occurring in the factorization.

EXAMPLE 11.5.3. Consider the 20 newforms of weight 2 and level dividing 33, i.e., the 20 α's such that π_α has conductor dividing 33. (See Examples 6.1.4, 6.3.5 and 10.3.3.) Since 33 is square-free, it follows from the list of possible conductors in §11.2 that no $\pi_{\alpha,p}$ is supercuspidal. Moreover if $\pi_{\alpha,p}$ is ramified, but its central character is not, then $\pi_{\alpha,p} \simeq \mathrm{sp}(\chi|\ |^{1/2}, \chi|\ |^{-1/2})$ for some χ. (Note that since $k = 2$, the central character has finite order and hence so does χ.) On the other hand, if both $\pi_{\alpha,p}$ and its central character are ramified, then $\pi_{\alpha,p} \simeq \pi(\mu_1, \mu_2)$ for some unitary μ_1 and μ_2, exactly one of which is ramified. As there are two newforms of level 33 for each of the 10 even Dirichlet characters, we find

- one π of conductor 11 with π_{11} special;
- one π of conductor 33 with π_3 and π_{11} special;
- eight π of conductor 33 with π_3 special and π_{11} principal series;
- ten π of conductor 33 with π_3 and π_{11} principal series.

In the first two cases, the central character is trivial, so the special representations are of the form $\mathrm{sp}(\chi|\ |^{1/2}, \chi|\ |^{-1/2})$ where χ may be either the trivial or the unramified quadratic character.

EXAMPLE 11.5.4. As another example, consider the unique newform of weight two, level 27 and trivial character. Since the central character is trivial and the conductor of π_3 is 27, we see that π_3 is supercuspidal. This follows again from the list of possible conductors; the conductor of a principal series or special representation π_p with unramified central character must be of the form $p^\delta \mathbf{Z}_p$ where δ is either even or 1.

REMARK 11.5.5. Recall that the L-function associated to a newform f has an Euler product and satisfies a functional equation (see Remark 5.0.2). The theory of Jacquet-Langlands offers another interpretation of the Euler product. We may view the L-function as being attached to the corresponding cuspidal automorphic representation and the Euler factor at a prime p can be described in terms of the corresponding local representations π_p (see [**JaLa**, Theorem 2.18], [**Gode**, §1.14], [**Gelb**, Theorem 6.15]). For example, if π_p is the unramified principal series $\pi(\mu_1, \mu_2)$, then the local factor is

$$L(\pi_p, s) = L(\mu_1, s)L(\mu_2, s) = (1 - \mu_1(p)p^{-s})^{-1}(1 - \mu_2(p)p^{-s})^{-1},$$

which according to (11.5.3) is simply the value at $s+1/2$ of the Euler factor at p of $L(f,s)$. The analysis of $L(\pi_p,s)$ in [**JaLa**] (see also [**Gode**, §1.14,1.15] and [**Gelb**, §6]) is in the spirit of Tate's analysis of the Euler factors of an L-function associated to a Hecke character [**Tate1**]. They define also an ϵ-factor $\epsilon(\pi_p,s)$ which plays a role in the local functional equation, and there is an analogous construction of L- and ϵ-factors at ∞ (see [**Gode**, §2.7,2.8], [**Gelb**, Theorem 6.16]). One feature of this point of view is that by [**JaLa**, Corollary 2.19] (or [**Gelb**, Theorem 6.14]), π_p is determined by its central character together with the pair of functions

$$\begin{aligned}(\chi,s) &\mapsto L(\pi_p\otimes(\chi\circ\det),s)\\ (\chi,s) &\mapsto \epsilon(\pi_p\otimes(\chi\circ\det),s)\end{aligned}$$

where χ runs over characters of $\mathbf{Q}_p^\times$.

12. Sheaves and cohomology

In this section we explain how modular forms can be viewed as sections of line bundles on modular curves. We also discuss the Eichler-Shimura isomorphism relating modular forms to the cohomology of modular curves ([**Shi1**, Chapter 8]), and the association of Galois representations to eigenforms for the Hecke operators ([**Shi1**, Chapter 7], [**Del1**], [**DeSe**]).

We shall usually try to state results for arbitrary weight, or weight $k>1$ when necessary. However the reader should be aware from the start that, as we indicate below, many of the results are much simpler in the setting of cusp forms of weight $k=2$. Then the relevant line bundle is simply the cotangent bundle, the cohomology groups are defined using constant coefficients, and the associated Galois representations are constructed from the Jacobian of the modular curve.

12.1. Line bundles.

PRIMARY REFERENCES:
[**Shi1**, Chapter 2] and [**Miy2**, §2.2–2.5].

We now explain how modular forms may be viewed as holomorphic sections of line bundles on modular curves. Much of the discussion is a reformulation in the language of sheaves of results found in [**Shi1**, §2.3–2.6], and we refer the reader there for more details.

Let k be an integer and Γ a congruence subgroup of $\mathrm{SL}_2(\mathbf{Z})$. We let X denote the modular curve $\Gamma\backslash\mathfrak{H}^*$ and Y the open subspace $\Gamma\backslash\mathfrak{H}$. We shall say that Γ is *k-small* (or simply *small* if k is fixed in the discussion), if the following two conditions are satisfied:

- if $k\neq 0$, then the image of Γ in $\mathrm{PSL}_2(\mathbf{Z})$ has no nontrivial elements of finite order;
- if k is odd, then $-1\notin\Gamma$ and the cusps of X are regular.

(If $-1\notin\Gamma$, then the cusp $\Gamma\cdot\gamma(\infty)$ with $\gamma\in\mathrm{SL}_2(\mathbf{Z})$ is *regular* if the stabilizer in Γ of $\gamma(\infty)$ is contained in $\gamma\{\left(\begin{smallmatrix}1 & *\\ 0 & 1\end{smallmatrix}\right)\}\gamma^{-1}$.) For example, if $N>4$, then $\Gamma_1(N)$ is k-small for all k. To prove this, note that if γ is an elliptic or parabolic element of $\Gamma_1(N)$, then $|\mathrm{tr}\,(\gamma)|\leq 2$ and $\mathrm{tr}\,(\gamma)\equiv 2 \bmod N$, and therefore $\mathrm{tr}\,(\gamma)=2$ (see [**Miy2**, Theorem 4.2.9]).

Now define an action of $\mathrm{SL}_2(\mathbf{Z})$ on $\mathfrak{H}\times\mathbf{C}$ by the formula

$$\alpha\cdot(z,\xi)=(\alpha(z),(cz+d)^k\xi)$$

for $\alpha = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbf{Z})$, $z \in \mathfrak{H}$ and $\xi \in \mathbf{C}$. If Γ is small, then the quotient $\Gamma\backslash(\mathfrak{H}\times\mathbf{C})$, with the natural projection map to Y, has the structure of a complex line bundle over Y. Moreover, it extends to a line bundle over X using the trivialization over a neighborhood of the cusps (see §9.1) defined by

$$\Gamma \cdot \gamma(z,\xi) \mapsto (\Gamma \cdot \gamma(z), \xi)$$

for $\gamma \in \mathrm{SL}_2(\mathbf{Z})$ and $z = x + iy$ with $y > 0$. We let G_k denote the resulting line bundle and write $\psi : G_k \to X$ for the projection map.

Next we consider the sheaf $\mathcal{G}_k$ on X of holomorphic sections of G_k. Thus $\mathcal{G}_k$ is an invertible sheaf of $\mathcal{O}_X$-modules where $\mathcal{O}_X = \mathcal{G}_0$ is the sheaf of holomorphic functions on X. If f is a modular form of weight k with respect to Γ, then we can define an element of $\mathcal{G}_k(Y)$ by $\Gamma \cdot z \mapsto \Gamma \cdot (z, f(z))$. The condition that f be holomorphic at the cusps translates to the condition that this extends to a holomorphic section $\phi_f : X \to G_k$. We find that $f \mapsto \phi_f$ establishes a natural bijection between the spaces $\mathcal{M}_k(\Gamma)$ and $\mathcal{G}_k(X) = H^0(X, \mathcal{G}_k)$.

If Γ is not small, then choose a small congruence subgroup Γ' normal in Γ and let π denote the natural projection map from $X' = \Gamma'\backslash\mathfrak{H}^*$ to X. Replacing Γ by Γ' in the definition of $\mathcal{G}_k$, we obtain an invertible sheaf $\mathcal{G}'_k$ of $\mathcal{O}_{X'}$-modules on X' and an action of Γ on the sheaf $\pi_*\mathcal{G}'_k$ which factors through Γ/Γ'. We also find that the action of γ on $\pi_*\mathcal{G}'_k(X) = \mathcal{G}'_k(X') = \mathcal{M}_k(\Gamma')$ is simply the operator $|[\gamma^{-1}]_k$ defined in §2.1. Thus $\mathcal{M}_k(\Gamma) = (\pi_*\mathcal{G}'_k)^\Gamma(X)$ where $(\pi_*\mathcal{G}'_k)^\Gamma$ is the subsheaf of $\pi_*\mathcal{G}'_k$ given by sections invariant under Γ. Thus for arbitrary k and Γ, we can write

$$\mathcal{M}_k(\Gamma) = H^0(X, \mathcal{G}_k) \tag{12.1.1}$$

where $\mathcal{G}_k$ is defined to be $(\pi_*\mathcal{G}'_k)^\Gamma$. We find that $\mathcal{G}_k$ is an invertible sheaf of $\mathcal{O}_X$-modules on X (unless k is odd and -1 is in Γ, in which case $\mathcal{M}_k(\Gamma)$ and $\mathcal{G}_k$ are both zero). It is independent of the choice of Γ' in the sense that a different choice produces a sheaf canonically isomorphic to $\mathcal{G}_k$. Moreover in the case that $\Gamma = \Gamma'$, this definition of $\mathcal{G}_k$ coincides with the previous one.

We can proceed similarly to interpret the cusp forms of weight k with respect to Γ as global sections of a certain invertible sheaf on X. Assume first that Γ is small. Let $\mathcal{C}_k \subset \mathcal{O}_X$ denote the sheaf of holomorphic functions on X which vanish at the cusps. We define $\mathcal{F}_k$ to be the invertible sheaf $\mathcal{G}_k \otimes_{\mathcal{O}_X} \mathcal{C}_k$ of $\mathcal{O}_X$-modules on X. Then $\mathcal{F}_k$ is naturally a subsheaf of $\mathcal{G}_k$ and we may identify $\mathcal{F}_k(X) \subset \mathcal{G}_k(X)$ with $\mathcal{S}_k(\Gamma) \subset \mathcal{M}_k(\Gamma)$. For arbitrary Γ, we again choose a small normal subgroup Γ', and consider the $\mathcal{O}_{X'}$-sheaf $\mathcal{F}'_k \subset \mathcal{G}'_k$ on X'. The action of Γ on $\pi_*\mathcal{G}'_k$ restricts to one on $\pi_*\mathcal{F}'_k$ and we let $\mathcal{F}_k = (\pi_*\mathcal{F}')^\Gamma$. Then $\mathcal{F}_k$ is independent of the choice of Γ and the definition agrees with the earlier one if $\Gamma = \Gamma'$. We now have

$$\mathcal{S}_k(\Gamma) = H^0(X, \mathcal{F}_k) \tag{12.1.2}$$

where $\mathcal{F}_k$ is an invertible $\mathcal{O}_X$-subsheaf of $\mathcal{G}_k$, and the identification is compatible with (12.1.1). The equation

$$\mathcal{F}_k = \mathcal{G}_k \otimes_{\mathcal{O}_X} \mathcal{C}_k \tag{12.1.3}$$

remains valid where $\mathcal{C}_k$ is defined as the sheaf of holomorphic functions which vanish at the regular cusps if k is odd, and at all cusps if k is even (see [**Shi1**, §2.4] or [**Miy2**, §2.3]).

EXAMPLE 12.1.1. If $k = 0$, then $\mathcal{G}_0 = \mathcal{O}_X$, $\mathcal{M}_0(\Gamma) = \mathcal{O}_X(X)$ is the space of constant functions $\mathbf{C}$ and $\mathcal{S}_0(\Gamma) = 0$.

EXAMPLE 12.1.2. More interesting, of course, is the case of $k = 2$. We find that $\mathcal{F}_2$ is isomorphic to Ω^1_X, the $\mathcal{O}_X$-sheaf of holomorphic differentials on X. For an explicit description of the isomorphism, first suppose that Γ is small and take $\omega \in \Omega^1_X(U)$ for an open subset U of X. Write $\varrho^*\omega = f(z)\,dz$ for a holomorphic function f on $\varrho^{-1}(U)$ where ϱ is the natural map $\mathfrak{H} \to X$. For $U^\circ = U \cap Y$, the holomorphic map $U^\circ \to G_2$ defined by $\Gamma \cdot z \mapsto \Gamma \cdot (z, f(z))$ is an element of $\mathcal{G}_2(U^\circ) = \mathcal{F}_2(U^\circ)$ which extends uniquely to an element ϕ_ω of $\mathcal{F}_2(U)$. Moreover $\omega \mapsto \phi_\omega$ defines an $\mathcal{O}_X(U)$-linear isomorphism $\Omega^1_X(U) \to \mathcal{F}_2(U)$ compatible with restriction, so

$$\Omega^1_X \cong \mathcal{F}_2. \tag{12.1.4}$$

If Γ is not small, then we still obtain (12.1.4) by choosing a suitable Γ' and checking that $\Omega^1_{X'} \cong \mathcal{F}'_2$ is compatible with the natural action of Γ. In general, (12.1.4) together with (12.1.2) gives an isomorphism ([**Shi1**, Corollary 2.17], [**Miy2**, Theorem 2.3.2])

$$\Omega^1_X(X) \xrightarrow{\sim} \mathcal{S}_2(\Gamma)$$

which is simply $\omega \mapsto f(z)$ where $\varrho^*\omega = f(z)\,dz$. In particular, note that the dimension of $\mathcal{S}_2(\Gamma)$ is just the genus of X which we recall is given by (9.1.2).

More generally, we can compute the dimension of $\mathcal{M}_k(\Gamma)$ and $\mathcal{S}_k(\Gamma)$ using the Riemann-Roch formula, provided $k \neq 1$. To do so, we first compute the degrees of the invertible sheaves $\mathcal{G}_k$ and $\mathcal{F}_k$. If Γ is k-small for all k, then we find that $G_1^{\otimes k}$ is naturally isomorphic to G_k and hence $\mathcal{G}_1^{\otimes_{\mathcal{O}_X} k} \cong \mathcal{G}_k$. Together with the isomorphisms (12.1.4) and (12.1.3), this gives

$$\mathcal{F}_k \cong \mathcal{G}_{k-2} \otimes_{\mathcal{O}_X} \Omega^1_X.$$

Moreover the formula (9.1.2) gives

$$\deg \mathcal{G}_k = (g-1)k + \nu_\infty \frac{k}{2} = \frac{k\mu}{12}; \quad \deg \mathcal{F}_k = (g-1)k + \nu_\infty(\frac{k}{2} - 1) = \frac{k\mu}{12} - \nu_\infty,$$

where g is the genus of X, μ is the index of the image of Γ in $\mathrm{PSL}_2(\mathbf{Z})$, and ν_∞ is the number of cusps on X. Thus if k is negative, so is the degree of $\mathcal{G}_k$, and $\mathcal{M}_k(\Gamma) = \mathcal{S}_k(\Gamma) = 0$. On the other hand, if $k \geq 2$ then the degree of $\mathcal{G}_k$ is greater than $2g - 2$, so

$$\dim_{\mathbf{C}} \mathcal{M}_k(\Gamma) = (g-1)(k-1) + \nu_\infty \frac{k}{2} = \frac{\mu}{12}(k-1) + \frac{\nu_\infty}{2} \tag{12.1.5}$$

by the Riemann-Roch formula. Similarly, if $k > 2$ then

$$\dim_{\mathbf{C}} \mathcal{S}_k(\Gamma) = (g-1)(k-1) + \nu_\infty(\frac{k}{2} - 1) = \frac{\mu}{12}(k-1) - \frac{\nu_\infty}{2}. \tag{12.1.6}$$

EXAMPLE 12.1.3. By Example 9.1.6, we see that if $N > 4$ and $k \geq 2$, then the dimension of $\mathcal{M}_k(\Gamma_1(N))$ is

$$(k-1)\frac{N^2}{24}\prod_{p|N}(1 - p^{-2}) + \frac{N}{4}\prod_{p|N}(1 - p^{-2} + v_p(N)(1 - p^{-1})^2).$$

For the dimension of $\mathcal{S}_k(\Gamma_1(N))$, change the "+" to a "−," and if $k = 2$ then add 1. One can write

$$\mathcal{M}_k(\Gamma_1(N)) = \mathcal{S}_k(\Gamma_1(N)) \oplus \mathcal{E}_k(\Gamma_1(N))$$

where $\mathcal{E}_k(\Gamma_1(N))$ is spanned by Eisenstein series (see §2.2) which can be described explicitly.

We shall briefly describe the situation when Γ is not k-small. For more details, see [**Shi1**, §2.4,2.6] (especially Theorems 2.23 and 2.25) or [**Miy2**, §2.5]. First suppose that $-1 \notin \Gamma$. Then we still have (12.1.4) and (12.1.3), but the natural map $\mathcal{G}_1^{\otimes_{\mathcal{O}_X} k} \to \mathcal{G}_k$ may fail to be an isomorphism at elliptic points and irregular cusps of X. Computing on stalks at these points gives instead that

$$\mathcal{G}_1^{\otimes_{\mathcal{O}_X} k} \cong \mathcal{G}_k \otimes_{\mathcal{O}_X} \mathcal{D}_k \otimes_{\mathcal{O}_X} \mathcal{E}_k \tag{12.1.7}$$

where $\mathcal{D}_k$ (respectively, $\mathcal{E}_k$) is the sheaf of holomorphic functions with zeroes of order at least $k/2$ (respectively, $k/3$ and $k/4$) at irregular cusps (respectively, elliptic points over $j = 0$ and 1728). We find that if $k \geq 2$, then $\mathcal{M}_k$ has dimension

$$(k-1)\frac{\mu}{12} + \delta_4(k)\frac{\nu_2}{4} + \delta_3(k)\frac{\nu_3}{3} + \delta_2(k)\frac{u'}{2} + \frac{u}{2}$$

if $k \geq 2$, where u (respectively, u') is the number of regular (respectively, irregular) cusps on X, $\delta_n(k) = n[k/n] - k + 1$ and the rest of the notation is as in §9.1. We have a similar expression for the dimension of $\mathcal{S}_k(\Gamma)$; replace "+" by "−" in the terms involving u and u', and add 1 if $k = 2$. If -1 is in Γ, then we assume k is even and obtain the same dimension formulas using $\mathcal{G}_2^{\otimes_{\mathcal{O}_X}(k/2)} \cong \mathcal{G}_k \otimes_{\mathcal{O}_X} \mathcal{E}_k$ instead of (12.1.7); no distinction is needed between regular and irregular cusps in this case.

EXAMPLE 12.1.4. We shall now give a complete description of the modular forms of level at most 4 in terms of the examples of §2.2.

First note that we have $\dim \mathcal{M}_k(\mathrm{SL}_2(\mathbf{Z})) = [k/12]+1$ for even positive k, unless $k \equiv 2 \bmod 12$ in which case the dimension is $[k/12]$. Similarly $\dim \mathcal{S}_k(\mathrm{SL}_2(\mathbf{Z})) = [k/12]$, unless $k > 12$ and $k \equiv 2 \bmod 12$ in which case the dimension is $[k/12] - 1$. One deduces that the Eisenstein series E_4 and E_6 of Example 2.2.1 are algebraically independent and that $\oplus_k \mathcal{M}_k(\mathrm{SL}_2(\mathbf{Z}))$ is isomorphic to a polynomial ring in the variables E_4 and E_6 (see [**Shi1**, Proposition 2.27] or [**Ser1**, VII.3.2]).

For even positive k, $\mathcal{M}_k(\Gamma_1(2))$ has dimension $[k/4] + 1$. Consider the algebra homomorphism

$$\phi : \mathbf{C}[X, Y] \to \oplus_k \mathcal{M}_k(\Gamma_1(2))$$

defined by $X \mapsto F_2$, $Y \mapsto E_4$, where F_2 is the Eisenstein series $E_2(z) - 2E_2(2z)$ of Example 2.2.6. One can check from the explicit Fourier expansions that F_2^2 and E_4 are linearly independent. It follows that so are F_2^3 and F_2E_4 and therefore E_6 is in the image. The injectivity of ϕ then follows from the algebraic independence of E_4 and E_6, and comparing dimensions for each k we conclude that ϕ is an isomorphism. Thus $\oplus_k \mathcal{M}_k(\Gamma_1(2))$ is generated as an algebra by F_2 and E_4.

Similarly we find that for $N = 3$ (respectively, 4) and any weight $k \neq 1$, $\mathcal{M}_k(\Gamma_1(N))$ has dimension $[k/3] + 1$ (respectively, $[k/2] + 1$). We also find that $\oplus_k(\mathcal{M}_k\Gamma_1(N))$ is a polynomial ring in two variables generated by the Eisenstein series $E_{1,N,\varepsilon}$ and $E_{3,N,\varepsilon}$ (respectively, $E_{1,N,\varepsilon}$ and F_2), where ε is the quadratic character of conductor N (see Example 2.2.2 for the definition and Fourier expansion).

Finally, in each of the cases with $N \leq 4$, $\oplus_k \mathcal{S}_k(\Gamma_1(N))$ is a principal ideal in the algebra of modular forms. The generators are Δ, $(\Delta(z)\Delta(2z))^{1/3}$, $(\Delta(z)\Delta(3z))^{1/4}$ and $E_{1,4,\varepsilon}^{-1}(\Delta(2z))^{1/2}$, respectively. (See Examples 2.2.7 and 2.2.8).

12.2. Cohomology.

PRIMARY REFERENCES:

[**Shi1**, Chapter 8], [**Hida3**, §6.1, 6.2] and [**Lang2**, Chapter VI].

We now turn to the Eichler-Shimura isomorphisms, which relate modular forms to the cohomology of modular curves. We maintain the notation of the preceding section. Again the results can be found in a somewhat different form in Shimura's text [**Shi1**, §8.2]; we consider the cohomology of the curves X and Y as well as that of the group Γ. (See also [**Hida3**, §6.2] and [**Del1**, (2.10)].)

REMARK 12.2.1. We postpone until §12.4 discussion of the Hecke action on cohomology, and we also restrict our attention for the moment to cohomology with complex coefficients.

We begin with the case of weight two, and explain later how the results are generalized to higher weight. Let us first consider $H^i(X, \mathbf{C})$ defined using singular cohomology or, equivalently, the cohomology of the constant sheaf $\mathbf{C}$ on X. By the de Rham theorem, $H^i(X, \mathbf{C})$ is naturally isomorphic to $H^i_{\mathrm{DR}}(X)$. Recall that $H^i_{\mathrm{DR}}(X)$ is the i^{th} cohomology group of the complex

$$0 \to \mathcal{C}^0(X) \to \mathcal{C}^1(X) \to \mathcal{C}^2(X) \to 0$$

where $\mathcal{C}^n$ denotes the sheaf of smooth complex-valued differential n-forms on X and the map $\mathcal{C}^n \to \mathcal{C}^{n+1}$ is differentiation. In particular $H^1(X, \mathbf{C})$ can be identified with the space of closed 1-forms on X modulo the space of exact 1-forms. Furthermore, according to the Hodge decomposition theorem, the natural map

$$H^{1,0}(X) \oplus H^{0,1}(X) \to H^1_{\mathrm{DR}}(X)$$

is an isomorphism where $H^{1,0}(X)$ (respectively, $H^{0,1}(X)$) is the space of holomorphic (respectively, antiholomorphic) 1-forms on X. Next recall that we have identified $H^{1,0}(X) = H^0(X, \Omega^1_X)$ with $\mathcal{S}_2(\Gamma)$. Note also that $f \mapsto \overline{f(z)}\, d\bar{z}$ defines a conjugate linear isomorphism $\mathcal{S}_2(\Gamma) \to H^{0,1}(X)$ and thus a $\mathbf{C}$-linear isomorphism $\overline{\mathcal{S}}_2(\Gamma) \cong H^{1,0}(X)$ where $\overline{\mathcal{S}}_2(\Gamma)$ denotes the complex vector space $\mathbf{C} \otimes_{\mathbf{C}} \mathcal{S}_2(\Gamma)$, the map $\mathbf{C} \to \mathbf{C}$ being complex conjugation. Thus we have a natural isomorphism

$$\mathcal{S}_2(\Gamma) \oplus \overline{\mathcal{S}}_2(\Gamma) \cong H^1(X, \mathbf{C}). \tag{12.2.1}$$

Moreover the cup product can be expressed in terms of the Petersson inner product (see (12.2.6) below).

Next we consider the cohomology of the non-compact curve Y. Let U be the intersection of Y with a sufficiently small neighborhood of the cusps of X. We find that the sequence

$$0 \to H^1(X, \mathbf{C}) \to H^1(Y, \mathbf{C}) \to H^1(U, \mathbf{C})$$

is exact, and the spaces have dimension $2g$, $2g + \nu_\infty - 1$ and ν_∞. We find also that the image of $H^1(X, \mathbf{C})$ in $H^1(Y, \mathbf{C})$ coincides with that of $H^1_c(Y, \mathbf{C}) \to H^1(Y, \mathbf{C})$ where $H^1_c(Y, \mathbf{C})$ is the cohomology with compact support. Note that this map is neither injective nor surjective if $\nu_\infty > 1$; we denote the image $H^1_p(Y, \mathbf{C})$.

Again we have a de Rham isomorphism $H^i(Y, \mathbf{C}) \cong H^i_{\mathrm{DR}}(Y)$, but now the natural map $H^0(Y, \Omega^1_Y) \to H^1_{\mathrm{DR}}(Y)$ is not injective. (Indeed $H^0(Y, \Omega^1_Y)$ is infinite-dimensional over $\mathbf{C}$ and $H^1(Y, \mathbf{C})$ is finite-dimensional.) Consider instead the composite

$$\mathcal{M}_2(\Gamma) = H^0(X, \mathcal{G}_2) \hookrightarrow H^0(Y, \Omega^1_Y) \to H^1(Y, \mathbf{C}) \tag{12.2.2}$$

where the middle injection is defined using the isomorphism $\mathcal{G}_2|_Y \cong \Omega^1_Y$ given by (12.1.3) and (12.1.4). We thus obtain a commutative diagram

$$\begin{array}{ccccccc} 0 & \to & \mathcal{S}_2(\Gamma) & \to & \mathcal{M}_2(\Gamma) & \to & \mathcal{M}_2(\Gamma)/\mathcal{S}_2(\Gamma) \\ & & \downarrow & & \downarrow & & \downarrow \\ 0 & \to & H^1(X, \mathbf{C}) & \to & H^1(Y, \mathbf{C}) & \to & H^1(U, \mathbf{C}) \end{array}$$

with exact rows. The map $\mathcal{M}_2(\Gamma) \to H^1(U, \mathbf{C})$ can be described explicitly in terms of residues at the cusps and we find the kernel is precisely $\mathcal{S}_2(\Gamma)$. Thus the last vertical map is injective, and since the first map is injective by (12.2.1), we see that so is (12.2.2). Comparing dimensions we conclude that we have produced an isomorphism

$$\mathcal{M}_2(\Gamma) \oplus \overline{\mathcal{S}}_2(\Gamma) \cong H^1(Y, \mathbf{C}) \tag{12.2.3}$$

We can also relate the cohomology of X and Y to that of the group Γ. If Γ is small, then $\mathfrak{H} \to Y$ is the universal cover and hence Γ is the fundamental group of Y. We therefore have a natural isomorphism

$$H^1(Y, \mathbf{C}) \cong H^1(\Gamma, \mathbf{C}) = \mathrm{Hom}(\Gamma, \mathbf{C}). \tag{12.2.4}$$

We check that (12.2.4) holds for arbitrary Γ by passing to a small normal subgroup Γ'. The isomorphism $H^1(Y', \mathbf{C}) \cong H^1(\Gamma', \mathbf{C})$ is compatible with the natural action of Γ and one checks that $H^1(Y, \mathbf{C})$ (respectively, $H^1(\Gamma, \mathbf{C})$) maps isomorphically to $H^1(Y', \mathbf{C})^\Gamma$ (respectively, $H^1(\Gamma', \mathbf{C})^\Gamma$). As in [**Shi1**, §8.2] or [**Hida1**, §3] (but note that we are working with coefficients in $\mathbf{C}$), one can give a very explicit description of the composite

$$\mathcal{M}_2(\Gamma) \oplus \overline{\mathcal{S}}_2(\Gamma) \xrightarrow{\sim} H^1(\Gamma, \mathbf{C}).$$

The form $f \in \mathcal{M}_2(\Gamma)$ is sent to the homomorphism $\Gamma \to \mathbf{C}$ defined by

$$\gamma \mapsto \int_{z_0}^{\gamma(z_0)} f(z)\, dz$$

for a fixed choice of base point $z_0 \in \mathfrak{H}$, and the map on $\overline{\mathcal{S}}_2(\Gamma)$ is described similarly by integrating antiholomorphic differentials. We can also identify the image of $H^1_p(Y, \mathbf{C})$, or equivalently of $\mathcal{S}_2(\Gamma) \oplus \overline{\mathcal{S}}_2(\Gamma)$, in $H^1(\Gamma, \mathbf{C})$ as $H^1_p(\Gamma, \mathbf{C})$, the group of parabolic cohomology classes. If M is a Γ-module, then we say that a class in $H^1(\Gamma, M)$ is *parabolic* if its image under restriction in $H^1(\Gamma_s, M)$ is trivial for each $s \in \mathbf{Q} \cup \{\infty\}$ (or equivalently, for each s in a set of representatives for the cusps of X), where Γ_s denotes the stabilizer in Γ of s (see [**Shi1**, §8.1], [**Hida2**, §4]). Note that $H^1_p(\Gamma, \mathbf{C})$ can be identified with $\mathrm{Hom}(\Gamma/N, \mathbf{C})$ where N is the normal subgroup generated by the Γ_s.

Next we briefly explain how this generalizes to weight $k \geq 2$ by replacing $\mathbf{C}$ with a certain $(k-1)$-dimensional representation of $\mathrm{SL}_2(\mathbf{Z})$. (See [**Shi1**, §8.2].) We let $V_k = \mathrm{Symm}^{k-2}_{\mathbf{C}}(\mathbf{C}^2)$ with an action of $\mathrm{SL}_2(\mathbf{Z})$ gotten from the standard one on $\mathbf{C}^2$. For f in $\mathcal{M}_k(\Gamma)$ we define a class in $H^1(\Gamma, V_k)$ by the cocycle

$$\gamma \mapsto \int_{z_0}^{\gamma(z_0)} f(z) \begin{pmatrix} z \\ 1 \end{pmatrix}^{k-2} dz. \tag{12.2.5}$$

Here z_0 is a basepoint, v^{k-2} denotes the image of $v \otimes \cdots \otimes v$ in $\mathrm{Symm}^{k-2}(\mathbf{C}^2)$ and the integral is that of a vector-valued differential. Together with a similar construction for antiholomorphic differentials, we obtain a $\mathbf{C}$-linear map

$$\beta : \mathcal{M}_k(\Gamma) \oplus \overline{\mathcal{S}}_k(\Gamma) \to H^1(\Gamma, V_k)$$

which restricts to

$$\beta_p : \mathcal{S}_k(\Gamma) \oplus \overline{\mathcal{S}}_k(\Gamma) \to H^1_p(\Gamma, V_k).$$

THEOREM 12.2.2. *β and β_p are isomorphisms.*

We have already covered the case $k = 2$; for $k > 2$ we sketch a proof which is a variant of the one presented in [**Shi1**, §8.2]; see also [**Hida3**, §6.2]. First one reduces to the case where Γ is small. Then $\pi : \mathfrak{H} \to Y$ is the universal cover, and the covering group Γ acts naturally on $(\pi^*\mathbf{V}_k)(\mathfrak{H}) \cong V_k$ where $\mathbf{V}_k$ is the locally constant sheaf of continuous sections from Y to $\Gamma\backslash(\mathfrak{H} \times V_k)$, where V_k is given the discrete topology. Since $\mathfrak{H}$ is contractible, the natural maps

$$H^i(\Gamma, V_k) \to H^i(Y, \mathbf{V}_k)$$

are isomorphisms [**Mum1**, §2, Appendix]. These groups vanish for $i \neq 1$ assuming $k > 2$. Let

$$\alpha : \mathcal{M}_k(\Gamma) \oplus \overline{\mathcal{S}}_k(\Gamma) \to H^1(Y, \mathbf{V}_k)$$

denote the composite of β with $H^1(\Gamma, V_k) \xrightarrow{\sim} H^1(Y, \mathbf{V}_k)$.

Now consider the restriction map $r : H^1(Y, \mathbf{V}_k) \to H^1(U \cap Y, \mathbf{V}_k)$ where U is again a suitable neighborhood of the cusps of X. We find that r is surjective with kernel $H^1_p(Y, \mathbf{V}_k)$ (the image of $H^1_c(Y, \mathbf{V}_k)$) and that the kernel of $r \circ \alpha$ is precisely $\mathcal{S}_k(\Gamma) \oplus \overline{\mathcal{S}}_k(\Gamma)$. (See [**Hida2**, §5] for such an argument in the context of group cohomology.) Thus α restricts to a homomorphism

$$\alpha_p : \mathcal{S}_k(\Gamma) \oplus \overline{\mathcal{S}}_k(\Gamma) \to H^1_p(Y, \mathbf{V}_k).$$

That α and α_p are isomorphisms follows on combining the assertions

- α_p is injective;
- $\dim_{\mathbf{C}} H^1(Y, \mathbf{V}_k) = \dim_{\mathbf{C}} \mathcal{M}_k(\Gamma) + \dim_{\mathbf{C}} \mathcal{S}_k(\Gamma)$.

To prove the first assertion, we use the cup product to construct a pairing on $H^1_p(Y, \mathbf{V}_k)$ which is compatible with the Petersson inner product discussed in §3.6. First note that Γ acts trivially on $\wedge^2_{\mathbf{C}}(V_3)$, so the standard alternating pairing $v \otimes w \mapsto \det(v, w)$ defines a Γ-linear map $\pi : V_3 \otimes_{\mathbf{C}} V_3 \to \mathbf{C}$, where $\mathbf{C}$ has trivial Γ-action. Next one checks that there is a unique Γ-linear map

$$\pi_k : V_k \otimes_{\mathbf{C}} V_k \to \mathbf{C}$$

such that $\pi_k(v^{k-2} \otimes w^{k-2}) = \pi(v \otimes w)^{k-2}$ (using the notation introduced following (12.2.5)). This then defines a homomorphism $\mathbf{V}_k \otimes_{\mathbf{C}} \mathbf{V}_k \to \mathbf{C}$ of sheaves on Y. The composite

$$H^1_c(Y, \mathbf{V}_k) \otimes_{\mathbf{C}} H^1(Y, \mathbf{V}_k) \xrightarrow{\cup} H^2_c(Y, \mathbf{V}_k \otimes_{\mathbf{C}} \mathbf{V}_k) \xrightarrow{\pi_k^*} H^2_c(Y, \mathbf{C}) \cong \mathbf{C}$$

induces the desired pairing ϕ_k. Taking $f_i, g_i \in \mathcal{S}_k(\Gamma)$ for $i = 1, 2$, we find (see [**Shi1**, (8.2.18)])

$$\phi_k(\alpha_p(f_1, \overline{g}_1) \otimes \alpha_p(f_2, \overline{g}_2)) = C_k(\langle f_1, g_2 \rangle + (-1)^{k-1}\langle f_2, g_1 \rangle), \tag{12.2.6}$$

where $C_k \neq 0$ depends only on k and we have written $\overline{g}_i$ for $1 \otimes g_i \in \overline{\mathcal{S}}_k(\Gamma)$. (This formula holds also for $k = 2$.) The injectivity of α_p then follows from the nondegeneracy of the Petersson inner product.

To prove the second assertion, note that by (12.1.5) and (12.1.6) we have

$$\dim_{\mathbf{C}} \mathcal{M}_k(\Gamma) + \dim_{\mathbf{C}} \mathcal{S}_k(\Gamma) = (2g - 2 + \nu_\infty)(k - 1) = -\chi(Y) \dim_{\mathbf{C}}(V_k)$$

where $\chi(Y) = \sum(-1)^i \dim_{\mathbf{C}} H^i(Y, \mathbf{C})$ is the Euler characteristic of Y. We can then appeal to the Mayer-Victoris sequence for sheaf cohomology to check that

$$\sum(-1)^i \dim_{\mathbf{C}} H^i(Y, \mathbf{V}_k) = \chi(Y) \dim_{\mathbf{C}}(V_k),$$

and we are done since $H^i(Y, \mathbf{V}_k) = 0$ for $i \neq 1$.

REMARK 12.2.3. Note that we may also regard the quotient $\Gamma\backslash(\mathfrak{H} \times V_k)$ as a vector bundle over Y and consider the $\mathcal{O}_Y$-sheaf $\mathcal{V}_k \cong \mathbf{V}_k \otimes_{\mathbf{C}} \mathcal{O}_Y$ of holomorphic sections. The map $\mathfrak{H} \times \mathbf{C} \to \mathfrak{H} \times V_k$ defined by

$$(z, \xi) \mapsto (z, \xi \begin{pmatrix} z \\ 1 \end{pmatrix}^{k-2})$$

induces a morphism $(\mathcal{G}_{k-2})|_Y \to \mathcal{V}_k$ of $\mathcal{O}_Y$-sheaves on Y. Tensoring with $\mathcal{G}_2|_Y \xrightarrow{\sim} \Omega^1_Y$ (see (12.1.4)), we obtain $\mathcal{G}_k|_Y \to \mathcal{V}_k \otimes_{\mathcal{O}_Y} \Omega^1_Y$. The restriction of α to $\mathcal{M}_k(\Gamma)$ can then be described as the composite

$$\mathcal{M}_k(\Gamma) \to \mathcal{G}_k(Y) \to (\mathcal{V}_k \otimes_{\mathcal{O}_Y} \Omega^1_Y)(Y) \to H^1(Y, \mathbf{V}_k),$$

the last map coming from the de Rham isomorphism.

12.3. The q-expansion principle.

PRIMARY REFERENCES:

[**DeRa**, §VII.3], [**Katz1**, Chapter 1] and [**Maz1**, §II.4,II.5].

We now discuss some of the theory of modular forms with coefficients in rings other than $\mathbf{C}$. Such a theory is useful, for example, in the study of congruences between eigenvalues of Hecke operators.

Let $\Gamma = \Gamma_0(N)$ or $\Gamma_1(N)$ and consider the injective map

$$\mathcal{M}_k(\Gamma) \to \mathbf{C}[[q]] \tag{12.3.1}$$

sending a modular form to its q-expansion, i.e., its Fourier expansion at ∞ as in (2.1.1). Let $\mathbf{M}_k(\Gamma; \mathbf{Z})$ denote the preimage of $\mathbf{Z}[[q]]$, i.e., set of elements of $\mathcal{M}_k(\Gamma)$ with Fourier coefficients in $\mathbf{Z}$. For an arbitrary ring A, we write $\mathbf{M}_k(\Gamma; A)$ for $\mathbf{M}_k(\Gamma; \mathbf{Z}) \otimes A$. Since $\mathbf{M}_k(\Gamma; \mathbf{Z}) \to \mathbf{Z}[[q]]$ has torsion-free cokernel, the map $\mathbf{M}_k(\Gamma; A) \to A[[q]]$ obtained by tensoring with A is also injective. We define $\mathbf{S}_k(\Gamma; A)$ similarly using cusp forms, and we identify it with an A-submodule of $\mathbf{M}_k(\Gamma; A)$. (Note that $\mathbf{S}_k(\Gamma; \mathbf{Z}) = \mathbf{M}_k(\Gamma; \mathbf{Z}) \cap \mathcal{S}_k(\Gamma)$.)

Let us naively call $\mathbf{M}_k(\Gamma; A)$ (respectively, $\mathbf{S}_k(\Gamma; A)$) the A-module of modular forms (respectively, cusp forms) with coefficients in A. The definition is naive in that we have not shown that $\mathbf{M}_k(\Gamma; \mathbf{Z})$ contains bases for $\mathcal{M}_k(\Gamma)$ and $\mathcal{S}_k(\Gamma)$, and we need this in order to identify $\mathbf{M}_k(\Gamma; \mathbf{C})$ with $\mathcal{M}_k(\Gamma)$. The existence of such bases is due to Shimura; see [**Shi1**, Theorem 3.52] for the case of $\mathcal{S}_k(\Gamma)$ with $k \geq 2$. Here, however, we shall explain how to deduce the general result from the q-expansion principle of Deligne-Rapoport [**DeRa**, §VII.3] and Katz [**Katz1**, Chapter 1] (see also [**Katz2**, Chapter 2]).

The injectivity of $\mathbf{M}_k(\Gamma; A) \to A[[q]]$ may be viewed as a naive version of the q-expansion principle. To state a more powerful version, we need an algebraic notion of a modular form with coefficients in an arbitrary ring A. We begin by regarding the sheaves $\mathcal{G}_k$ (see §12.1) as arising naturally in the context of the moduli problems discussed in §7.2 and §8.

For simplicity, we restrict our attention for the moment to $\Gamma_1(N)$ with $N > 4$. Recall from §8.2 that there is a universal elliptic curve with a point of order N over the model $\mathcal{Y}_1(N)$ for $Y_1(N) = \Gamma_1(N)\backslash\mathfrak{H}$. In the consideration of q-expansions

it is more convenient to use the model $\mathcal{Y}_\mu(N)$ of Variant 8.2.2. Let $\mathcal{E}_{\text{univ}}$ now denote the universal elliptic curve (over $\mathcal{Y}_\mu(N)$), and i_{univ} the canonical immersion $(\boldsymbol{\mu}_N)_{\mathcal{Y}_\mu(N)} \hookrightarrow \mathcal{E}_{\text{univ}}$. We let E_{univ} denote the complex points of $\mathcal{E}_{\text{univ}}$; thus E_{univ} is a complex analytic family of elliptic curves over $Y_1(N)$. We choose $e^{2\pi i/N}$ as our Nth root of unity and let $P_{\text{univ}} : Y_1(N) \to E_{\text{univ}}$ be the corresponding family of points of order N given by i_{univ}.

We can describe E_{univ} and P_{univ} concretely as follows. Define a right action of $\mathbf{Z} \times \mathbf{Z}$ on $\mathfrak{H} \times \mathbf{C}$ by

$$(z, \zeta) \cdot (m, n) = (z, \zeta + mz + n)$$

for $m, n \in \mathbf{Z}$, $z \in \mathfrak{H}$ and $\zeta \in \mathbf{C}$. The quotient is naturally a family of elliptic curves over $\mathfrak{H}$, and $z \mapsto (z, 1/N)$ defines a family of points of order N. Now define a left action of $\Gamma_1(N)$ on the quotient so that $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ sends the orbit of (z, ζ) to that of $(\gamma(z), (cz + d)^{-1}\zeta)$. The quotient

$$\Gamma_1(N)\backslash((\mathfrak{H} \times \mathbf{C})/(\mathbf{Z} \times \mathbf{Z})),$$

viewed as an elliptic curve over $Y_1(N)$ can be identified with E_{univ}, and the section defined by $z \mapsto 1/N$ can be identified with P_{univ}.

The line bundle $G_1|_{Y_1(N)}$ on $Y_1(N)$ (see §12.1) can now be identified with restriction along the zero-section of the relative cotangent bundle of E_{univ} over $Y_1(N)$. To make this identification precise, note that the latter bundle is canonically

$$\Gamma_1(N)\backslash(\mathfrak{H} \times V), \tag{12.3.2}$$

where V is the cotangent space of $\mathbf{C}$ at the origin and the action of $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ on $\mathfrak{H} \times V$ is given by $(z, d\zeta) \mapsto (\gamma(z), (cz + d)d\zeta)$. We identify this with $G_1|_{Y_1(N)}$ via

$$(z, \xi) \leftrightarrow (z, 2\pi i \xi d\zeta).$$

We can extend the moduli-theoretic description of G_1 to the cusps by considering the universal generalized elliptic curve with an immersion of $\boldsymbol{\mu}_N$ (see Variant 9.3.6). Again denote the universal curve $\mathcal{E}_{\text{univ}}$ (now over $\mathcal{X}_\mu(N)$) and consider its complex points; these form a complex analytic family of generalized elliptic curves E_{univ} over $X_1(N)$. We can again give a concrete description of E_{univ}; in particular, its fiber over the cusp $\Gamma_1(N) \cdot \infty$ is simply $\mathbf{C}/\mathbf{Z}$, the point of order N being $1/N$ mod $\mathbf{Z}$. The description of E_{univ} in these terms near other cusps is slightly more complicated (see §9.3 and [**DeRa**, VII.4]), and we will not go into detail here. However, the cotangent bundle of E_{univ} over $X_1(N)$ restricted to the zero section depends only on the identity component, and can again be identified with G_1.

An advantage of the moduli-theoretic description of G_1 is that the base need not be $\mathbf{C}$. Indeed we can construct an invertible sheaf on $\mathcal{X}_\mu(N)$ which is a "canonical model" for the line bundle G_1. Let ω denote the pull-back along the zero section $\mathcal{X}_\mu(N) \to \mathcal{E}_{\text{univ}}$ of the sheaf $\Omega^1_{\mathcal{E}_{\text{univ}}/\mathcal{X}_\mu(N)}$. Then ω is an invertible sheaf on $\mathcal{X}_\mu(N)$, and the complex analytic sheaf on $X_1(N)$ associated to $\omega_{\mathbf{C}}$ can be identified with the sheaf we denoted $\mathcal{G}_1$ in §12.1. We also have $\omega^{\otimes k}$ as a model for $\mathcal{G}_k$. The Gauss-Manin connection yields an isomorphism

$$\omega^{\otimes 2}|_{\mathcal{Y}_\mu(N)} \to \Omega^1_{\mathcal{Y}_\mu(N)/\mathbf{Z}}. \tag{12.3.3}$$

(see [**DeRa**, §VI.4.5], [**Katz1**, A1.3] and [**Schl**, 2.4]). Moreover the complement of $\mathcal{Y}_\mu(N)$ in $\mathcal{X}_\mu(N)$ defines a divisor $\mathcal{Z}_\mu(N)$ and we write $\mathcal{L}$ for the corresponding

invertible sheaf on $\mathcal{X}_\mu(N)$. Then (12.3.3) extends to an isomorphism

$$\omega^{\otimes 2} \to \Omega^1_{\mathcal{X}_{\mu(N)}/\mathbf{Z}} \otimes \mathcal{L}. \tag{12.3.4}$$

The isomorphism is compatible with that of (12.1.4) and allows us to regard $\omega^{\otimes(k-2)} \otimes \Omega^1_{\mathcal{X}_\mu(N)/\mathbf{Z}}$ as a model for $\mathcal{F}_k$.

For an arbitrary ring A, we define *a modular form over* A (of weight k with respect to $\Gamma_1(N)$) to be an element of

$$H^0(\mathcal{X}_\mu(N)_A, \omega_A^{\otimes k}).$$

Similarly, we define *a cusp form over* A as an element of

$$H^0(\mathcal{X}_\mu(N)_A, \omega_A^{\otimes(k-2)} \otimes \Omega^1_{\mathcal{X}_\mu(N)_A/A}).$$

We write $\mathcal{M}_k(\Gamma_1(N); A)$ for the A-module of modular forms over A. We write $\mathcal{S}_k(\Gamma_1(N); A)$ for the cusp forms which we regard as a submodule of $\mathcal{M}_k(\Gamma_1(N); A)$ via (12.3.4).

REMARK 12.3.1. Note the terminology *over* A to distinguish this from the naive definition of modular forms *with coefficients in* A.

Identifying $\mathcal{G}_k$ with the complex analytic sheaf associated to $\omega_{\mathbf{C}}^{\otimes k}$, we obtain natural isomorphisms

$$\mathcal{M}_k(\Gamma_1(N); \mathbf{C}) \cong \mathcal{M}_k(\Gamma_1(N)); \qquad \mathcal{S}_k(\Gamma_1(N); \mathbf{C}) \cong \mathcal{S}_k(\Gamma_1(N)). \tag{12.3.5}$$

Base change arguments (see [**Katz1**, §1.7] and [**Maz1**, II.3]) together with Theorem 9.3.7 yield the following result.

THEOREM 12.3.2. *If B is an A-algebra and either of the following hold*

- *B is flat over A;*
- *$k > 1$ and N is invertible in B,*

then the natural maps

$$\mathcal{M}_k(\Gamma_1(N); A) \otimes_A B \to \mathcal{M}_k(\Gamma_1(N); B);$$
$$\mathcal{S}_k(\Gamma_1(N); A) \otimes_A B \to \mathcal{S}_k(\Gamma_1(N); B)$$

are isomorphisms.

REMARK 12.3.3. The definition we have given for a modular form over A is most convenient for the applications below. However it is not necessarily the most suitable if, for example, A is a field of characteristic p dividing N. Moreover we have restricted our attention here to $\Gamma_1(N)$ with $N > 4$. For discussion of various notions of modular forms over a ring A, base-change and q-expansion in a more general context, see [**DeRa**, VII.3], [**Katz1**, §1.7, 1.8], [**Katz2**, §II.2.2], [**Maz1**, II.4] and [**Gross**, §10].

We now explain how Deligne and Rapoport's algebraic description of the cusps allows us to define the q-expansion of a modular form over A. More details can be found in [**DeRa**, VII.3] and [**Katz1**, A.1.3]. See also [**Gross**, §2] for statements in the context of modular forms with respect to $\Gamma_1(N)$.

In our discussion in §9.3 of the work of Deligne-Rapoport [**DeRa**], we described how the cusps of $X_1(N)$ correspond to degenerate elliptic curves. Moreover we indicated how Tate curves can be used to describe the completion of $\mathcal{X}_1(N)$ along D, the cuspidal divisor. However in the present discussion we shall continue to use the models discussed in Variant 9.3.6.

Now consider the point s_∞ of $\mathcal{X}_\mu(N)(\mathbf{Z})$ arising from the generalized elliptic curve $\mathbf{P}^1$ (over $\mathbf{Z}$) with its canonical embedding of $\boldsymbol{\mu}_N$. The image of s_∞ in $\mathcal{X}_\mu(N)(\mathbf{C}) \cong X_1(N)$ is the cusp we have identified with $\Gamma_1(N) \cdot \infty$. The map $\operatorname{Spec} \mathbf{Z} \to \mathcal{X}_\mu(N)$ is a closed immersion, and we write $\hat{\mathcal{X}}_\mu(N)$ for the completion of $\mathcal{X}_\mu(N)$ along the image. The Tate curve E_q over $\mathbf{Z}[[q]]$ has a canonical immersion of $\boldsymbol{\mu}_N$ so that the composite

$$\operatorname{Spec} \mathbf{Z} \to \operatorname{Spec} \mathbf{Z}[[q]] \to \mathcal{X}_\mu(N)$$

is s_∞, where the first map is the closed immersion defined by $q \mapsto 0$. This gives rise to a morphism of formal schemes

$$j_\infty : \operatorname{Spf} \mathbf{Z}[[q]] \to \hat{\mathcal{X}}_\mu(N),$$

which is in fact an isomorphism (see [**DeRa**, VII.2]). Moreover the isomorphism identifies the completion of ω with the sheaf on $\operatorname{Spf} \mathbf{Z}[[q]]$ corresponding to the $\mathbf{Z}[[q]]$-module $e^*\Omega^1_{E_q/\mathbf{Z}[[q]]}$, where e is the zero section of E_q. But this is a free $\mathbf{Z}[[q]]$-module with a canonical generator denoted ω_{can} in [**Katz1**, A.1.3]. Using $\omega_{\text{can}}^{\otimes k}$ as a generator for the completion of $\omega^{\otimes k}$ and working over an arbitrary ring A, one obtains the q-expansion homomorphism

$$\phi_{\infty,A} : \mathcal{M}_k(\Gamma_1(N); A) \to A[[q]]. \tag{12.3.6}$$

The restriction to $\mathcal{S}_k(\Gamma_1(N); A)$ maps to $qA[[q]]$. The maps $\phi_{\infty,A}$ are functorial in A, and in the case $A = \mathbf{C}$, this becomes the usual q-expansion at ∞ in (2.1.1).

We now state the q-expansion principle of [**DeRa**, Theorem VII.3.9] in our context. It is proved using Theorem 9.3.7 and the arguments of Deligne-Rapoport or Katz, [**Katz1**, §1.6].

THEOREM 12.3.4. 1. *The q-expansion homomorphism $\phi_{A,\infty}$ is injective for every ring A.*

2. *If A is a subring of B, then the commutative diagram*

$$\begin{array}{ccc} \mathcal{M}_k(\Gamma_1(N), A) & \xrightarrow{\phi_{\infty,A}} & A[[q]] \\ \downarrow & & \downarrow \\ \mathcal{M}_k(\Gamma_1(N), B) & \xrightarrow{\phi_{\infty,B}} & B[[q]] \end{array}$$

is Cartesian; i.e., the image of $\mathcal{M}_k(\Gamma_1(N); A)$ in $\mathcal{M}_k(\Gamma_1(N); B)$ is precisely the set of modular forms whose q-expansions at ∞ have coefficients in A.

3. *The above assertions hold with $\mathcal{M}_k$ replaced by $\mathcal{S}_k$.*

The first part of the theorem states that a modular form over A is determined by its q-expansion at ∞. The second part of the theorem shows in particular that the image of $\mathcal{M}_k(\Gamma_1(N); \mathbf{Z})$ in $\mathcal{M}_k(\Gamma_1(N); \mathbf{C}) = \mathcal{M}_k(\Gamma_1(N))$ is precisely $\mathbf{M}_k(\Gamma_1(N); \mathbf{Z})$. More generally if R is a subring of $\mathbf{C}$, we may identify $\mathcal{M}_k(\Gamma_1(N); R)$ with the set of modular forms whose Fourier coefficients at ∞ lie in R. Analogous statements hold for cusp forms by the third part of the theorem.

REMARK 12.3.5. For each cusp s of $X_1(N)$ and $\mathbf{Z}[1/N, e^{2\pi i/N}]$-algebra A, one can define a corresponding q-expansion homomorphism $\phi_{s,A}$ with values in $A[[q^{1/h}]]$ (for suitable h). One then obtains the analogue of Theorem 12.3.4 (see [**DeRa**, §VII.3]). In particular, if f is in $\mathcal{M}_k(\Gamma_1(N), A)$, then its q-expansion $\phi_{s,A}(f)$ is identically zero for some cusp s if and only if it is identically zero for all cusps s. We have also that if the q-expansion at one cusp has coefficients in a $\mathbf{Z}[1/N, e^{2\pi i/N}]$-subalgebra B of A, then so does it at all cusps.

An element of $\mathcal{M}_k(\Gamma_1(N); A)$ is in $\mathcal{S}_k(\Gamma_1(N); A)$ if and only if the constant term of the q-expansion vanishes at all cusps.

For $f \in \mathbf{M}_k(\Gamma_1(N); \mathbf{Z}) \subset \mathcal{M}_k(\Gamma_1(N))$, it need not be the case that the Fourier expansions have integer coefficients at cusps other than $\Gamma_1(N) \cdot \infty$. One can show however that the denominators of the coefficients of $\phi_{s,A}(f) \in \mathbf{Z}[1/N, e^{2\pi i/N}][[q]]$ are bounded. (See [**DeRa**, Corollary VII.3.11], [**Katz1**, A.1.2].)

REMARK 12.3.6. Had we used the model $\mathcal{X}_1(N)$ for $X_1(N)$ of §9.3 rather than that of Variant 9.3.6, the cusp $\Gamma_1(N) \cdot \infty$ would not be defined over $\mathbf{Q}$. We would then have had to restrict our attention throughout to algebras over $\mathbf{Z}[1/N, e^{2\pi i/N}]$.

We now record some consequences of the q-expansion principle. First, we combine Theorem 12.3.4 with Theorem 12.3.2 to obtain the following.

THEOREM 12.3.7. *The natural maps*

$$\begin{array}{ccc} \mathbf{M}_k(\Gamma_1(N); A) & \to & \mathcal{M}_k(\Gamma_1(N); A); \\ \mathbf{S}_k(\Gamma_1(N); A) & \to & \mathcal{S}_k(\Gamma_1(N); A) \end{array}$$

are injective, and are isomorphisms provided one of the following holds

- *A is flat over $\mathbf{Z}$;*
- *$k > 1$ and N is invertible in A.*

Note especially that this holds if $k > 1$ and A is a field of characteristic prime to N. (Recall that we are assuming $N > 4$.)

Note also that Theorem 12.3.7 holds if $A = \mathbf{C}$ yielding the following corollary.

COROLLARY 12.3.8. *For all positive integers N and k, the space $\mathcal{M}_k(\Gamma_1(N))$ (respectively, $\mathcal{S}_k(\Gamma_1(N))$) has a basis in $\mathbf{M}_k(\Gamma_1(N); \mathbf{Z})$ (respectively, $\mathbf{S}_k(\Gamma_1(N); \mathbf{Z})$).*

Note that we have removed the assumption that $N > 4$. Indeed Example 12.1.4 shows that for $N \leq 4$, the spaces are spanned by monomials in forms with integer Fourier coefficients (see (2.2.5) and Example 2.2.7).

COROLLARY 12.3.9. *Suppose that $f = \sum a_n q^n$ is in $\mathcal{M}_k(\Gamma_1(N))$ (respectively, $\mathcal{S}_k(\Gamma_1(N))$) and σ is an automorphism of $\mathbf{C}$. Then there is a form f^σ in $\mathcal{M}_k(\Gamma_1(N))$ (respectively, $\mathcal{S}_k(\Gamma_1(N))$) with Fourier expansion $\sum a_n^\sigma q^n$.*

This follows from Corollary 12.3.8, or in case $N > 4$ directly from the q-expansion principle Theorem 12.3.4 applied to $\sigma : A \to B$ with $A = B = \mathbf{C}$.

COROLLARY 12.3.10. *Suppose that $k > 0$ and $f = \sum a_n q^n$ is in $\mathcal{M}_k(\Gamma_1(N))$. Let K be the the subfield of $\mathbf{C}$ generated by $\{ a_n \mid n > 0 \}$. Then a_0 is in K.*

The proof is as follows ([**Shi6**, Proposition 1.3]). If σ is a field automorphism of $\mathbf{C}$ fixing K, then the constant $a_0^\sigma - a_0 = f^\sigma - f$ is in $\mathcal{M}_k(\Gamma_1(N))$ and hence is equal to 0.

Again assume that $N > 4$ and let $\mathcal{E}_{\text{univ}}$ denote the universal generalized elliptic curve over $\mathcal{X}_\mu(N)$ with immersion i_{univ} of $\boldsymbol{\mu}_N$. For d in $(\mathbf{Z}/N\mathbf{Z})^\times$, we denote by $\langle d \rangle$ the automorphism of $\mathcal{X}_\mu(N)$ corresponding to the pair $(\mathcal{E}_{\text{univ}}, di_{\text{univ}})$. Then $\langle d \rangle$ is a model over $\mathbf{Z}$ for $\langle d \rangle : X_1(N) \to X_1(N)$. Moreover we may identify $\langle d \rangle^* \omega$ with ω and thus obtain an action of $(\mathbf{Z}/N\mathbf{Z})^\times$ on

$$\mathcal{M}_k(\Gamma_1(N); \mathbf{Z}) = H^0(\mathcal{X}_\mu(N), \omega^{\otimes k}).$$

More generally we can define in this way actions of $(\mathbf{Z}/N\mathbf{Z})^\times$ on $\mathcal{M}_k(\Gamma_1(N); A)$ and $\mathcal{S}_k(\Gamma_1(N); A)$ for arbitrary A. The action is functorial in A and respects the

inclusion of $\mathcal{S}_k(\Gamma_1(N); A)$ in $\mathcal{M}_k(\Gamma_1(N); A)$. Moreover it is compatible via (12.3.5) with the action of $(\mathbf{Z}/N\mathbf{Z})^\times$ on $\mathcal{M}_k(\Gamma_1(N))$ of the operators denoted $\langle d\rangle_k$ in §2.1.

In particular, it follows that

PROPOSITION 12.3.11. 1. *The subsets* $\mathbf{M}_k(\Gamma_1(N); \mathbf{Z})$ *and* $\mathbf{S}_k(\Gamma_1(N); \mathbf{Z})$ *are preserved by the operators* $\langle d\rangle$ *for* $d \in (\mathbf{Z}/N\mathbf{Z})^\times$.

2. *If* f *is in* $\mathcal{M}_k(N, \varepsilon)$, *then* f^σ *is in* $\mathcal{M}_k(N, \varepsilon^\sigma)$, *where* f^σ *is defined in* Corollary 12.3.9 *and* ε^σ *denotes* $\sigma \circ \varepsilon$.

The second assertion follows from the fact that $\langle d\rangle_k$ commutes with $f \mapsto f^\sigma$. Note that both assertions hold for all $N \geq 1$, since for $N \leq 4$ each $\langle d\rangle$ acts by ± 1 on $\mathcal{M}_k(\Gamma_1(N))$.

Now consider the "trace" map

$$\sum_{d \in (\mathbf{Z}/N\mathbf{Z})^\times} \langle d\rangle_k : \mathcal{M}_k(\Gamma_1(N)) \to \mathcal{M}_k(\Gamma_0(N)).$$

The map is surjective as its restriction to $\mathcal{M}_k(\Gamma_0(N))$ is multiplication by $\phi(N) = |(\mathbf{Z}/N\mathbf{Z})^\times|$. By Proposition 12.3.11, we see that $\mathbf{M}_k(\Gamma_1(N); \mathbf{Z})$ is mapped to $\mathbf{M}_k(\Gamma_0(N); \mathbf{Z}) = \mathbf{M}_k(\Gamma_1(N); \mathbf{Z}) \cap \mathcal{M}_k(\Gamma_0(N))$. The same assertions hold for cusp forms and we deduce the following from Corollary 12.3.8.

COROLLARY 12.3.12. *Let* $\Gamma = \Gamma_0(N)$ *or* $\Gamma_1(N)$ *with* $N \geq 1$. *Then* $\mathcal{M}_k(\Gamma)$ (*respectively,* $\mathcal{S}_k(\Gamma)$) *has a basis in* $\mathbf{M}_k(\Gamma; \mathbf{Z})$ (*respectively,* $\mathbf{S}_k(\Gamma; \mathbf{Z})$).

This holds also for Γ satisfying $\Gamma_1(N) \subset \Gamma \subset \Gamma_0(N)$. One also finds that the spaces $\mathcal{M}_k(N, \varepsilon)$ and $\mathcal{S}_k(N, \varepsilon)$ are spanned by forms with q-expansions in $\mathbf{Z}[\varepsilon]$ where $\mathbf{Z}[\varepsilon]$ denotes the ring generated by the values of ε.

12.4. Hecke action.

PRIMARY REFERENCES:

[**Shi1**, Chapters 3,8], [**Del1**, §3] and [**Hida3**, §6.3].

For f in $\mathcal{M}_k(\Gamma_1(N))$ and $n \geq 0$, let us write $a_n(f)$ for the nth Fourier coefficient in the q-expansion of f at ∞. For each positive integer m, Proposition 3.4.3 gives

$$a_n(T_m f) = \sum d^{k-1} a_{mn/d^2}(\langle d\rangle f), \tag{12.4.1}$$

the sum being over positive divisors d of (m, n) which are relatively prime to N. So by Proposition 12.3.11, we have

PROPOSITION 12.4.1. *Let* k, N *and* m *be positive integers, and let* $\Gamma = \Gamma_0(N)$ *or* $\Gamma_1(N)$. *If* f *is in* $\mathbf{M}_k(\Gamma; \mathbf{Z})$, *then so is* $T_m f$, *and similarly for* $\mathbf{S}_k(\Gamma; \mathbf{Z})$.

REMARK 12.4.2. Note that we do not need to appeal to Proposition 12.3.11 in the case of $\Gamma = \Gamma_0(N)$; indeed Proposition 12.4.1 is immediate from (12.4.1).

Let $\tilde{\mathbb{T}}$ be as in Proposition 3.5.1, i.e., the subring of $\operatorname{End} \mathcal{M}_k(\Gamma)$ generated by the Hecke operators T_m for all $m \geq 1$, or equivalently by the $\{T_p, \langle q\rangle_k\}$ for all primes p and all primes $q \nmid N$. By Proposition 12.4.1 we may regard $\mathbf{M}_k(\Gamma; \mathbf{Z})$ as a $\tilde{\mathbb{T}}$-module. Thus for an arbitrary ring A, we may regard

$$\mathbf{M}_k(\Gamma; A) = \mathbf{M}_k(\Gamma; \mathbf{Z}) \otimes A$$

as a $\tilde{\mathbb{T}} \otimes A$-module.

Appealing to Corollary 12.3.12, we see that the map $\tilde{\mathbb{T}} \to \operatorname{End} \mathbf{M}_k(\Gamma; \mathbf{Z})$ is injective. Therefore

COROLLARY 12.4.3. *The ring $\tilde{\mathbb{T}}$ is a finitely generated free* $\mathbf{Z}$*-module.*

REMARK 12.4.4. See the discussion following Proposition 12.4.10 for an alternate proof in the case $k \geq 2$ using the Eichler-Shimura isomorphism.

The result has the following application to Hecke eigenvalues. (See for example [**Shi1**, §3.5] or [**Shi3**, §1].) Suppose that $f \in \mathcal{M}_k(\Gamma)$ is a simultaneous eigenform for the operators in $\tilde{\mathbb{T}}$ and consider the eigencharacter $\theta : \tilde{\mathbb{T}} \to \mathbf{C}$ defined by $f|T = \theta(T)f$. The image of the ring $\tilde{\mathbb{T}}$ in $\mathbf{C}$ is finitely generated as a $\mathbf{Z}$-module, hence is contained in the ring of algebraic integers of a number field. Moreover, for $\sigma \in \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ and $d \in (\mathbf{Z}/N\mathbf{Z})^\times$, it follows from (12.4.1) and the compatibility of $f \mapsto f^\sigma$ with $f \mapsto \langle d \rangle f$ that $T_m f^\sigma = (T_m f)^\sigma$.

COROLLARY 12.4.5. *Let k and N be positive integers and ε a mod N Dirichlet character. Suppose that $f = \sum a_n q^n$ is a normalized eigenform in $\mathcal{M}_k(N, \varepsilon)$ (respectively, $\mathcal{S}_k(N, \varepsilon)$) for the Hecke operators T_n, for all $n \geq 1$. Then there is a number field whose ring of integers contains the Fourier coefficients a_n for all $n \geq 1$. For $\sigma \in \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$, the form f^σ is a normalized eigenform in $\mathcal{M}_k(N, \varepsilon^\sigma)$ (respectively, $\mathcal{S}_k(N, \varepsilon^\sigma)$) for the operators T_n for all $n \geq 1$. Moreover if f is a newform, then so is f^σ.*

REMARK 12.4.6. If f is a newform then the field K generated by the eigenvalues a_n is either totally real or CM (i.e., a totally imaginary quadratic extension of a totally real field). This follows from [**Shi1**, Proposition 3.56] (see [**Shi3**, §1] or [**Shi5**, Lemma 2]).

REMARK 12.4.7. If I is an ideal of a ring A, we say that two forms f and g in $\mathbf{M}_k(\Gamma; A)$ are congruent mod I if their images in $\mathbf{M}_k(\Gamma; A/I)$ coincide, or equivalently, if the coefficients of the associated q-expansions $\sum a_n q^n$, $\sum b_n q^n$ in $A[[q]]$ satisfy $a_n \equiv b_n \bmod I$ for all $n \geq 0$.

Most interesting is the case where A is the ring of integers of a number field and f and g are eigenforms for the action of the Hecke operators. The study of such congruences arises naturally in the context of the associated Galois representations (see Remark 12.5.5) in the recent work of Ribet [**Rib4**] and Wiles [**Wil2**]. For earlier work on the subject, see for example [**Ser2**], [**SwDy**], [**Katz1**], [**DoOh**], [**Hida1**] and [**Rib3**].

We shall now describe the natural action of the Hecke operators on some of the objects we related to modular forms in the preceding sections.

Let us first consider the case of cusp forms of weight two with respect to $\Gamma = \Gamma_1(N)$ or $\Gamma_0(N)$. Let Δ denote the corresponding semigroup $\Delta_1(N)$ or $\Delta_0(N)$ in the notation of §3.1. Let $X = \Gamma \backslash \mathfrak{H}^*$. For δ in Δ, we write Γ^δ for $\Gamma \cap \delta\Gamma\delta^{-1}$ and X^δ for $X = \Gamma^\delta \backslash \mathfrak{H}^*$. (Recall from §3.2 that Γ_δ denotes $\delta^{-1}\Gamma\delta \cap \Gamma$.) Let $\pi : X^\delta \to X$ be the canonical projection. and let $\pi_\delta : X^\delta \to X$ be the map induced by $z \mapsto \delta^{-1}z$. Recall that the isomorphism (12.1.4) identifies $\mathcal{F}_2$ with the sheaf Ω^1_X of holomorphic differentials and hence $\mathcal{S}_2(\Gamma)$ with $H^0(X, \Omega^1_X)$. The Hecke operator $\Gamma\delta\Gamma$ on the space $\mathcal{S}_2(\Gamma)$ is then given by

$$H^0(X, \Omega^1_X) \to H^0(X^\delta, \Omega^1_{X^\delta}) \to H^0(X, \Omega^1_X), \tag{12.4.2}$$

the first map being the pullback π^*, the second being the trace $\pi_{\delta,*}$ on differentials. (See [**Shi1**, (7.2.6)].)

REMARK 12.4.8. Recall that for $\Gamma = \Gamma_1(N)$ with $N > 4$, we gave in §12.3 an algebraic definition of the A-module of modular forms over A, denoted $\mathcal{M}_k(\Gamma; A)$. According to theorem 12.3.7, the natural inclusion $\mathbf{M}_k(\Gamma; A) \to \mathcal{M}_k(\Gamma; A)$ is an isomorphism under suitable hypotheses. Hence $\mathcal{M}_k(\Gamma; A)$ inherits an action of the Hecke operators.

Recall that we have already given a geometric description of the action of the operators $\langle d \rangle$ on $\mathcal{M}_k(\Gamma; A)$, and it is compatible with the one on $\mathbf{M}_k(\Gamma; A)$. One can do this also for the operators T_p, at least if p is invertible in A; see [**Katz1**, §1.11].

Now we discuss the Hecke action on the cohomology groups considered in §12.2 (see [**Shi1**, §8.3]).

Recall that in the case of weight $k = 2$, (12.2.3) and (12.2.4) related $H^1(\Gamma, \mathbf{C})$ to the space of modular forms with respect to Γ. We now consider $H^1(\Gamma, \mathbf{Z})$ and define on it an action of the abstract Hecke ring $R(\Gamma, \Delta)$. For $\delta \in \Delta$, we define an endomorphism of $H^1(\Gamma, \mathbf{Z})$ as the composite

$$H^1(\Gamma, \mathbf{Z}) \to H^1(\Gamma^\delta, \mathbf{Z}) \to H^1(\Gamma_\delta, \mathbf{Z}) \to H^1(\Gamma, \mathbf{Z}) \tag{12.4.3}$$

where the first map is restriction, the second is gotten from conjugation by δ and the last is the transfer (or trace) map. The map depends only on the double coset $\Gamma\delta\Gamma$, and the image of a class x is denoted $x|(\Gamma\delta\Gamma)$ (see [**Shi1**, §8.3], [**Hida1**, §3]). Extending linearly, we obtain the desired action of $R(\Gamma, \Delta)$.

Recall that for $k \geq 2$, we let V_k denote the Γ-module $\mathrm{Symm}_{\mathbf{C}}^{k-2}(\mathbf{C}^2)$. Now consider $M_k = \mathrm{Symm}_{\mathbf{Z}}^{k-2}(\mathbf{Z}^2)$ with its action not only of Γ, but also of Δ. For a double coset $\Gamma\delta\Gamma$ in $R(\Gamma, \Delta)$, we define an endomorphism of $H^1(\Gamma, M_k)$ by a composition generalizing (12.4.3), but let us instead give a more explicit description of the endomorphism

$$\begin{array}{ccc} H^1(\Gamma, M_k) & \to & H^1(\Gamma, M_k) \\ x & \mapsto & x|(\Gamma\delta\Gamma) \end{array} \tag{12.4.4}$$

following [**Shi1**, §8.3]. We let u be a cocycle representing x and we decompose $\Gamma\alpha\Gamma$ as a disjoint union of $\Gamma\delta_i$ with $i = 1, \ldots, r$. Now for $\gamma \in \Gamma$ and for each i, we have $\delta_i \gamma \delta_{j(i)}^{-1} \in \Gamma$ for some $j(i)$. We then define a map $v : \Gamma \to M_k$ by

$$v(\gamma) = \sum_{i=1}^{r} \delta_i^\iota \cdot u(\delta_i \gamma \delta_{j(i)}^{-1}),$$

where ι is the main involution of $M_2(\mathbf{Z})$, i.e., the anti-involution defined by $\beta^\iota + \beta = (\mathrm{tr}\,\beta)I$. Then v is a cocycle and its cohomology class depends only x and the double coset $\Gamma\delta\Gamma$; we define $x|(\Gamma\delta\Gamma)$ to be this class.

One finds that the action of the double cosets extends linearly to define an action of the Hecke ring $R(\Gamma, \Delta)$, and that $H_p^1(\Gamma, M_k)$ is preserved by $R(\Gamma, \Delta)$. (See §12.2 for the definition of the parabolic cohomology groups $H_p^1(\Gamma, M_k)$.)

REMARK 12.4.9. In the situations where the group cohomology can be identified with a cohomology group for the modular curve, the double coset operator has a description analogous to (12.4.3). (See [**Hida1**, §3].)

In particular, for $k = 2$ we may identify (12.4.3) with the composite

$$H^1(Y, \mathbf{Z}) \xrightarrow{\pi^*} H^1(Y^\delta, \mathbf{Z}) \xrightarrow{\pi_{\delta,*}} H^1(Y, \mathbf{Z}) \tag{12.4.5}$$

where $Y = \Gamma\backslash\mathfrak{H}$ and $Y^\delta = \Gamma^\delta\backslash\mathfrak{H}$. This works also using the compactified curves to describe the action on $H^1_p(\Gamma, \mathbf{Z})$.

For $k > 2$, assume $\Gamma = \Gamma_1(N)$ with $N > 4$. Then the maps π and π_δ are covering maps and there is a canonical isomorphism

$$H^1(\Gamma, M_k) \cong H^1(Y, \mathbf{M})$$

where $\mathbf{M}$ is the locally constant sheaf of continuous sections $Y \to \Gamma\backslash(\mathfrak{H} \times M_k)$. Writing $\mathbf{M}^\delta$ for the corresponding sheaf on Y^δ, $x \mapsto x|(\Gamma\alpha\Gamma)$ becomes

$$H^1(Y, \mathbf{M}) \to H^1(Y^\delta, \mathbf{M}^\delta) \to H^1(Y^\delta, \pi_\delta^*\mathbf{M}^\delta) \to H^1(Y, \mathbf{M}), \tag{12.4.6}$$

where the maps are defined as follows. The first map is just π^* together with the canonical identification of $\pi^*\mathbf{M}$ with $\mathbf{M}^\delta$. The second map is given by a map on sheaves defined by δ^ι. The last is a trace. Combined with a similar construction for cohomology with compact support, one obtains also a description of the endomorphism $\Gamma\delta\Gamma$ on $H^1_p(\Gamma, M_k)$.

We can similarly define an action of $R(\Gamma, \Delta)$ on $H^1(\Gamma, V_k)$ preserving $H^1_p(\Gamma, V_k)$ (and indeed on $H^1(\Gamma, M_k \otimes A)$ for any abelian group A). The action is compatible with the canonical maps

$$\begin{array}{ccccc} H^1(\Gamma, M_k) & \to & H^1(\Gamma, M_k) \otimes \mathbf{C} & \cong & H^1(\Gamma, V_k) \\ H^1_p(\Gamma, M_k) & \to & H^1_p(\Gamma, M_k) \otimes \mathbf{C} & \cong & H^1_p(\Gamma, V_k). \end{array}$$

More importantly, we have (see [**Shi1**, Proposition 8.5])

PROPOSITION 12.4.10. *The Eichler-Shimura isomorphisms β and β_p of* Theorem 12.2.2 *are compatible with the action of $R(\Gamma, \Delta)$.*

This gives another proof, due to Shimura [**Shi1**, §3.5], of Corollary 12.4.3 in the case $k \geq 2$. Indeed $H^1(\Gamma, M_k)$ is finitely generated (using for example that Γ has a subgroup of finite index which is a finitely generated free group). Thus the image of $R(\Gamma, \Delta)$ in $\operatorname{End} H^1(\Gamma, M_k)$ is finitely generated. Now observe that $\bar{\mathbb{T}}$ is a quotient of that image.

REMARK 12.4.11. We note a variant in the case $k = 2$ which makes use of an important observation. Recall that the Jacobian $J = J_1(N)$ (respectively, $J_0(N)$) of $X = X_1(N)$ (respectively, $X_0(N)$) can be identified with $\operatorname{Hom}(W, \mathbf{C})/L$, where $W = H^0(X, \Omega^1_{X/\mathbf{C}})$ and L is the image of $H_1(X, \mathbf{Z})$ in $\operatorname{Hom}(W, \mathbf{C})$ under the canonical map defined by integration (see §10). We have explained how the modular correspondences on the curve X give rise to endomorphisms of J; in fact they define an action of $R(\Gamma, \Delta) \cong \mathbf{T}_N$ on J. This in turn defines an action of $\mathbf{T}_N$ on $\operatorname{Cot}_0(J)$, the cotangent space of J at the origin, and this space is canonically isomorphic to

$$H^0(J, \Omega^1_J) \cong H^0(X, \Omega^1_X) \cong \mathcal{S}_2(\Gamma)$$

(see §12.1). The action of $\mathbf{T}_N$ on $\mathcal{S}_2(\Gamma)$ is precisely the one we first considered in §3.4. Moreover, the map $\operatorname{End} J \to \operatorname{End}(\operatorname{Cot}_0 J)$ is injective. We may thus identify the image $\mathbb{T}$ of $\mathbf{T}_N$ in $\operatorname{End} \mathcal{S}_2(\Gamma)$ with the image of $\mathbf{T}_N$ in $\operatorname{End} J$. As the latter is finitely generated over $\mathbf{Z}$ (indeed it can be identified with a subring of $\operatorname{End} L$), so $\mathbb{T}$ is also finitely generated.

REMARK 12.4.12. The description of the Hecke action on the cohomology of modular curves extends to the adelic setting (see Remark 11.1.1). It is then natural to consider the direct limit of cohomology groups of X_U over all open compact

$U \subset G_{\mathbf{f}}$. Using coefficients in compatible systems of sheaves, the direct limit is an admissible $G_{\mathbf{f}}$-module which can be related to the ones considered in §11. See [**Del1**, §3].

We next consider the structure of some of the Hecke modules we have been discussing. We restrict our attention to the context of cusp forms, fix a weight $k > 0$ and group $\Gamma = \Gamma_0(N)$ or $\Gamma_1(N)$. We let $\mathbb{T}$ denote the image of $\bar{\mathbb{T}}$ in $\operatorname{End} \mathcal{S}_k(\Gamma)$. (Recall from Proposition 3.5.1 that if $k > 1$, this coincides with the image of $\mathbf{T}_N$, and is thus consistent with the notation of Remark 12.4.11.) Now regard $\mathbf{S}_k(\Gamma; A)$ as a module for $\mathbb{T} \otimes A$ using Proposition 12.4.1.

PROPOSITION 12.4.13. *For every A, $\mathbf{S}_k(\Gamma; A)$ is isomorphic to $\operatorname{Hom}_A(\mathbb{T}, A)$.*

This is proved in the case $A = \mathbf{Z}$ by showing that $(f, T) \mapsto a_1(f|T)$ is a perfect pairing; the general case follows on extending scalars. (See [**Shi1**, §3.5] and [**Rib2**, §2].)

Recall that we use $\overline{\mathcal{S}}_k(\Gamma)$ to denote the complex vector space $\mathcal{S}_k(\Gamma) \otimes_{\mathbf{C}} \mathbf{C}$ where the map $\mathbf{C} \to \mathbf{C}$ is complex conjugation. For $g \in \mathcal{S}_k(\Gamma)$ we write $\overline{g}$ for $g \otimes 1$ in $\overline{\mathcal{S}}_k(\Gamma)$. Then the $\mathbf{C}$-linear pairing $(f, \overline{g}) \mapsto \langle f, W_N g \rangle$ defines an isomorphism of $\mathbb{T} \otimes \mathbf{C}$-modules

$$\overline{\mathcal{S}}_k(\Gamma) \cong \operatorname{Hom}_{\mathbf{C}}(\mathcal{S}_k(\Gamma), \mathbf{C})$$

(see §4).

Moreover $\overline{\mathcal{S}}_k(\Gamma)$ is isomorphic to $\mathcal{S}_k(\Gamma)$ as a module for $\mathbb{T} \otimes \mathbf{C}$ as each is isomorphic to $\mathbf{S}_k(\Gamma; \mathbf{Z}) \otimes \mathbf{C}$. Hence $\mathcal{S}_k(\Gamma)$ is also free of rank one over $\mathbb{T} \otimes \mathbf{C}$. In fact, an explicit generator is given by

$$\sum_{M|N} \sum_{j=1}^{j_M} \iota_{N/M}^* g_j^M$$

in the notation of Remark 6.3.4.

We thus have

PROPOSITION 12.4.14. *If A is a field of characteristic 0, then $\mathbf{S}_k(\Gamma; A)$ is free of rank one over $\mathbb{T} \otimes A$. If $k \geq 2$, then $H^1_p(\Gamma, M_k) \otimes A$ is free of rank two over $\mathbb{T} \otimes A$.*

In the case $A = \mathbf{C}$, this follows from the above discussion together with the Eichler-Shimura isomorphism (Theorem 12.2.2) and its Hecke compatibility (Proposition 12.4.10). The general case then follows from that of $A = \mathbf{C}$.

We close the section with a brief discussion of the structure of the Hecke ring $\mathbb{T}$. Since $\mathbb{T} \otimes \mathbf{Q}$ is a finitely generated $\mathbf{Q}$-algebra, it canonically decomposes as the product

$$\mathbb{T} \otimes \mathbf{Q} = \prod_{\mathfrak{p}} \mathbb{T}_{\mathfrak{p}} \tag{12.4.7}$$

of its localizations at minimal prime ideals $\mathfrak{p}$. These are the primes ideals $\mathfrak{p}$ of $\mathbb{T}$ such that $\mathfrak{p} \cap \mathbf{Z} = 0$. Writing $\overline{\mathbf{Q}}$ for the field of algebraic numbers in $\mathbf{C}$, each such $\mathfrak{p}$ is the kernel of a homomorphism $\mathbb{T} \to \overline{\mathbf{Q}}$, determined up to Galois conjugacy. In turn, each such homomorphism is realized as an eigencharacter θ_f for a unique normalized eigenform f in $\mathcal{S}_k(\Gamma)$ (see [**Shi1**, Chapter 3]). Thus the factors in (12.4.7) correspond to Galois conjugacy classes of normalized eigenforms (see Proposition

12.4.5). Proposition 12.4.13 provides more information about the structure. It implies the existence of a (non-canonical) isomorphism

$$\mathbb{T} \otimes \mathbf{Q} \cong \mathrm{Hom}_{\mathbf{Q}}(\mathbb{T}, \mathbf{Q}) \tag{12.4.8}$$

and hence

$$\mathbb{T}_{\mathfrak{p}} \cong \mathrm{Hom}_{\mathbf{Q}}(\mathbb{T}_{\mathfrak{p}}, \mathbf{Q}). \tag{12.4.9}$$

Similarly $\mathbb{T} \otimes \mathbf{Z}_\ell$ is a product of local rings $(\mathbb{T} \otimes \mathbf{Z}_\ell)_{\mathfrak{m}}$ where $\mathfrak{m}$ runs through the maximal ideals of $\mathbb{T}$ containing ℓ. These maximal ideals are in one-to-one correspondence with $\mathrm{Gal}(\overline{\mathbf{F}}_\ell/\mathbf{F}_\ell)$-conjugacy classes of normalized $\mathbb{T}$-eigenforms in $\mathbf{S}_k(\Gamma; \overline{\mathbf{F}}_\ell)$. (Recall that $\mathbf{S}_k(\Gamma; \overline{\mathbf{F}}_\ell)$ is defined as $\mathbf{S}_k(\Gamma; \mathbf{Z}) \otimes \overline{\mathbf{F}}_\ell$. We thus obtain a natural action of $\mathrm{Gal}(\overline{\mathbf{F}}_\ell/\mathbf{F}_\ell)$ as well as a q-expansion map to $\overline{\mathbf{F}}_\ell[[q]]$. As usual, normalized means that the coefficient of q is 1.)

The factor $(\mathbb{T} \otimes \mathbf{Z}_\ell)_{\mathfrak{m}}$ may be identified with $\mathbb{T}_{\mathfrak{m}}$, the completion of $\mathbb{T}$ at $\mathfrak{m}$. It is a finite flat $\mathbf{Z}_\ell$-algebra and $\mathbb{T}_{\mathfrak{m}} \otimes \mathbf{Q}_\ell$ can be identified with the product of $\mathbb{T}_{\mathfrak{p}} \otimes_{\mathbf{Q}} \mathbf{Q}_\ell$ where $\mathfrak{p}$ runs over the minimal primes contained in $\mathfrak{m}$. Two minimal primes $\mathfrak{p}_1$ and $\mathfrak{p}_2$ are contained in the same $\mathfrak{m}$ if and only if the corresponding eigenforms are, up to $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ conjugacy, congruent modulo a prime over ℓ (in the sense of Remark 12.4.7).

REMARK 12.4.15. The rings $\mathbb{T}_{\mathfrak{m}}$ thus contain information about congruences between modular forms. Their structure, much finer than that of the rings $\mathbb{T}_{\mathfrak{p}}$, plays an important role in the work of Wiles [**Wil2**]. Henceforth in this remark we restrict our attention to the case $k = 2$; this is the case with which Wiles is concerned and in which the structure is best understood. Combining the q-expansion principle with properties of the Jacobian, Mazur [**Maz1**, §9,14,15] proves the analogue of (12.4.9),

$$\mathbb{T}_{\mathfrak{m}} \cong \mathrm{Hom}_{\mathbf{Z}_\ell}(\mathbb{T}_{\mathfrak{m}}, \mathbf{Z}_\ell), \tag{12.4.10}$$

under certain hypotheses. The result has since been generalized by several authors; see Remark 12.5.7 for a brief discussion of Mazur's method and the hypotheses required. The existence of such an isomorphism (12.4.10) is known to be equivalent to the ring $\mathbb{T}_{\mathfrak{m}}$ being Gorenstein. An even stronger ring-theoretic property of $\mathbb{T}_{\mathfrak{m}}$ is established by Taylor and Wiles [**TaWi**] under certain hypotheses. This stronger property, that $\mathbb{T}_{\mathfrak{m}}$ be a complete intersection, was a crucial ingredient in Wiles' proof of the Shimura-Taniyama-Weil conjecture for semistable elliptic curves.

12.5. ℓ-adic representations.

PRIMARY REFERENCES:

[**Shi1**, Chapter 7], [**Del1**], [**Ser3**, Part I] and [**Cara**, §0].

In this subsection we discuss how Galois representations are attached to modular forms.

Let k and N be positive integers. Let f be an element of $\mathcal{S}_k(\Gamma_1(N))$ which is a normalized eigenform for the Hecke operators in $\mathbf{T}_N$. Recall that this is the ring generated by the operators T_p for all primes p, and S_p for all primes p not dividing N. Let K be a number field containing K_f (the field generated by the eigenvalues of these Hecke operators acting on f), and let $\mathcal{O}$ be its ring of integers. Let ε be the Nebentypus character, and write θ for the eigencharacter $\mathbf{T}_N \to K$ defined by the action on $\mathbf{C}f$, i.e.,

$$\begin{aligned} T_p &\mapsto a_p(f) \\ S_p &\mapsto p^{k-2}\varepsilon(p). \end{aligned}$$

A construction due to Shimura [**Shi1**, Chapter 7] for $k = 2$, Deligne [**Del1**] for $k > 2$, and Deligne and Serre [**DeSe**] for $k = 1$, attaches to f a certain compatible family of ℓ-adic Galois representations. This family consists of representations

$$\rho_\lambda : \mathrm{Gal}\,(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_2(K_\lambda) \tag{12.5.1}$$

indexed by the primes λ of K. Each ρ_λ is characterized up to isomorphism by the following

- ρ_λ is continuous and unramified at primes p not dividing $N\ell$ where ℓ is the rational prime which λ divides;
- for each prime p not dividing $N\ell$, the characteristic polynomial of $\rho_\lambda(\mathrm{Frob}_p)$ is

$$X^2 - \theta(T_p)X + p\theta(S_p). \tag{12.5.2}$$

The determinant of ρ_λ is thus $\varepsilon\chi_\ell^{k-1}$, where χ_ℓ denotes the ℓth cyclotomic character. (We have used ε to denote the finite order character of $\mathrm{Gal}\,(\overline{\mathbf{Q}}/\mathbf{Q})$ corresponding to the Dirichlet character ε.) In particular, ρ_λ is *odd* in the sense that $\det \rho_\lambda(c) = -1$ for any complex conjugation c.

REMARK 12.5.1. Our convention here is that Frob_p is an arithmetic Frobenius element at p. To obtain such an element, choose a preimage in $\sigma_p \in \mathrm{Gal}\,(\overline{\mathbf{Q}}_p/\mathbf{Q}_p)$ of the Frobenius automorphism of the residue field $\overline{\mathbf{F}}_p$. Now choose an embedding $\overline{\mathbf{Q}} \to \overline{\mathbf{Q}}_p$ and let Frob_p be the image of σ_p under the inclusion

$$\mathrm{Gal}\,(\overline{\mathbf{Q}}_p/\mathbf{Q}_p) \to \mathrm{Gal}\,(\overline{\mathbf{Q}}/\mathbf{Q}). \tag{12.5.3}$$

The conjugacy class of $\rho_\lambda(\mathrm{Frob}_p)$ is independent of the choice of such an element.

REMARK 12.5.2. The term "compatible" refers to the fact that for primes p not dividing N, the characteristic polynomial of $\rho_\lambda(\mathrm{Frob}_p)$ for λ is independent of the prime λ not dividing p. (See [**Del3**, §9].)

REMARK 12.5.3. The representation depends only on the newform associated to f, and thus can be viewed as arising from the corresponding automorphic representation.

Using the continuity of ρ_λ and the compactness of $\mathrm{Gal}\,(\overline{\mathbf{Q}}/\mathbf{Q})$ we find that there is a lattice in K_λ^2 stable under the action of $\mathrm{Gal}\,(\overline{\mathbf{Q}}/\mathbf{Q})$. This lattice yields a representation

$$\mathrm{Gal}\,(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_2(\mathcal{O}_\lambda);$$

reducing mod λ, we obtain

$$\mathrm{Gal}\,(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_2(\overline{\mathbf{F}})$$

where $\overline{\mathbf{F}}$ is an algebraic closure of $\mathcal{O}/\lambda$. The isomorphism class of the representation so defined is not necessarily independent of the choice of lattice. However its semi-simplification is independent of the choice, and we denote it $\bar{\rho}_\lambda$. It may be characterized as the unique continuous semisimple representation unramified outside $N\ell$ such that for all $p \not| N\ell$, the characteristic polynomial of $\bar{\rho}_\lambda(\mathrm{Frob}_p)$ is given by (12.5.2) mod λ.

REMARK 12.5.4. The representations ρ_λ are known to be irreducible [**Rib1**, Theorem 2.3], but ρ_λ may be reducible. One is particularly interested in the irreducible ρ_λ. Serre has conjectured [**Ser4**] that all continuous, odd, irreducible representations

$$\mathrm{Gal}\,(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_2(\overline{\mathbf{F}}_\ell)$$

arise from modular forms by the above construction. See H. Darmon's article in this volume.

REMARK 12.5.5. Consider two eigenforms f_1 and f_2 as above, with weights k_1 and k_2 and levels N_1 and N_2, and let λ be a prime of a field K containing $K_{f_1}K_{f_2}$. From the above characterization of $\bar{\rho}_\lambda$, we see that the associated representations $\rho_{1,\lambda}$ and $\rho_{2,\lambda}$ are isomorphic if and only if $a_n(f_1) \equiv a_n(f_2) \mod \lambda$ for all integers n relatively prime to $N_1N_2\ell$.

Now we briefly explain the construction of the representations ρ_λ in the case $k = 2$ (see [**Shi1**, §7.6]). The construction proceeds by considering the Jacobian $J_1(N)$ of the modular curve $X_1(N)$. Let $J_1(N)[\ell^n]$ denote the kernel of multiplication by ℓ^n in $J_1(N)$, and let $\mathrm{Ta}_\ell(J_1(N))$ denote the ℓ-adic Tate module of $J_1(N)$, i.e.,

$$\varprojlim_n J_1(N)[\ell^n]$$

where the maps used to define the inverse limit are multiplication by ℓ. Then $\mathrm{Ta}_\ell(J_1(N))$ is a free $\mathbf{Z}_\ell$-module of rank $2g$ where g is the genus of $X_1(N)$. The action of $\mathbf{T}_N$ on $J_1(N)$ induces an action of $\mathbf{T}_N$ on $\mathrm{Ta}_\ell(J_1(N))$. Moreover the action factors through $\mathbb{T}$, which acts faithfully on $J_1(N)$ and hence on $\mathrm{Ta}_\ell(J_1(N))$ (see Remark 12.4.11). One checks also that

$$W_\ell(J_1(N)) = \mathrm{Ta}_\ell(J_1(N)) \otimes_{\mathbf{Z}_\ell} \mathbf{Q}_\ell$$

is free of rank two over $\mathbb{T} \otimes \mathbf{Q}_\ell$. Indeed this is a variant of Proposition 12.4.14 provided by the canonical isomorphism between $H_1(X_1(N), \mathbf{Z}_\ell)$ and $\mathrm{Ta}_\ell(J_1(N))$.

Next we consider the Galois action on $J_1(N)$. Recall that $X_1(N)$ has a canonical model over $\mathbf{Z}[1/N]$ which we denoted $\mathcal{X}_1(N)$. Its Jacobian $\mathcal{J} = \mathcal{J}_1(N)_{\mathbf{Z}[1/N]}$ is an abelian scheme over $\mathbf{Z}[1/N]$, and is a model for $J_1(N)$. We may thus identify $J_1(N)[\ell^n]$ with $\mathcal{J}(\overline{\mathbf{Q}})[\ell^n]$ and obtain an action of $\overline{\mathbf{Q}}$ on $J_1(N)[\ell^n]$, hence on $\mathrm{Ta}_\ell(J_1(N))$ and hence on $W_\ell(J_1(N))$. Moreover, the existence of models for T_p and $\langle q \rangle$ as endomorphisms of $\mathcal{J}$ shows that the action of $\mathrm{Gal}\,(\overline{\mathbf{Q}}/\mathbf{Q})$ on $W_\ell(J_1(N))$ is compatible with that of $\mathbb{T}$.

We are now ready to define ρ_λ as the representation on the $K_\lambda[\mathrm{Gal}\,(\overline{\mathbf{Q}}/\mathbf{Q})]$-module

$$W_\ell(J_1(N)) \otimes_{\mathbb{T}\otimes\mathbf{Q}_\ell} K_\lambda,$$

where the map $\mathbb{T} \otimes \mathbf{Q}_\ell \to K_\lambda$ is defined by the eigencharacter θ.

To see that ρ_λ has the desired properties, we use the Eichler-Shimura relation. If p is a prime not dividing N, then $\mathcal{J}_{\mathbf{Q}}$ has good reduction at p. Moreover we can consider the finite flat group scheme $\mathcal{J}[\ell^n]_{\mathbf{Z}_p}$, the kernel of multiplication by ℓ^n on $\mathcal{J}_{\mathbf{Z}_p}$. If $p \neq \ell$, then this finite flat group scheme over $\mathbf{Z}_p$ is etale, and the natural maps

$$\mathcal{J}[\ell^n](\mathbf{Q}_p^{\mathrm{unr}}) \leftarrow \mathcal{J}[\ell^n](\mathbf{Z}_p^{\mathrm{unr}}) \to \mathcal{J}[\ell^n](\overline{\mathbf{F}}_p)$$

are isomorphisms ([**SeTa**, Lemma 2]). The isomorphisms respect the action of $\mathrm{Gal}\,(\overline{\mathbf{Q}}_p/\mathbf{Q}_p)$, which factors through that of $\mathrm{Gal}\,(\overline{\mathbf{F}}_p/\mathbf{F}_p)$. Moreover the isomorphisms and Galois action are compatible with the action of the Hecke operators.

Now recall the Eichler-Shimura relation (10.2.3)

$$\mathcal{T}_{p,\mathbf{F}_p} = \mathrm{Frob} + \langle p\rangle_{\mathbf{F}_p,*}\mathrm{Ver}, \tag{12.5.4}$$

which we presented as an identity of endomorphisms of $\mathcal{J}_{\mathbf{F}_p}$. Since the endomorphism Frob induces Frob_p on points, and since $\mathrm{Ver}\,\mathrm{Frob} = p$, we obtain the equation

$$T_p = \mathrm{Frob}_p + \langle p\rangle p\mathrm{Frob}_p^{-1}$$

on $\mathcal{J}[\ell^n](\overline{\mathbf{F}}_p) \cong \mathcal{J}[\ell^n](\overline{\mathbf{Q}}_p)$. The equation

$$\mathrm{Frob}_p^2 - T_p\mathrm{Frob}_p + p\langle p\rangle$$

follows, and then so does the formula

$$\rho_\lambda(\mathrm{Frob}_p)^2 - \theta(T_p)\rho_\lambda(\mathrm{Frob}_p) + p\theta(S_p).$$

One can use the Weil pairing to show that Frob_p and $\langle p\rangle p\mathrm{Frob}_p^{-1}$ have the same trace and deduce that this in fact the characteristic polynomial.

REMARK 12.5.6. As a variant (see [**Shi4**, Theorem 1]), we could let $\mathfrak{p} = \ker\theta$ and consider the quotient $A = J_1(N)/\mathfrak{p}J_1(N)$. Then A is an abelian variety defined over $\mathbf{Q}$ and the action of $\mathbb{T}$ on $J_1(N)$ induces one of $\mathbb{T}/\mathfrak{p}$ on A. Let $K = K_f$ and identify $\mathbb{T}/\mathfrak{p} \otimes \mathbf{Q}$ with K via θ. Then we find that the dimension of A is $[K : \mathbf{Q}]$. Moreover $\mathrm{Ta}_\ell(A)\otimes_{\mathbf{Z}_\ell}\mathbf{Q}_\ell$ is free of rank two over $K\otimes\mathbf{Q}_\ell$. Thus it can be written as a product over the primes λ of K dividing ℓ. The factors, viewed as $K_\lambda[\mathrm{Gal}\,(\overline{\mathbf{Q}}/\mathbf{Q})]$-modules, give rise to the representations ρ_λ.

Note that in the case that f has rational coefficients, K equals $\mathbf{Q}$ and A is an elliptic curve.

REMARK 12.5.7. We can now say a little more about the hypotheses and proof for (12.4.10). First assume that ℓ does not divide $2N$, and that $\bar\rho_\lambda$ is irreducible. Let $\mathbf{m}$ be the the preimage of λ under $\theta : \mathbb{T} \to \mathcal{O}$, and let $\mathbf{F} = \mathbb{T}/\mathbf{m}$. Consider $J_1(N)[\mathbf{m}]$, where $[\mathbf{m}]$ denotes the intersection of the kernels of elements of $\mathbf{m}$. An analysis of the action of $\mathbf{F}[\mathrm{Gal}\,(\overline{\mathbf{Q}}/\mathbf{Q})]$ shows that $J_1(N)[\mathbf{m}]$ is a direct sum of copies of a model over $\mathbf{F}$ for ρ_λ. (See [**BLRi**].)

In short, Mazur's argument in [**Maz1**, §14] uses Dieudonné theory to compare $\mathcal{J}_{\mathbf{F}_\ell}[\ell]$ with $\mathcal{S}_2(\Gamma_1(N);\mathbf{F}_\ell)$. One obtains

$$\dim_{\mathbf{F}}\mathcal{S}_2(\Gamma_1(N);\mathbf{F}_\ell)\otimes_{\mathbb{T}}\mathbf{F} = \frac{1}{2}\dim_{\mathbf{F}}J_1(N)[\mathbf{m}] = \dim_{\mathbf{F}}\mathcal{S}_2(\Gamma_1(N);\mathbf{F}_\ell)[\mathbf{m}],$$

which by the q-expansion principle is one (see Proposition 12.4.13). Applying Nakayama's lemma, one deduces that $\mathcal{S}_2(\Gamma_1(N);\mathbf{Z}_\ell)_{\mathbf{m}}$ is free over $\mathbb{T}_{\mathbf{m}}$, and (12.4.10) follows. This proves also the "multiplicity one" result that $J_1(N)[\mathbf{m}]$ is a model over $\mathbf{F}$ for ρ_λ. It gives also an integral version of Proposition 12.4.14, namely that $H_1(X_1(N),\mathbf{Z})_{\mathbf{m}}$ is free of rank two over $\mathbb{T}_{\mathbf{m}}$.

Though technically more difficult, the generalizations to many cases in which ℓ divides $2N$ are based on the same principle (see [**Edi2**, §9] and [**Wil2**, §2.1]).

REMARK 12.5.8. For weight $k > 2$, the representations ρ_λ were constructed by Deligne [**Del1**] using ℓ-adic cohomology groups in the place of W_ℓ. The definition of the ℓ-adic sheaf used, call it $\mathbb{V}_k$, mirrors that of the sheaves $\mathcal{V}_k$ and $\mathbf{V}_k$ which appeared in §12.2. Very roughly speaking, $\mathbb{V}_k$ (respectively, $\mathcal{V}_k$, $\mathbf{V}_k$) comes from Symm^{k-2} of the ℓ-adic Tate module, (respectively, de Rham complex, singular cohomology) of the universal elliptic curve.

REMARK 12.5.9. The case of $k = 1$ is of a somewhat different nature. Deligne and Serre [**DeSe**] (see also [**Ser3**, §3]) construct a representation using the representations associated to congruent eigenforms of higher weight. However in this case, one actually obtains a continuous, odd, irreducible representation

$$\rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_2(\mathbf{C}),$$

unramified outside N. The characteristic polynomial of $\rho(\mathrm{Frob}_p)$ for p not dividing N is then $X^2 - \theta(T_p)X + \theta(\langle p \rangle)$ (cf. (12.5.2)). The image of ρ is finite and a basis can be chosen so that the image of ρ is in $\mathrm{GL}_2(\mathcal{O})$. (So there is again a "compatible family" of ρ_λ, each being ρ itself.)

An important feature of the case $k = 1$ is the existence of converse results due to Langlands [**Lngl2**] and Tunnell [**Tunn**]. If ρ is as above, and its image is solvable, then ρ arises from a cusp form of weight one.

REMARK 12.5.10. The Ramanujan-Petersson conjecture (see Remarks 5.0.1 and 11.5.2) follows from the fact that the roots of (12.5.2) have absolute value $p^{(k-1)/2}$. In the case $k = 2$, this follows from the Eichler-Shimura theory on applying the Weil conjectures to the abelian variety $\mathcal{J}_1(N)_{\mathbf{F}_p}$ (see [**Shi1**, Theorem 7.12]). For $k > 2$, one uses Deligne's cohomological version of the Weil conjectures (see [**Del1**, §5]) and for $k = 1$, one uses that ρ has finite image, [**Ser3**, §5].

From now on, let us assume that f is a newform (see Remark 12.5.3).

Let p be a prime not dividing $N\ell$ and consider the the restriction of ρ_λ to a decomposition group D_p, meaning the image of an embedding as in (12.5.3). This restriction is completely determined by $\rho_\lambda(\mathrm{Frob}_p)$, whose characteristic polynomial is determined by the eigenvalues of S_p and T_p. We may view this relationship as an equality of local factors of L-functions. The local factor at p of the L-function attached to the representation ρ_λ is defined as

$$\det(I - \rho_\lambda(\mathrm{Frob}_p)p^{-s})^{-1}, \tag{12.5.5}$$

(see [**Del3**, §9]), and this is the same as the local factor at p of $L(f, s)$. Moreover in the case $k = 2$, this is related to the L-function of the abelian variety A in Remark 12.5.6, see [**Shi1**, §7.5].

REMARK 12.5.11. Note that the characteristic polynomial of $\rho_\lambda(\mathrm{Frob}_p)$ only determines the semisimplification of $\rho_\lambda|_{D_p}$. In the case $k = 1$, the restriction is semisimple as its image is finite. In the case $k = 2$, the restriction is semisimple because this is known, by work of Tate (see [**Mil1**, Theorem 5.1]), to hold in general for the representation of $\mathrm{Gal}(\overline{\mathbf{F}}_p/\mathbf{F}_p)$ on the ℓ-adic Tate module of an abelian variety over $\mathbf{F}_p$. It is not known whether the representations $\rho_\lambda|_{D_p}$ are semisimple for $k > 2$ (where p is a prime not dividing $N\ell$).

Let π be the automorphic representation corresponding to f. Recall that each local factor π_p is determined by the eigenvalues of T_p and S_p, provided p does not divide the conductor N (see Example 11.2.5). The representation $\rho_\lambda|_{D_p}$ is thus related to the representation π_p. It is via such a relationship that one also describes the representations $\rho_\lambda|_{D_p}$ at ramified primes $p \neq \ell$. (See [**Cara**, §0.5], [**PSh1**] and the discussion at the end of §4 of [**Lngl1**].) This relationship is expressed by the local Langlands correspondence; the proof of the existence of this correspondence was completed by Kutzko, [**Kutz**]. The local Langlands correspondence is a bijection between irreducible, admissible representations of $\mathrm{GL}_2(\mathbf{Q}_p)$ and two-dimensional F-semisimple complex representations of the Weil-Deligne group at

p. (See [**Tate2**] and [**Del3**, §8] for the definitions of the Weil-Deligne group and F-semi-simplicity; see [**Kudla**] and [**Kna3**] for further discussion of the local Langlands correspondence.) If the representation of the Weil-Deligne group is defined over K, one can then associate a continuous representation

$$\mathrm{Gal}\,(\overline{\mathbf{Q}}_p/\mathbf{Q}_p) \to \mathrm{GL}_2(K_\lambda).$$

We shall only vaguely describe the local Langlands correspondence by saying that it respects L and ϵ factors, the L-factor of the Galois representation being as in (12.5.5) (but restricted to the coinvariants under inertia).

REMARK 12.5.12. With our choices of conventions for the automorphic representations in §11 and Galois representations above, we are implicitly choosing different conventions for the local Langlands correspondence than usually used in the literature.

REMARK 12.5.13. The analogue of the local Langlands correspondence in the context of GL_1 is provided by local class field theory. Moreover, the central character of an irreducible, admissible representation of $\mathrm{GL}_2(\mathbf{Q}_p)$ corresponds via class-field theory to the determinant of the corresponding Galois representation. If χ is a character of $\mathbf{Q}_p^\times$ with values in $K^\times$, we will write χ^A for the character $D_p \cong \mathrm{Gal}\,(\overline{\mathbf{Q}}_p/\mathbf{Q}_p) \to K_\lambda$ corresponding to χ via local class field theory.

Given a newform f, a rational prime p and a prime λ of K not dividing p, the local Langlands correspondence associates (via the factor π_p of the automorphic representation) a continuous representation

$$\mathrm{Gal}\,(\overline{\mathbf{Q}}_p/\mathbf{Q}_p) \to \mathrm{GL}_2(K_\lambda).$$

Work of Deligne, Langlands [**Lngl1**, §7] and Carayol [**Cara**] establishes that this representation is isomorphic to the F-semisimplification of $\rho_\lambda|_{D_p}$. The result, [**Cara**, Théorème (A)], has the following corollary.

THEOREM 12.5.14. *For each prime λ not dividing p,*

- *the Artin conductor of $\rho_\lambda|_{D_p}$ is the power of p dividing N;*
- *the Euler factor at p of $L(f,s)$ coincides with*

$$L(\rho_\lambda|_{D_p}, s).$$

The first assertion follows from the fact that the local Langlands correspondence respects conductors, the second from its compatibility with the formation of L-functions. (A similar statement holds for ϵ-factors.)

We now discuss the meaning of the Deligne-Langlands-Carayol theorem in specific cases.

- If π_p is the principal series $\pi(\mu_1|\ |^{1/2}, \mu_2|\ |^{1/2})$, then the semisimplification of $\rho_\lambda|_{D_p}$ is isomorphic to $\mu_1^A \oplus \mu_2^A$ (extending scalars if necessary).
- If π_p is the special representation $\mathrm{sp}(\chi|\ |^{1/2}, \chi|\ |^{-1/2})$, then $\rho_\lambda|_{D_p}$ is isomorphic to $\chi^A \otimes \sigma$ where σ can be characterized up to isomorphism as the ramified representation of the form

$$\begin{pmatrix} \chi_\ell & * \\ 0 & 1 \end{pmatrix}.$$

- If π_p is supercuspidal, then we will only remark that $\rho_\lambda|_{D_p}$ is irreducible.

Note that if p does not divide N, the description of $\rho_\lambda|_{D_p}$ is just a reformulation of the fact that the representation is unramified and that $\rho_\lambda(\mathrm{Frob}_p)$ has characteristic polynomial (12.5.2).

Now we examine the situation when $k = 2$ and the conductor of f is divisible by p but not p^2 (see [**Cas1**]). We have seen then that there are two types of possibilities for π_p. We give an indication of the proof in each case that the representation $\rho_\lambda|_{D_p}$ is as described by the local Langlands correspondence.

If the central character of π is unramified at p, then $\pi_p \cong \mathrm{sp}(\chi|\ |^{1/2}, \chi|\ |^{-1/2})$ for some unramified character χ of finite order. One can then apply the results of Deligne-Rapoport and Raynaud discussed in §10.3. In particular, the abelian variety A of Remark 12.5.6 must have multiplicative reduction at p as it is a subquotient of A_1/A_2 in the notation of Theorem 10.3.1. An analysis of the action of $\mathrm{Gal}(\overline{\mathbf{F}}_p/\mathbf{F}_p)$ on the character group of the torus T in (10.3.1) shows that $\mathrm{Frob}_p = p\langle p\rangle^{-1}T_p$ on $T(\overline{\mathbf{F}}_p)$. Applying general results about abelian varieties with multiplicative reduction ([**Ray1**] or [**Mum2**]), one deduces that the invariants under inertia at p of $\mathrm{Ta}_\ell(A) \otimes_{\mathbf{Z}_\ell} \mathbf{Q}_\ell$, viewed as a $(K_{f,\ell})[\mathrm{Gal}(\overline{\mathbf{Q}}_p/\mathbf{Q}_p)]$-module, are free of rank one over $K_{f,\ell}$ and that Frob_p acts via $p\theta(\langle p\rangle^{-1}T_p)$. Combined with the knowledge of the determinant and the formula $\theta(T_p^2) = \theta(\langle p\rangle)$, it follows that $\rho_\lambda|_{D_p}$ has the desired form.

On the other hand, suppose that the central character is ramified at p. Then $\pi_p = \pi(\mu_1|\ |^{1/2}, \mu_2|\ |^{1/2})$ with μ_1 unramified and μ_2 of conductor p. In this case, the abelian variety A is a subquotient of $J_1(Np)/A_2$ and acquires good reduction over $\mathbf{Q}_p(\zeta_p)$. One deduces from this that $(\rho_\lambda)|_{D_p}$ is a sum of two characters, each of conductor dividing p. One knows also that the determinant is as predicted, so it suffices to identify one of these characters as the unramified character μ_1^A. This was carried out using the methods described by Langlands in [**Lngl1**].

REMARK 12.5.15. The restriction of ρ_λ to a decomposition group D_ℓ is more difficult to describe. If $k = 2$ and ℓ does not divide N, then $\rho_\lambda|_{D_\ell}$ arises from an ℓ-divisible group over $\mathbf{Z}_\ell$. This follows from the fact that $\mathcal{J}_1(N)_{\mathbf{Q}}$ has good reduction at ℓ, and hence its ℓ-divisible group extends to one over $\mathbf{Z}_\ell$.

For arbitrary k and N, if $\theta(T_\ell)$ is a unit mod λ, then $\rho_\lambda|_{D_\ell}$ is "ordinary" in the sense that it is of the form

$$\begin{pmatrix} \chi_1 & * \\ 0 & \chi_2 \end{pmatrix}$$

where χ_2 is unramified (see [**Wil1**, Theorem 2.2]). Moreover $\chi_2(\mathrm{Frob}_\ell)$ is $\theta(T_\ell)$ if ℓ divides N, and is the unit root of the polynomial

$$X^2 - \theta(T_\ell)X + \ell\theta(S_\ell)$$

if ℓ does not divide N.

13. Shimura-Taniyama-Weil Conjecture

PRIMARY REFERENCES:
[**SDBi**], [**Maz2**], [**Kna2**, Chapter XIII] and [**Wil2**].

Given an elliptic curve E over $\mathbf{Q}$, we say that it is *modular* if there is a non-constant map $\mathcal{X}_0(N)_{\mathbf{Q}} \longrightarrow E$ for some positive integer N.

The Shimura-Taniyama-Weil conjecture asserts

CONJECTURE 13.0.1. *Every elliptic curve E defined over $\mathbf{Q}$ is modular.*

Through work of Wiles and Taylor [**Wil2**], [**TaWi**], [**Diam**], this is now known for a large class of elliptic curves, including all those with semistable reduction at the primes 3 and 5. As their methods and results are discussed elsewhere in this volume, we content ourselves here with a discussion of a number of equivalent conditions for E to be modular. Before listing them in Theorem 13.0.5, we recall in the form of remarks several definitions and results, some of them discussed earlier in the paper.

REMARK 13.0.2. Let E be an elliptic curve defined over $\mathbf{Q}$, and let N_E denote its conductor. For each prime p let $A_p = p + 1 - B_p$ where B_p is the number of projective solutions over $\mathbf{F}_p$ of the minimal Weierstrass equation for E. Let $\varepsilon(p) = 1$ or 0 according to whether or not E has good reduction at p. The Hasse-Weil L-function $L(E, s)$ is defined by the Euler product [**Sil2**, §II.10]

$$\prod_p (1 - A_p p^{-s} + \varepsilon(p) p^{1-2s})^{-1}.$$

The local factor at p, $L_p(E, s)$ can be described as follows:

- if E has good reduction at p, then $B_p = \#\mathcal{E}(\mathbf{F}_p)$ where $\mathcal{E}$ is the Néron model of E over $\mathbf{Z}$, p does not divide N_E and $L_p(E, s) = (1 - A_p p^{-s} + p^{1-2s})^{-1}$;
- if E has split (respectively, non-split) multiplicative reduction at p, then $p \| N_E$ and $L_p(E, s) = (1 - p^{-s})^{-1}$ (respectively, $(1 + p^{-s})^{-1}$);
- if E has additive reduction at p, then $p^2 | N_E$ and $L_p(E, s) = 1$.

For primes $\ell \neq p$, the local factor $L_p(E, s)$ coincides with $L_p(\rho_{E,\ell}, s)$ where $\rho_{E,\ell}$ is the representation of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ on $\mathrm{Ta}_\ell(E) \otimes_{\mathbf{Z}_\ell} \mathbf{Q}_\ell$. In particular, for primes p not dividing $N_E \ell$, $\rho_{E,\ell}(\mathrm{Frob}_p)$ has characteristic polynomial

$$X^2 - A_p X + p.$$

REMARK 13.0.3. Given a newform f of weight 2, level N, trivial character and rational q-expansion, we have seen how the theory of Eichler and Shimura associates to f an elliptic curve over $\mathbf{Q}$. This is the elliptic curve denoted A in Remark 12.5.6, and it may be regarded as a quotient of $\mathcal{J}_0(N)_{\mathbf{Q}}$, as well as of $\mathcal{J}_1(N)_{\mathbf{Q}}$, via the natural maps

$$\mathcal{J}_0(N)_{\mathbf{Q}} \to \mathcal{J}_1(N)_{\mathbf{Q}} \to A.$$

Writing $\rho_{f,\ell}$ for the representation denoted ρ_λ in (12.5.1), we have $\rho_{f,\ell} \cong \rho_{E,\ell}$ for all ℓ by construction. Moreover by the Deligne-Langlands-Carayol Theorem 12.5.14, we have $L(s, E) = L(s, f)$ and $N_E = N$.

REMARK 13.0.4. Suppose $E/\mathbf{Q}$ is a modular elliptic curve with a nonconstant morphism $\varphi : \mathcal{X}_0(N)_{\mathbf{Q}} \longrightarrow E$ where $\varphi(i\infty) = O$. Then E has good reduction at primes p not dividing N. Let ω be a Néron differential for E, i.e., one of the two generators of $H^0(\mathcal{E}, \Omega^1_{\mathcal{E}}) \cong \mathbf{Z}$ where $\mathcal{E}$ is the Néron model of E over $\mathbf{Z}$. Its pullback $\varphi^*\omega$ defines an element h of $\mathcal{S}_2(\Gamma_0(N))$. One can deduce from the Eichler-Shimura relation that h is a $\mathbf{T}^{(N)}$-eigenform with eigenvalues $\lambda_p \in \mathbf{Z}$ of T_p satisfying $\lambda_p = A_p$ for all p not dividing N. (See [**SDBi**, §3].)

We now list several equivalent conditions for an elliptic curve over $\mathbf{Q}$ to be modular.

THEOREM 13.0.5. *For an elliptic curve E over $\mathbf{Q}$ of conductor N_E, the following are equivalent.*

$\mathbf{X}_w$ *There exist a positive integer N and a non-constant holomorphic mapping*

$$X_1(N) \longrightarrow E(\mathbf{C}).$$

$\mathbf{J}_w$ *There exist a positive integer N and a non-constant holomorphic mapping*

$$J_1(N) \longrightarrow E(\mathbf{C}).$$

$\mathbf{R}_w$ *There exist positive integers N and D, and a $\mathbf{T}^{(ND)}$-eigenform f in $\mathcal{S}_2(\Gamma_1(N))$ with coefficients in a number field K such that*

$$\rho_{E,\ell} \otimes_{\mathbf{Q}_\ell} K_\lambda \cong \rho_{f,\lambda},$$

for some prime λ of K.

$\mathbf{L}_w$ *There exist positive integers N and D and a $\mathbf{T}^{(ND)}$-eigenform f in $\mathcal{S}_2(\Gamma_1(N))$ such that*

$$L_p(s, f) = L_p(s, E),$$

for all primes p not dividing ND.

$\mathbf{X}_s$ *There exists a surjective morphism*

$$\mathcal{X}_0(N_E)_{\mathbf{Q}} \longrightarrow E$$

of curves over $\mathbf{Q}$.

$\mathbf{J}_s$ *There exists a surjective homomorphism*

$$\mathcal{J}_0(N_E)_{\mathbf{Q}} \longrightarrow E$$

of abelian varieties over $\mathbf{Q}$.

$\mathbf{R}_s$ *There exists a newform f in $\mathbf{S}_2(\Gamma_0(N_E); \mathbf{Z})$ such that*

$$\rho_{E,\ell} \cong \rho_{f,\ell}$$

for all primes ℓ.

$\mathbf{L}_s$ *There exists a newform f in $\mathbf{S}_2(\Gamma_0(N_E); \mathbf{Z})$ such that*

$$L(s, f) = L(s, E).$$

In each case, it is clear that the strong assertion (s) implies the corresponding weak assertion (w). We discuss the remaining equivalences.

If $\mathbf{X}_s$ holds for E, then Albanese functoriality (see §10.1) defines a surjective morphism of Jacobians. Conversely, if $\mathbf{J}_s$ holds, one chooses a basepoint to define a map $i : \mathcal{X}_0(N)_{\mathbf{Q}} \to \mathcal{J}_0(N)_{\mathbf{Q}}$ and checks that the composite with $\mathcal{J}_0(N)_{\mathbf{Q}} \to E$ is nonconstant and hence surjective. The equivalence between $\mathbf{J}_w$ and $\mathbf{X}_w$ is similar.

If $\mathbf{L}_s$ holds, then for p not dividing $N_E\ell$, the characteristic polynomials of the images of Frob_p coincide under $\rho_{E,\ell}$ and $\rho_{f,\ell}$ (see (12.5.2) and Remark 13.0.2). Applying the Cebotarev density theorem and continuity of the representations, it follows that the characteristic polynomials coincide for all elements of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$. Then $\mathbf{R}_s$ follows from the irreducibility of the representations. The proof that $\mathbf{L}_w$ implies $\mathbf{R}_w$ is similar. Moreover the converse holds, and we may replace "some prime λ" with "all primes λ" in the statement of $\mathbf{R}_w$.

By Remark 13.0.4, $\mathbf{X}_s$ provides a $\mathbf{T}^{(ND)}$-eigenform h for which $\mathbf{L}_w$ is satisfied. Replacing h with the associated newform f and applying the results of Deligne, Langlands and Carayol (see Remark 13.0.3), we find that $\mathbf{R}_w$ implies $L(E, s) = L(f, s)$. Moreover f has trivial character, conductor N_E and integer Fourier coefficients, so we conclude that $\mathbf{L}_s$ holds.

We now have

$$\begin{array}{ccccccc} \mathbf{J}_{\mathrm{s}} & \Leftrightarrow & \mathbf{X}_{\mathrm{s}} & \Rightarrow & \mathbf{X}_{\mathrm{w}} & \Leftrightarrow & \mathbf{J}_{\mathrm{w}} \\ \Downarrow & & & & & & \\ \mathbf{R}_{\mathrm{s}} & \Leftrightarrow & \mathbf{L}_{\mathrm{s}} & \Leftrightarrow & \mathbf{L}_{\mathrm{w}} & \Leftrightarrow & \mathbf{R}_{\mathrm{w}}. \end{array}$$

That $\mathbf{R}_{\mathrm{s}}$ implies $\mathbf{J}_{\mathrm{s}}$ follows from Faltings' isogeny theorem [**Falt**, §5, Corollary 2]. Indeed E is isogenous to the elliptic curve A associated to the newform f by Eichler-Shimura (see Remark 13.0.3).

Finally, we sketch the proof that $\mathbf{X}_{\mathrm{w}}$ implies $\mathbf{L}_{\mathrm{w}}$ (see also [**Maz2**], especially the appendix). If E has complex multiplication, then $\mathbf{L}_{\mathrm{w}}$ is known by work of Deuring [**Deur**] and Hecke (see [**Shi2**]). So we may assume that E does not have complex multiplication. One shows first that the map $X_1(N) \to E(\mathbf{C})$ is algebraic and in fact defined over some number field F. We thus obtain a surjective map $\mathcal{J}_1(N)_F \to E_F$, and hence $A_F \to E_F$ where A is the abelian variety associated (by a construction as in Remark 12.5.6) to some $\mathbf{T}^{(N)}$-eigenform $f = \sum a_n q^n$ in $\mathcal{S}_2(\Gamma_1(N))$. Replacing N by a divisor, we can assume that f is a newform. We now have (for any ℓ) a surjection

$$\mathrm{Ta}_\ell(A) \otimes_{\mathbf{Z}_\ell} \overline{\mathbf{Q}}_\ell \to \mathrm{Ta}_\ell(E) \otimes_{\mathbf{Z}_\ell} \overline{\mathbf{Q}}_\ell$$

of $\overline{\mathbf{Q}}_\ell[\mathrm{Gal}\,(\overline{F}/F)]$-modules. The representation of $\mathrm{Gal}\,(\overline{\mathbf{Q}}/\mathbf{Q})$ on $\mathrm{Ta}_\ell(A) \otimes_{\mathbf{Z}_\ell} \overline{\mathbf{Q}}_\ell$ decomposes as a direct sum of $\rho_{f,\lambda} \otimes_{K_\lambda} \overline{\mathbf{Q}}_\ell$, indexed by pairs (λ, ι) where λ is a prime of $K = K_f$ over ℓ and ι is an embedding $K_\lambda \hookrightarrow \overline{\mathbf{Q}}_\ell$. Using that E does not have complex multiplication, one finds that $\rho_{E,\ell} \otimes_{\mathbf{Q}_\ell} \overline{\mathbf{Q}}_\ell$ restricted to $\mathrm{Gal}\,(\overline{F}/F)$ is irreducible. We then deduce that the restrictions to $\mathrm{Gal}\,(\overline{F}/F)$ of $\rho_{E,\ell} \otimes_{\mathbf{Q}_\ell} \overline{\mathbf{Q}}_\ell$ and $\rho_{f,\lambda} \otimes_{K_\lambda} \overline{\mathbf{Q}}_\ell$ are isomorphic (for some λ and ι). One next shows that the representations of $\mathrm{Gal}\,(\overline{\mathbf{Q}}/\mathbf{Q})$ are isomorphic, but with $\rho_{f,\lambda}$ replaced by a twist by some finite order character $\chi : \mathrm{Gal}\,(\overline{\mathbf{Q}}/\mathbf{Q}) \to \overline{\mathbf{Q}}_\ell^\times$. Thus for all but finitely many primes p,

$$L_p(E,s) = \left(1 - \chi(p) a_p p^{-s} + \chi(p)^2 \varepsilon(p) p^{1-2s}\right)^{-1}$$

where we use χ also to denote the corresponding Dirichlet character (which in fact takes values in $K^\times$). Now $\mathbf{L}_{\mathrm{w}}$ follows from the fact that $\sum \chi(n) a_n q^n$ is a $\mathbf{T}^{(N'D)}$-eigenform in $\mathcal{S}_2(\Gamma_1(N'))$ for some N' and D; see [**Shi1**, Proposition 3.64].

References

CONFERENCE PROCEEDINGS

[Ant1] W. Kuijk (ed.), *Modular forms of one variable I,* Lecture Notes in Math. **320**, Springer-Verlag, Berlin, Heidelberg and New York, 1973.

[Ant2] P. Deligne, W. Kuyk (eds.), *Modular forms of one variable II,* Lecture Notes in Math. **349**, Springer-Verlag, Berlin, Heidelberg and New York, 1973.

[Ant3] W. Kuyk, J.-P. Serre (eds.), *Modular forms of one variable III,* Lecture Notes in Math. **350**, Springer-Verlag, Berlin, Heidelberg and New York, 1973.

[Ant4] B. Birch, W. Kuyk (eds.), *Modular forms of one variable IV,* Lecture Notes in Math. **476**, Springer-Verlag, Berlin, Heidelberg and New York, 1975.

[Bonn1] J.-P. Serre, D. Zagier (eds.), *Modular functions of one variable, V,* Lecture Notes in Math. **601**, Springer-Verlag, Berlin, Heidelberg and New York, 1977.

[Bonn2] J.-P. Serre, D. Zagier (eds.), *Modular functions of one variable, VI,* Lecture Notes in Math. **627**, Springer-Verlag, Berlin, Heidelberg and New York, 1977.

[Boul] A. Borel, G. Mostow (eds.), *Algebraic groups and discontinuous groups,* Proc. Symp. Pure Math. **9**, Amer. Math. Soc., Providence, 1966.

[Cor1] A. Borel, W. Casselman (eds.), *Automorphic forms, representations and L-functions,* Proc. Symp. Pure Math. **33**, Part 1, Amer. Math. Soc., Providence, 1979.

[Cor2] A. Borel, W. Casselman (eds.), *Automorphic forms, representations and L-functions,* Proc. Symp. Pure Math. **33**, Part 2, Amer. Math. Soc., Providence, 1979.

[CoSi] G. Cornell and J. H. Silverman (eds.), *Arithmetic geometry,* Springer-Verlag, Berlin, Heidelberg and New York, 1986.

[Durh] A. Fröhlich (ed.), *Algebraic number fields (L-functions and Galois properties),* Academic Press, New York, 1977.

[Seat] U. Jannsen, et al (eds.), *Motives,* Proc. Symp. Pure Math. **55**, Amer. Math. Soc., Providence, 1994.

BOOKS AND ARTICLES

[Artin] M. Artin, *Néron models,* Chapter VIII of [**CoSi**], pp. 213–230.

[AtLe] A. O. L. Atkin and J. Lehner, *Hecke operators on* $\Gamma_0(m)$, Math. Ann. **185** (1970), 134–160.

[AtLi] A. O. L. Atkin and W. W. Li, *Twists of newforms and pseudo-eigenvalues of W-operators,* Inv. Math. **48** (1978), 221–243.

[Birch] B. Birch, *Some calculations of modular relations,* in [**Ant1**], pp. 175–186.

[Borel] A. Borel, *Introduction to automorphic forms,* in [**Boul**], pp. 199–210.

[BoJa] A. Borel and H. Jacquet, *Automorphic forms and automorphic representations,* in [**Cor1**], pp. 189–202.

[BLRa] S. Bosch, W. Lütkebohmert and M. Raynaud, *Néron models,* Ergebnisse der Math., 3.Folge **21**, Springer-Verlag, Berlin, Heidelberg and New York, 1990.

[BLRi] N. Boston, H. Lenstra, K. Ribet, *Quotients of group rings arising from two-dimensional representations,* C. R. Acad. Sci. Paris, Série I **312** (1991), 323–328.

[Cara] H. Carayol, *Sur les représentations ℓ-adiques associées aux formes modulaires de Hilbert,* Ann. Sci. E. N. S. **19** (1986), 409–468.

[Cart] P. Cartier, *Representations of p-adic groups: A survey,* in [**Cor1**], pp. 111–156.

[Cas1] W. Casselman, *On representations of* GL_2 *and the arithmetic of modular curves,* in [**Ant2**], pp. 107–141.

[Cas2] W. Casselman, *On some results of Atkin and Lehner,* Math. Ann. **201** (1973), 301–313.

[Cas3] W. Casselman, GL_n, in [**Durh**], pp. 663–704.

[Del1] P. Deligne, *Formes modulaires et représentations ℓ-adiques,* Séminaire Bourbaki, no. 355, Lecture Notes in Math. **179**, Springer-Verlag, Berlin, Heidelberg and New York, 1971, pp. 139–172.

[Del2] P. Deligne, *Travaux de Shimura,* Séminaire Bourbaki, no. 389, Lecture Notes in Math. **244**, Springer-Verlag, Berlin, Heidelberg and New York, 1971, pp. 123–165.

[Del3] P. Deligne, *Les constantes locales des équations fonctionnelles des fonctions L,* in [**Ant2**], pp. 501–595.

[Del4] P. Deligne, *Courbes elliptiques: formulaire (d'aprés J. Tate),* in [**Ant4**], pp. 53–73.

[Del5] P. Deligne, *La conjecture de Weil. I,* Publ. Math. IHES **43** (1974), 273–307.

[DeRa] P. Deligne and M. Rapoport, *Les schémas de modules de courbes elliptiques,* in [**Ant2**], pp. 143–316.

[DeSe] P. Deligne, J.-P. Serre, *Formes modulaires de poids* 1, Ann. Sci. Ec. Norm. Sup. **7** (1974), 507–530.

[Deur] M. Deuring, *Die Zetafunktion einer algebraischen Kurve vom Geschechte Eins I–IV,* Nachr. Akad. Wiss. Göttingen (1953), 85–94, (1955), 13–52, (1956), 37–76, (1957), 55–80.

[Diam] F. Diamond, *On deformation rings and Hecke rings,* preprint.

[DoOh] K. Doi, M. Ohta, *On some congruences between cusp forms on* $\Gamma_0(N)$, in [**Bonn2**], pp. 91–105.

[Drin] V. G. Drinfeld, *Elliptic modules,* Math. Sbornik (Russian) **94** (1974), 594–627. (English translation: Math. USSR, Sbornik **23**, No. 4 (1974), 561–592.

[Edi1] B. Edixhoven, *L'action de l'algèbre de Hecke sur les groupes de composantes des jacobiennes des courbes modulaires est "Eisenstein,"* Astérisque **196–197** (1991), 159–170.

[Edi2] B. Edixhoven, *The weight in Serre's conjectures on modular forms,* Inv. Math. **109** (1992), 563–594.

[Eich] M. Eichler, *Quaternäre quadratische Formen und die Riemannsche Vermutung für die Kongruenzzetafunktion,* Arch. Math. **5** (1954), 355–366.

[Falt] G. Faltings, *Endlichkeitssätze für abelsche Varietäten über Zahlkörpern,* Inv. Math. **73** (1983), 349–366. (English translation: Chapter II of [**CoSi**], pp. 9–27.)

[FaCh] G. Faltings and C.-L. Chai, *Degeneration of abelian varieties,* Ergebnisse der Math., 3.Folge **22**, Springer-Verlag, Berlin, Heidelberg and New York, 1990.

[Flath] D. Flath, *Decomposition of representations into tensor products,* in [**Cor1**], pp. 179–183.

[Gelb] S. Gelbart, *Automorphic forms on adele groups,* Ann. Math. Stud. **83**, Princeton Univ. Press, Princeton, 1975.

[GeJa] S. Gelbart and H. Jacquet, *Forms of* GL_2 *from the analytic point of view,* in [**Cor1**], pp. 213–252.

[GGPS] I. Gelfand, M. Graev and I. Piatetski-Shapiro, *Representation theory and automorphic functions,* Saunders, Philadelphia, 1969.

[Gode] R. Godement, *Notes on Jacquet-Langlands theory,* Institute for Advanced Study, Princeton, 1970.

[Gross] B. H. Gross, *A tameness criterion for Galois representations associated to modular forms mod p,* Duke Math. J. **61** (1990), 445–517.

[Gro1] A. Grothendieck, *Fondements de la géométrie algébrique,* Séminaire Bourbaki, no. 232, W. A. Benjamin, New York, 1966.

[Gro2] A. Grothendieck, *Séminaire de géométrie algébrique 7, I,* Lecture Notes in Math. **288**, Springer-Verlag, Berlin, Heidelberg and New York, 1972.

[HaCh] Harish-Chandra, *Representations of semisimple Lie groups, I,* Trans. AMS **75** (1953), 185–243.

[Hart] R. Hartshorne, *Algebraic geometry,* Graduate Texts in Math. **52**, Springer-Verlag, Berlin, Heidelberg and New York, 1977.

[Hecke] E. Hecke, *Mathematische Werke,* Vandenhoeck and Ruprecht, Göttingen, 1959.

[Hec1] E. Hecke, *Theorie der Eisensteinschen Reihen höherer Stufe und ihre Anwendung auf Funktionentheorie und Arithmetik,* Abh. Math. Sem. Hamburg **5** (1927), 199–224. (No. 24 in [**Hecke**], pp. 461–486.)

[Hec2] E. Hecke, *Über die Bestimmung Dirichletscher Reihen durch ihre Funktionalgleichung,* Math. Ann. **112** (1936), 664–699. (No. 33 in [**Hecke**], pp. 591–626.)

[Hec3] E. Hecke, *Über Modulfunktionen und die Dirichletschen Reihen mit Eulerscher Produktentwicklung. I, II,* Math. Ann. **114** (1937) 1–28, 316–351. (No. 35,36 in [**Hecke**], pp. 644–707.)

[Hers] I. N. Herstein, *Topics in algebra,* John Wiley and Sons, New York, 1964.

[Hida1] H. Hida, *Congruences of cusp forms and special values of their zeta functions,* Inv. Math. **63** (1981), 225–261.

[Hida2] H. Hida, *Galois representations into* $\mathrm{GL}_2(\mathbf{Z}_p[[X]])$ *attached to ordinary cusp forms,* Inv. Math. **85** (1986), 545–613.

[Hida3] H. Hida, *Elementary theory of L-functions and Eisenstein series,* London Math. Soc. Student Texts **26**, Cambridge Univ. Press, Cambridge, 1993.

[Huse] D. Husemöller, *Elliptic curves,* Graduate Texts in Math. **111**, Springer-Verlag, Berlin, Heidelberg and New York, 1986.
[Huxl] M. N. Huxley, *Scattering matrices for congruence subgroups,* in *Modular Forms,* R. Rankin (ed.), John Wiley and Sons, New York, 1984, pp. 141–156.
[Igu1] J.-I. Igusa, *Fibre systems of Jacobian varieties, III: Fibre systems of elliptic curves,* Amer. J. Math. **81**, 453–476 (1959).
[Igu2] J.-I. Igusa, *Kroneckerian model of fields of elliptic modular functions,* Amer. J. Math. **81**, 561–577 (1959).
[Igu3] J.-I. Igusa, *On the algebraic theory of elliptic modular functions,* J. Math. Soc. Japan **20** (1968), 96–106.
[Ihara] Y. Ihara, *On modular curves over finite fields,* in *Proceedings of the international colloquium on discrete subgroups of Lie groups and applications to moduli,* Bombay, 1973, pp. 161–202.
[JaLa] H. Jacquet, R. P. Langlands *Automorphic forms on* GL_2, Lecture Notes in Math. **114**, Springer-Verlag, Berlin, Heidelberg and New York, 1970.
[Katz1] N. Katz, *p-adic properties of modular schemes and modular forms,* in [**Ant3**], pp. 70–189.
[Katz2] N. Katz, *p-adic interpolation of real analytic Eisenstein series,* Ann. Math. **104** (1976), 459–571.
[KaMa] N. Katz and B. Mazur, *Arithmetic moduli of elliptic curves,* Ann. Math. Studies **108**, Princeton Univ. Press, Princeton, 1985.
[Kna1] A. W. Knapp, *Representations of* $GL_2(\mathbf{R})$ *and* $GL_2(\mathbf{C})$, in [**Cor1**], pp. 87–91.
[Kna2] A. W. Knapp, *Elliptic curves,* Princeton Math. Notes **40**, Princeton Univ. Press, Princeton, 1993.
[Kna3] A. W. Knapp, *The local Langlands correspondence: the Archimedean case,* in [**Seat**], Part 2, pp. 393–410.
[Kobl] N. Koblitz, *Introduction to elliptic curves and modular forms,* Graduate Texts in Math. **97**, Springer-Verlag, Berlin, Heidelberg and New York, 1984.
[Koike] M. Koike, *Congruences between cusp froms and linear representations of the Galois group,* Nagoya Math. J., **64** (1976), 63–85.
[Kubo] T. Kubota, *The elementary theory of Eisenstein series,* John Wiley and Sons, New York, 1973.
[Kudla] S. Kudla, *The local Langlands correspondence: the non-Archimedean case,* in [**Seat**], Part 2, pp. 365–391.
[Kutz] P. Kutzko, *The local Langlands conjecture for* GL(2) *of a local field,* Ann. Math. **112** (1980), 381–412.
[Lang1] S. Lang, *Algebraic number theory,* Addison-Wesley, Reading, MA, 1970.
[Lang2] S. Lang, *Introduction to modular forms,* Grundl. Math. Wiss. **222**, Springer-Verlag, Berlin, Heidelberg and New York, 1976.
[Lngl1] R. P. Langlands, *Modular forms and ℓ-adic representations,* in [**Ant2**], pp. 361–500.
[Lngl2] R. P. Langlands, *Base change for* GL(2), Ann. Math. Stud. **96**, Princeton Univ. Press, Princeton, 1980.
[LiOe] S. Ling, J. Oesterlé, *The Shimura subgroup of* $J_0(N)$, Astérisque **196–197** (1991), 171–203.
[Maass] H. Maass, *Über eine neue Art von nicht analytischen automorphen Funktionen und die Bestimmung von Dirichlet Reihen durch funktionale Gleichungen,* Math. Ann. **121** (1949), 141–183.
[Maz1] B. Mazur, *Modular curves and the Eisenstein ideal,* Publ. Math. I.H.E.S. **47** (1977), 33–186.
[Maz2] B. Mazur, *Number theory as gadfly,* Amer. Math. Monthly **98** (1991), 593–610.
[MaWi] B. Mazur and A. Wiles, *Class fields of abelian extensions of* $\mathbf{Q}$, Inv. Math. **76** (1984), 179–330.
[Mil1] J.S. Milne, *Points on Shimura varieties mod p,* in [**Cor2**], pp. 165–184.
[Mil2] J.S. Milne, *Abelian varieties,* Chapter V of [**CoSi**], pp. 103–150.
[Mil3] J.S. Milne, *Abelian varieties,* Chapter VII of [**CoSi**], pp. 167–212.
[Miy1] T. Miyake, *On automorphic forms on* GL_2 *and Hecke operators,* Ann. Math. **94** (1971), 174–189.
[Miy2] T. Miyake, *Modular forms,* Springer-Verlag, Berlin, Heidelberg and New York, 1989.
[Mum1] D. Mumford, *Abelian varieties,* Oxford Univ. Press, Oxford, 1970.

[Mum2] D. Mumford, *An analytic construction of degenerating abelian varieties over complete rings*, **24**, Fasc. 3 (1972), 239–272. (Also appendix to [**FaCh**].)
[Neron] A. Néron, *Modèles minimaux des variétés abéliennes sur les corps locaux et globaux*, Publ. Math. I.H.E.S. **21** (1964), 5–128.
[Ogg] A. Ogg, *Modular forms and Dirichlet series*, W. A. Benjamin, New York, 1969.
[PSh1] I. Piatetski-Shapiro, *Zeta-functions of modular curves*, in [**Ant2**], pp. 317–360.
[PSh2] I. Piatetski-Shapiro, *Multiplicity one theorems*, in [**Cor1**], pp. 209–212.
[Ray1] M. Raynaud, *Variétés abéliennes et géométrie rigide*, in *Actes, Congr. Int. Math., Vol. 1*, 1970, pp. 473–477.
[Ray2] M. Raynaud, *Spécialization du foncteur de Picard*, Publ. Math. I.H.E.S. **38** (1971), 27–76.
[Ray3] M. Raynaud, *Jacobienne des courbes modulaires et opérateurs de Hecke*, Astérisque **196–197** (1991), 9–25.
[Rib1] K. Ribet, *Galois representations attached to eigenforms with Nebentypus*, in [**Bonn2**], pp. 17–52.
[Rib2] K. Ribet, *Mod p Hecke operators and congruences between modular forms*, Inv. Math. **71** (1983), 193–205.
[Rib3] K. Ribet, *Congruence relations between modular forms*, in *Proc. Int. Congr. Math.*, 1983, pp. 503–514.
[Rib4] K. Ribet, *On modular representations of* $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ *arising from modular forms*, Inv. Math. **100** (1990), 431–476.
[Roga] J. D. Rogawski, *Modular forms, the Ramanujan conjecture and the Jacquet-Langlands correspondence*, appendix to: A. Lubotzky, *Discrete groups, expanding graphs and invariant measures*, Progress in Math. **125** Birkhäuser, Boston, 1994, pp. 135–176.
[Rosen] M. Rosen, *Abelian varieties over* $\mathbf{C}$, Chapter IV of [**CoSi**], pp. 79–101.
[Sata] I. Satake, *Spherical functions and Ramanujan conjecture*, in [**Boul**], pp. 258–264.
[Schl] A. Scholl, *Modular forms and de Rham cohomology; Atkin-Swinnerton-Dyer congruences*, Inv. Math. **79** (1985), 49–77.
[Schn] B. Schoeneberg, *Elliptic modular functions*, Grundl. Math. Wiss. **203**, Springer-Verlag, Berlin, Heidelberg and New York, 1974.
[Ser1] J.-P. Serre, *A course in arithmetic*, Graduate Texts in Math. **7**, Springer-Verlag, Berlin, Heidelberg and New York, 1973.
[Ser2] J.-P. Serre, *Congruences et formes modulaires*, Séminaire Bourbaki, no. 416, Lecture Notes in Math. **317** Springer-Verlag, Berlin, Heidelberg and New York, 1973, pp. 319–338.
[Ser3] J.-P. Serre, *Modular forms of weight one and Galois representations*, in [**Durh**], pp. 193–268.
[Ser4] J.-P. Serre, *Sur les représentations modulaires de degré 2 de* $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$, Duke Math. J. **54** (1987), 179–230.
[SeTa] J.-P. Serre, J. Tate, *Good reduction of abelian varieties*, Ann. Math. **88** (1968), 492–517.
[Shatz] S. Shatz, *Group schemes, formal groups and p-divisible groups*, Chapter III of [**CoSi**], pp. 29–78.
[Shi1] G. Shimura, *Introduction to the arithmetic theory of automorphic functions*, Iwanami Shoten and Princeton Univ. Press, Princeton, 1971.
[Shi2] G. Shimura, *On elliptic curves with complex multiplication as factors of the jacobians of modular function fields*, Nagoya Math. J., **43** (1971), 199–208.
[Shi3] G. Shimura, *Class fields over real quadratic fields and Hecke operators*, Ann. Math. **95** (1972), 130–190.
[Shi4] G. Shimura, *On the factors of the jacobian variety of a modular function field*, J. Math. Soc. Japan **25**, No. 3 (1973), 523–544.
[Shi5] G. Shimura, *The special values of the zeta functions associated with cusp forms*, Comm. Pure Appl. Math. **29** (1976), 783–804.
[Shi6] G. Shimura, *The special values of zeta functions associated with Hilbert modular forms*, Duke Math. J. **45** (1978), 637–679.
[Shi7] G. Shimura, *On the Eisenstein series of Hilbert modular groups*, Rev. Mat. Iberoamer. **1**, No. 3 (1985), 1–42.
[Shi8] G. Shimura, *Yutaka Taniyama and his time*, Bull. London Math. Soc. **21** (1989), 186–196.

[Sil1] J. H. Silverman, *The arithmetic of elliptic curves,* Graduate Texts in Math. **106**, Springer-Verlag, Berlin, Heidelberg and New York, 1986.

[Sil2] J. H. Silverman, *Advanced topics in the arithmetic of elliptic curves,* Graduate Texts in Math. **151**, Springer-Verlag, Berlin, Heidelberg and New York, 1994.

[SwDy] H. P. F. Swinnerton-Dyer, *On ℓ-adic representations and congruences for coefficients of modular forms,* in [**Ant3**], pp. 1–57.

[SDBi] H. P. F. Swinnerton-Dyer, B. Birch, *Elliptic curves and modular functions,* in [**Ant4**], pp. 2–32.

[Tate1] J. Tate, *Fourier analysis in number fields and Hecke's zeta function,* in *Algebraic number theory,* J.W.S. Cassels and A. Fröhlich (eds.), Academic Press, New York, 1968.

[Tate2] J. Tate, *Number-theoretic background,* in [**Cor2**], pp. 3–26.

[TaWi] R. Taylor and A. Wiles, *Ring theoretic properties of certain Hecke algebras,* Ann. Math. **141** (1995), 553–572.

[Tunn] J. Tunnell, *Artin's conjecture for representations of octasahedral type,* Bull. AMS (N.S.) **5** (1981), 173–175.

[Wal1] N. Wallach, *Representations of reductive Lie groups,* in [**Cor1**], pp. 71–86.

[Wal2] N. Wallach, *Real reductive groups, I,* Academic Press, New York, 1988.

[Weil] A. Weil, *Über die Bestimmung Dirichletscher Richen durch Funktionalgleichungen,* Math. Ann. **168** (1967), 149–156.

[Wil1] A. Wiles, *On ordinary λ-adic representations associated to modular forms,* Inv. Math. **94** (1988), 529–573.

[Wil2] A. Wiles, *Modular elliptic curves and Fermat's Last Theorem,* Ann. Math. **141** (1995), 443–551.

D.P.M.M.S., University of Cambridge, Cambridge CB2 1SB, United Kingdom
E-mail address: fdiamond@pmms.cam.ac.uk

Department of Mathematics, University of Toronto, Toronto, Ontario, Canada M5S 1A1
E-mail address: imj@math.toronto.edu