Modular Forms
Spring 2011 Notes (and beyond)

Kimball Martin

May 2, 2022

Preface

The first part of these notes are based on a graduate course on modular forms I gave in Spring 2011 at the University of Oklahoma. The second part of these notes, still in progress, consists of additional topics I did not cover in the course, but which I might imagine covering in a second semester.

Preface to first part

In the first part, I have tried to give an introduction to modular forms with a view towards classical applications, such as quadratic forms and functions on Riemann surfaces, as opposed to "modern applications" (in the sense of requiring a more modern perspective) such as Fermat's last theorem and the congruent number problem.

At the same time, I have tried to give a suitable introduction to lead into a one-semester course on automorphic forms and representations (in Fall 2011), which meant a slightly different balance of material than in a course wholly focusing on classical modular forms.

I also tried to keep the prerequisites as minimal as possible while attempting to meet both of these goals.

For those considering using these notes: you can get an idea of the contents by looking at the table of contents and skimming through, so I won't elaborate on them here, but just say the following points, which distinguish the presentation from some other treatments:

(1) My general philosophy is to find a balance between simplicity and completeness, focusing on what I think is important to understanding the ideas and being able to see applications, rather than favoring either generalities or minutia—of course my preferred balance may be different than yours. (2) I try to give geometric motivation to the definition of modular forms. (3) I primarily focus on modular forms on $\Gamma_0(N)$ and do not even introduce modular forms with character (nebentypus). (4) I try to be fairly explicit with arithmetic (e.g., working out Fourier expansions of Eisenstein series with level and explicit formulas for representation numbers of quadratic forms), though I don't go overboard (or even as far along the board as I wanted). (5) I finish the first part with a brief treatment of L-functions, which I will hopefully expand on later, possibly in the second part.

A warning and an apology are in order.

The warning: several of the sections were written in a rush, and may have some (hopefully not serious) errors. Please email me if you find any mistakes, so I can correct them.

The apology: due to weather and travel, we missed many lectures, and as a result there are many things missing that I would have liked to include, such as Siegel modular forms

and the structure of modular functions. Further, there are many details that I would like to have included which I had not the time for. In addition, I realize now I could have done a better job filling in some prerequisite material. Upon teaching this course again, or possibly earlier if I am inspired, I plan to revise these notes. (I have made minor revisions to the first part, but no serious ones since.) If you have any comments or suggestions, I would be happy to hear them.

I would like to thank my students, for asking many questions and pointing out many mistakes in early versions of these notes, in particular: Kumar Balasubramanian, Jeff Breeding, James Broda, Shayna Grove, Catherine Hall, Daniel McLaury, Salam Turki and Jeremy West. I am also grateful to Victor Manuel Aricheta and Roberto Miatello for pointing out additional errors.

Preface to second part

Around 2014–2015, I started thinking again to write up notes on some additional topics I did not cover in the one-semester modular forms course, e.g.: newforms, half-integral weight forms (though I didn't even do odd weight in the first part!), quaternionic modular forms, Eichler–Shimura theory, modularity of elliptic curves, Hilbert modular forms, Siegel modular forms. (Though, perhaps surprising to some, I still have no desire to cover modular forms with character—there's no accounting for taste, you know). I was contemplating teaching a second semester course in modular forms in Spring 2016, but there was no modular forms course in Fall 2015, and I instead decided to teach a course on *(Quaternion) Algebras in Number Theory*. The notes for that course should eventually discuss quaternionic modular forms and the Jacquet–Langlands correspondence, at least in simple situations.

Still, I made some progress in 2015–2016 by adding an unpolished chapter on newforms to begin the second part, which is a slow work in progress (currently in progress: a chapter on Hilbert modular forms). While I don't plan on doing a comprehensive treatment of the topics in the second part, or necessarily give complete proofs, I hope to give a more-or-less working introduction to these topics.

Contents

Part I

The basic course

Chapter 1

Introduction

There are many starting points for the theory of modular forms. They are a fundamental topic lying at the intersection of number theory, harmonic analysis and Riemann surface theory. Even from a number theory point of view, there are several ways to motivate the theory of modular forms. One is via the connection with elliptic curves, made famous through Wiles' solution to Fermat's Last Theorem. This connection has many amazing implications in number theory, but we will emphasize the role of modular forms in the theory of quadratic forms.

Let Q be a *quadratic form* of rank k over $\mathbb{Z}$. This means Q is a homogeneous polynomial of degree 2 over $\mathbb{Z}$ in k variables with a certain nondegeneracy condition. For example, Q might be a *diagonal form*

$$Q(x_1, \dots, x_k) = a_1 x_1^2 + a_2 x_2^2 + \cdots + a_k x_k^2 \tag{1.0.1}$$

where each $a_i \neq 0$; or it may be something like $Q(x, y) = x^2 + 2xy + 3y^2$. Regardless, the fundamental question about Q is

Question 1.1. *What numbers does Q represent? In other words, for which n does*

$$Q(x_1, \dots, x_k) = n$$

have a solution in $\mathbb{Z}$?

There is a more quantitative version of this question as follows.

Question 1.2. *Let $r_Q(n)$ denote the number of solutions (in $\mathbb{Z}$) to*

$$Q(x_1, \dots, x_k) = n.$$

Determine $r_Q(n)$.

Note that an answer to Question 1.2 provides an answer to Question 1.1, since Question 1.1 is simply asking, when is $r_Q(n) > 0$? (Sometimes $r_Q(n)$ is infinite, and instead one counts solutions up to some equivalence, but we will not go into this here.)

Let us consider the specific examples of forms

$$Q_k(x_1, \dots, x_k) = x_1^2 + x_2^2 + \cdots + x_k^2,$$

and write $r_{Q_k}(n)$ simply as $r_k(n)$. So Question 1.1 is simply the classical question, what numbers are sums of k squares? Lagrange proved that every positive integer can be written as a sum of four squares (and therefore ≥ 4 squares by just taking $x_i = 0$ for $i > 4$). The cases of sums of two squares and sums of three squares were answered by Fermat and Gauss. So a complete answer to Question 1.1 for the forms Q_k has been known since the time of Gauss (who did fundamental work on quadratic forms), but the answer for $k > 4$ is not so interesting, as it is trivially encoded in the answer for $k = 4$.

Hence at least for these forms, we see the question of determining $r_k(n)$ is much more interesting. Furthermore, Question 1.1 is typically much more difficult for arbitrary quadratic forms Q than it is for Q_k, and a general method for answering Question 1.2 will provide us a way to answer Question 1.1.

Now let us briefly explain how one might try to find a formula for $r_k(n)$. Jacobi considered the *theta function*

$$\vartheta(z) = \sum_{n=-\infty}^{\infty} q^{n^2}, \quad q = e^{2\pi i z}. \tag{1.0.2}$$

This function is well defined for $z \in \mathfrak{H} = \{x + iy : x, y \in \mathbb{R}, y > 0\}$. Then

$$\vartheta^2(z) = \left(\sum_{\ell=-\infty}^{\infty} q^{\ell^2}\right)\left(\sum_{m=-\infty}^{\infty} q^{m^2}\right) = \sum_{\ell,m} q^{\ell^2+m^2} = \sum_{n\geq 0} r_2(n) q^n.$$

Similarly,

$$\vartheta^k(z) = \sum_{n\geq 0} r_k(n) q^n. \tag{1.0.3}$$

(More generally, if Q is the diagonal form in (1.0.1), then we formally have $\prod_{i=1}^{k} \vartheta(a_i z) = \sum r_Q(n) q^n$.) It is not too difficult to see that ϑ^k satisfies the identities

$$\vartheta^k(z+1) = \vartheta^k(z), \quad \vartheta^k\left(\frac{-1}{4z}\right) = \left(\frac{2z}{i}\right)^{\frac{k}{2}} \vartheta^k(z). \tag{1.0.4}$$

Indeed, the first identity is obvious because q is invariant under $z \mapsto z + 1$.

The space of (holomorphic) functions on $\mathfrak{H}$ satisfying the transformation properties is (1.0.4) is defined to be the space of *modular forms* $M_{k/2}(4)$ of *weight* $k/2$ and *level* 4. The theory of modular forms will tell us that $M_{k/2}(4)$ is a finite-dimensional vector space.

For example, when $k = 4$, $M_2(4)$ is a 2-dimensional vector space, and one can find a basis in terms of *Eisenstein series*. Specifically, consider the Eisenstein series

$$G(z) = -\frac{1}{24} + \sum_{n=1}^{\infty} \sigma(n) q^n,$$

where $\sigma(n)$ is the divisor function $\sigma(n) = \sum_{d|n} d$. Then a basis of $M_2(4)$ is

$$f(z) = G(z) - 2G(2z), \quad g(z) = G(2z) - 2G(4z).$$

Hence $\vartheta^4(z)$ is a linear combination of $f(z)$ and $g(z)$. How do we determine what combination? Simply compare the first two coefficients of q^n in $af(z) + bg(z)$ with $\vartheta^4(z)$, and one sees that

$$\vartheta^4(z) = 8f(z) + 16g(z).$$

Expanding this out, one sees that

$$\vartheta^4(z) = \sum_{n\geq 0} r_4(n)q^n = 1 + 8\sum_{n\geq 1}\sigma(n)q^n - 32\sum_{n\geq 1}\sigma(n)q^{4n}. \tag{1.0.5}$$

Consequently

$$r_4(n) = \begin{cases} 8\sigma(n) & 4\nmid n \\ 8\sigma(n) - 32\sigma(n/4) & 4|n. \end{cases}$$

If one wishes, one can write this as a single formula

$$r_4(n) = 8\left(2 + (-1)^n\right)\sum_{d|n, 2\nmid d} d. \tag{1.0.6}$$

In particular, it is obvious that $r_4(n) > 0$ for all n, in other words, we have Lagrange's theorem that every positive integer is a sum of four (not necessarily nonzero) integer squares. Furthermore, we have a simple formula for the number of representations of n as a sum of four squares, in terms of the divisors of n.

In this course, we will develop the theory of modular forms, and use this to derive various formulas of the above type. Along the way, we will give some other applications of modular forms. We will also introduce the theory of L-functions, which are an important tool in the theory of modular forms, and are fundamental in the connection with elliptic curves. Time permitting, we will introduce generalizations of modular forms, such as Siegel modular forms and automorphic forms.

As much as possible, we will try to keep the prerequisites to a minimum. Certainly, working knowledge of linear algebra is expected, as well as some familiarity with groups and rings. Familiarity with elementary number theory is helpful but not necessary (modular arithmetic will be used, as well as some standard notation from elementary number theory). We may at times discuss some aspects of basic algebraic number theory, but these discussions should be sufficiently limited to not greatly affect the flow of the text if ignored.

Particularly in the beginning, we will be discussing geometric ideas, as this provides motivation for studying modular forms (why study functions on the upper-half plane $\mathfrak{H}$ satisfying seeming strange transformation laws as in (1.0.4)?). Here, familiarity with such things as Riemann surfaces, isometry groups and universal covers will be helpful, but we will develop the needed tools as we go. Some basic notions of point-set topology (open sets, continuity, etc.) will be assumed.

Perhaps most helpful will be a solid course on complex analysis (fractional linear transformations, holomorphy, meromorphy, Cauchy's integral formula, etc.) and familiarity of Fourier analysis. For those lacking (or forgetting) this analysis background, I will recall the necessary facts as we go, but the reader should refer to texts on complex analysis or Fourier analysis for further details and proofs.

There are a variety (but not a plethora) of exercises intertwined with the text. Most of them are not too difficult, and I encourage you to think about all of them, whether or not you decide it's worth your while to work out the details. I have starred certain exercises which I consider particularly important. You may find a larger selection of exercises from the references listed below.

There are many good references for the basic theory of modular forms. Unfortunately, none of them do exactly what we want, which is why I am writing my own notes. Also unfortunately, the terminology and notation among the various texts is not standarized. On the other hand, they are better at what they do than my notes (and likely with fewer errors), so you are encouraged to refer to them throughout the course.

- [Ser73] Serre, J.P. *A course in arithmetic.* A classic streamlined introduction to modular forms of level 1. Many of the details you need to work out for yourself.
- [Kob93] Koblitz, Neal. *Introduction to elliptic curves and modular forms.* A solid introduction to modular forms of both integral and half-integral weight (or arbitrary level), if slightly dense. The goal is to present them in connection with elliptic curves and show how they are used in Tunnell's solution, assuming the weak Birch–Swinnerton-Dyer conjecture, of the ancient *congruent number problem.*
- [Kil08] Kilford, L.J.P. *Modular forms: a classical and computational introduction.* A new book, and it seems like a good introduction to modular forms. Has errata online. The one thing lacking for our course is that it does not cover L-functions.
- [Zag08] Zagier, Don. *Elliptic modular forms and their applications,* in "The 1-2-3 of modular forms." A beautiful overview of modular forms (primarly level 1) and their applications. Available online.
- [Lan95] Lang, Serge. *Introduction to modular forms.* It's Serge Lang. Covers some advanced topics.
- [Apo90] Apostol, Tom. *Modular functions and Dirichlet series in number theory.* A nice classical analytic approach to modular forms.
- [Mil] Milne, J.S. *Modular functions and modular forms.* Online course notes. This treatment, somewhat like [DS05] or [CSS97], has a more geometric focus (e.g., modular curves).
- [DS05] Diamond, Fred and Shurman, Jerry. *A first course in modular forms.* An excellent book, perhaps requiring more geometric background than others, focusing on the connections of modular forms, elliptic curves, modular curves and Galois representations. Available online.
- [Iwa97] Iwaniec, Henryk. *Topics in classical automorphic forms.* An excellent and fairly elementary analytic approach using classical automorphic forms. Many interesting applications are presented.
- [Miy06] Miyake, Toshitsune. *Modular forms.* A fairly advanced presentation of the theory of modular forms, starting with automorphic forms on adeles. Contains useful material not easily found in many texts. Available online.
- [Bum97] Bump, Daniel. *Automorphic forms and representations.* A thorough (at least as much as possible in 550 pp.) text on automorphic forms and representations on $\mathrm{GL}(n)$. The first quarter of the book treats classical modular forms.

- [Ste07] Stein, William. *Modular forms: a computational approach.* Perhaps a useful reference for those wanting to do computations with modular forms.
- [CSS97] Cornell, Silverman and Stevens (ed.s). *Modular forms and Fermat's last theorem.* A nice collection of articles giving an overview of the theory of elliptic curves, modular curves, modular forms and Galois representations, and how they are used to prove Fermat's last theorem.

Chapter 2

Elliptic functions

Before we introduce modular forms, which, as explained in the introduction, are functions on the upper half-plane, or the *hyperbolic plane*, satisfying certain transformation laws, it may be helpful to get a basic understanding of *elliptic functions*, which are functions on $\mathbb{C}$ satisfying certain simpler transformation laws. That is our goal for this chapter.

Elliptic functions are a classical topic in complex analysis, and their theory can be found in several books on the subject, such as [Ahl78], [Lan99], [FB09] (available online) or [Sta09] (available online), as well as many texts on elliptic curves and modular forms (e.g., [Kob93], [Iwa97]). In fact, the complex analysis books [FB09] and [Sta09] discuss modular forms. Since elliptic functions are not a focus for this course, but rather a tool for motivation of the theory of modular forms, we will not strive for rigorous proofs, but merely conceptual understanding. Put another way, this chapter is a sort of summary of the pre-history of modular forms.

We will explain later how the theory of elliptic functions is essentially a "Euclidean version" of the theory of modular forms, and in fact modular forms first arose from the study of elliptic functions.

2.1 Complex analysis review: Holomorphy

First let us review some basic facts from the theory of functions of one complex variable.

Let $\mathbb{C}$ denote the complex plane. Throughout these notes, z will denote a complex number, and unless stated otherwise, x and y will denote the real and imaginary parts, $x = \mathrm{Re}(z)$ and $y = \mathrm{Im}(z)$, of z. I.e., $z = x + iy$ where $x, y \in \mathbb{R}$.

Let $f : \mathbb{C} \to \mathbb{C}$. As vector spaces, we clearly have $\mathbb{C} \simeq \mathbb{R}^2$ via the linear isomorphism $z \mapsto (x, y)$, hence we may also view f as a function from $\mathbb{R}^2$ to $\mathbb{R}^2$. In fact, $\mathbb{C}$ and $\mathbb{R}^2$ are isomorphic as topological spaces, so the notion of continuity is the same whether we regard f as a function on $\mathbb{C}$ or $\mathbb{R}^2$. The conceptual difference between functions on $\mathbb{C}$ and $\mathbb{R}^2$ arises when we study the notion of differentiablilty, and the difference arises because one can multiply complex numbers, but not (naturally) elements of $\mathbb{R}^2$.

Specifically, look at the condition

$$f'(z) = \lim_{h \to 0} \frac{f(z+h) - f(z)}{h}. \tag{2.1.1}$$

If we want to think of f as truly a function of $\mathbb{C}$, this means we should be taking $h \in \mathbb{C}$ in the above limit. In other words, unlike in the real case where h can only approach 0 from the left or right, here h can approach 0 along any line or path to the origin in $\mathbb{C}$. In particular, (2.1.1) should be true for $h \in \mathbb{R}$ and $h = ik \in i\mathbb{R}$. Note, restricting to $h \in \mathbb{R}$ in the limit above amounts to differentiating with respect to x, and restricting to $h \in i\mathbb{R}$ amounts to differentiating with respect to y.

We can write $f(z)$ uniquely as $f(z) = u(z) + iv(z)$ where $u : \mathbb{C} \to \mathbb{R}$ and $v : \mathbb{C} \to \mathbb{R}$, i.e., $u = \mathrm{Re}(f)$ and $v = \mathrm{Im}(f)$. We also write $u(z) = u(x, y)$ and $v(z) = v(x, y)$. Then condition (2.1.1) taking $h \in \mathbb{R}$ means

$$\begin{aligned} f'(z) &= \lim_{h\to 0} \frac{f(x+h+iy) - f(x+iy)}{h} \\ &= \lim_{h\to 0} \frac{u(x+h,y) + iv(x+h,y) - u(x,y) - iv(x,y)}{h} \\ &= \lim_{h\to 0} \frac{u(x+h,y) - u(x,y)}{h} + i \lim_{h\to 0} \frac{v(x+h,y) - v(x,y)}{h} \\ &= \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x}. \end{aligned}$$

Similarly, taking $h = ik \in i\mathbb{R}$ in (2.1.1) means

$$\begin{aligned} f'(z) &= \lim_{k\to 0} \frac{f(x+i(y+k)) - f(x+iy)}{ik} \\ &= \lim_{k\to 0} \frac{u(x,y+k) + iv(x,y+k) - u(x,y) - iv(x,y)}{ik} \\ &= \lim_{k\to 0} \frac{u(x,y+k) - u(x,y)}{ik} + i \lim_{k\to 0} \frac{v(x+k,y) - v(x,y)}{ik} \\ &= \frac{1}{i}\left(\frac{\partial u}{\partial y} + i\frac{\partial v}{\partial y}\right) = \frac{\partial v}{\partial y} - i\frac{\partial u}{\partial y}. \end{aligned}$$

Comparing the real and imaginary parts of these 2 expressions for $f'(z)$ gives the **Cauchy–Riemann equations**

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}. \tag{2.1.2}$$

Definition 2.1.1. *Let $U \subseteq \mathbb{C}$ be an open set. We say $f : U \to \mathbb{C}$ is* **(complex) differentiable***, or* **holomorphic***,* **at** *z if the limit $\lim_{h\to 0} \frac{f(z+h)-f(z)}{h}$ exists (for $h \in \mathbb{C}$). In this case, the* **derivative** *$f'(z)$ is defined to be the value of this limit.*

We say f is **holomorphic on** *U if f is holomorphic at each $z \in U$. If f is holomorphic on all of $\mathbb{C}$, we say f is* **entire.**

As we saw above, being holomorphic on an open set U means that the Cauchy–Riemann equations will hold (and the partial derivatives will be continuous). (This is in fact if and only if.) Contrast this to differentiable functions on $\mathbb{R}^2$: if one knows the partial derivatives exist and are continuous on an open set in $\mathbb{R}^2$, the function is (real) differentiable there. The Cauchy–Riemann equations give a much stronger condition for a function to be complex differentiable.

The biggest consequence of the Cauchy–Riemann equations is that if f is holomorphic on U, so is f'. Hence being differentiable once means being infinitely differentiable. This feature makes complex analysis *much* nicer than real analysis. In particular, any differentiable function on U has a Taylor series expansion around any $z_0 \in U$:

$$f(z) = f(z_0) + f(z - z_0) + \frac{f''(z_0)}{2!}(z - z_0)^2 + \frac{f'''(z_0)}{3!}(z - z_0)^3 + \cdots .$$

Recall that any power series $\sum_{n\geq 0} a_n(z - z_0)^n$ has a radius of convergence $R \in [0, \infty]$ such that the series converges absolutely for $|z - z_0| < R$ and diverges for $|z - z_0| > R$.

Definition 2.1.2. *Let U be an open set in $\mathbb{C}$ and $f : U \to \mathbb{C}$. We say f is* **analytic** *at $z_0 \in U$ if, for some $R > 0$, we can write*

$$f(z) = \sum_{n\geq 0} a_n(z - z_0)^n, \quad |z - z_0| < R.$$

We say f is **analytic** *on U if f is analytic at each $z_0 \in U$.*

Note if we can write $f(z)$ as a power series $\sum_{n\geq 0} a_n(z - z_0)^n$ about z_0, and this series has radius of convergence R, then f is analytic on the open disc $|z - z_0| < R$.

One of the main theorems of complex analysis is

Theorem 2.1.3. *Let U be an open set of $\mathbb{C}$ and $f : U \to \mathbb{C}$. The following are equivalent.*

(i) f is holomorphic on U;

(ii) f is infinitely differentiable on U;

(iii) f if analytic on U; and

(iv) If $u = \mathrm{Re}(f)$ and $v = \mathrm{Im}(f)$, then the partial derivatives of u and v with respect to x, y exist, are continuous, and satisfy the Cauchy–Riemann equations.

From the definition, it is clear differentiation over $\mathbb{C}$ satisfies the usual differentiation rules of calculus (sum, product, quotient and chain rules; as well as the power rule and derivative formulas for trigonometric and exponential functions).

For $n \in \mathbb{Z}$, the power functions $f(z) = z^n$ are well defined and entire for $n \geq 0$, and holomorphic on the punctured plane $\mathbb{C} - \{0\}$ for $n < 0$.

Exercise 2.1.4. *Consider $f(z) = \frac{1}{z}$. Show f satisfies the Cauchy–Riemann equations. Deduce that it is analytic on its domain, but there is no single power series expansion valid for all $z \in \mathbb{C} - \{0\}$.*

One can define e^z, $\sin(z)$ and $\cos(z)$ by the usual Maclaurin series expansions

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!},$$

$$\sin(z) = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{(2n+1)!},$$

and

$$\cos(z) = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n}}{(2n)!}.$$

These series have radius of convergence ∞, and hence define entire functions.

On the other hand, the logarithm cannot be extended to an entire function. The best one can do is make it well defined on the complex plane minus a ray from the origin. This involves choosing a "branch cut." Unless otherwise specified, we will chose our branch cuts so that the complex logarithm is holomorphic function on $\mathbb{C} - \mathbb{R}_{\leq 0}$

Power functions for non-integral exponents can be defined in terms of the logarithm. Specifically, one can formally define $z^a = e^{a \log z}$. This will be holomorphic when the logarithm is, which will typically be $\mathbb{C} - \mathbb{R}_{\leq 0}$ for us.

Some basic facts about analytic functions are recorded in the following.

Theorem 2.1.5. *Let $z_0 \in \mathbb{C}$ and f be defined on a neighborhood of z_0. Suppose f is analytic at z_0. Then f is analytic in a neighborhood U of z_0. Furthermore, we have the following.*

(a) If f' is nonzero on U, then f is conformal*, i.e., f preserves angles.*

(b) If $f'(z_0) \neq 0$, then f is locally invertible at z_0, i.e., there exist a function g which is analytic near $f(z_0)$ such that $g \circ f = \mathrm{id}$ near z_0.

(c) Let S be a subset of U containing an accumulation point. If $g : U \to \mathbb{C}$ is analytic and $f(z) = g(z)$ for $z \in S$, then $f(z) = g(z)$ for all $z \in U$.

The first part of the theorem follows from the fact we already remarked after Definition 2.1.2. Part (a) says that analytic function preserve a lot of geometry. Even though they do not in general preserve distance, this property of preserving angles forces certain rigid behavior of analytic functions. For instance, conformality implies that there is no analytic function mapping an open disc to an open square. Another example of this rigid behaviour is Liouville's function, which we will state later, but geometrically says that no analytic function can map the complex plane to any bounded region. Part (c) also describes a rigidity feature: an analytic function is determined by its values on merely a countable set of points (which contains an accumulation point).

2.2 Complex analysis review II: Zeroes and poles

The most basic analytic information about an analytic function is the location of its zeroes and poles. Let's start off discussing zeroes, although there's not much to say at the moment.

Definition 2.2.1. *Let f be analytic at z_0 such that $f(z_0) = 0$. We say f has a zero of* **order** *m at z_0 if $\lim_{z \to z_0} \frac{f(z)}{(z-z_0)^m}$ exists and is nonzero, or, equivalently, if the power series expansion of f at z_0 has the form*

$$f(z) = \sum_{n=m}^{\infty} a_n (z - z_0)^n$$

with $a_m \neq 0$.

Note that zeroes of a nonconstant analytic function must be isolated, i.e., they form a discrete set. To see this, suppose f is analytic on U, and let S be the set of zeroes of f. If f is nonconstant and S has an accumulation point, then Theorem 2.1.5(c) implies $f \equiv 0$ on U. In particular, in any bounded region, an nonzero analytic function has a finite number of zeroes.

Definition 2.2.2. *Let U be a neighborhood of z_0 and f be a nonconstant analytic function on $U - \{z_0\}$. We say f has a* **pole of order** *m at z_0 if $\frac{1}{f(z)}$ has a zero of order m at z_0, or, equivalently, $\lim_{z\to z_0}(z-z_0)^m f(z)$ exists and is nonzero. In this case, we write $f(z) = \lim_{z\to z_0} f(z) = \infty$.*

If one wants to be a little more formal about assigning a value of ∞ to f, consider the **Riemann sphere** $\hat{\mathbb{C}} = \mathbb{P}^1(\mathbb{C}) = \mathbb{C} \cup \{\infty\}$. (Here the open sets of $\hat{\mathbb{C}}$ are generated by the open sets of $\mathbb{C}$ together with the balls about infinity, $\{z : |z| > \epsilon\} \cup \{\infty\}$, for $\epsilon \in \mathbb{R}_{\geq 0}$. Hence $\hat{\mathbb{C}}$ is topologically a sphere.) Even though there are infinitely many "real directions" to go off to infinity in $\mathbb{C}$ (meaning picturing $\mathbb{C}$ as $\mathbb{R}^2$), we think of them all leading to the same point, ∞, on $\hat{\mathbb{C}}$.

In fact, one can push this idea further. If $f : \mathbb{C} \to \hat{\mathbb{C}}$, one can define a value for $f(\infty)$ and think of $f : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$, and one can talk about analytic maps from the Riemann sphere to itself.

We need a term for analytic functions with poles (since functions are not called analytic, or holomorphic, at their poles).

Definition 2.2.3. *Let $U \subseteq \mathbb{C}$ be an open set and $f : U \to \hat{\mathbb{C}}$. Let $S = f^{-1}(\infty) \subseteq U$ be the set of poles of f in U. If S is discrete and f is analytic on $U - S$, we say f is* **meromorphic** *on U.*

In particular, if f and g are holomorphic functions on U and $g \not\equiv 0$, then f/g is meromorphic on U. This follows from the fact that f/g is differentiable outside of the set of zeroes of S, which we remarked above is discrete. Hence all rational functions are meromorphic on $\mathbb{C}$. Specifically, if f and g are nonzero polynomials, then the number of zeroes (counting multiplicity, i.e., summing up the orders of zeroes) of f/g is $\deg(f)$ and the number of poles (again counting with multiplicity) is $\deg(g)$.

Just like a holomorphic function has a power series expansion about any point in its domain, a meromorphic function has a *Laurent series* expansion around any point in its domain.

Suppose $f : U \to \hat{\mathbb{C}}$ is meromorphic, and let $z_0 \in \mathbb{C}$. If $f(z_0) \neq \infty$, then we just have a usual power series expansion $f(z) = \sum_{n=0}^{\infty} a_n(z-z_0)^n$ valid near z_0. Suppose instead f has a pole of order m at z_0. Then $(z-z_0)^m f(z)$ is holomorphic near z_0, so we have a power series expansion

$$(z-z_0)^m f(z) = \sum_{n=0}^{\infty} a_n(z-z_0)^n.$$

This implies, in a neighborhood of z_0, we have the following **Laurent series expansion**

$$f(z) = \sum_{n=-m}^{\infty} a_{n+m}(z-z_0)^n = \sum_{n=-m}^{\infty} b_n(z-z_0)^n,$$

where we put $b_n = a_{n+m}$ for $n \geq -m$. (A Laurent series is simply a series of the above form, i.e., a power series where the exponents are allowed to start at a finite negative number.)

Exercise 2.2.4. *Write down a Laurent series for* $\frac{z+2}{z^2(z+1)}$ *about* $z_0 = 0$.

Proposition 2.2.5. *Let* $z_0 \in \mathbb{C}$ *and* U *be a neighborhood of* $\mathbb{C}$. *Suppose* $f : U - \{z_0\} \to \mathbb{C}$ *is analytic and* $\lim_{z\to z_0} |z - z_0|^m |f(z)| = 0$ *for some* $m \in \mathbb{N}$. *By defining* $f(z_0) = \lim_{z\to z_0} f(z)$, *the extension* $f : U \to \hat{\mathbb{C}}$ *is meromorphic on* U.

There are two cases in the proof. Either $\lim_{z\to z_0} f(z)$ exists as a finite complex number or not. In fact if $f(z)$ is bounded as $z \to z_0$, one can use Cauchy's integral formula (which we will recall later) to show the limit exists and the extension of f to U is analytic at z_0. In this case we say f has a **removable singularity**.

Otherwise, $g(z) = (z - z_0)^m f(z)$ has a removable singularity, so by the above $g(z)$ can be extended to be analytic on U. If $g \equiv 0$, then $f \equiv 0$, which cannot happen since we have assumed $\lim_{z\to z_0} f(z)$ is not finite. Hence $g(z)$ has a zero of some finite order $0 < k < m$ at z_0. Then $\frac{1}{f(z)} = \frac{(z-z_0)^m}{g(z)}$ has a zero of order $m - k$ at z_0, i.e., $f(z)$ has a pole of order $m - k$ at z_0, and is by definition meromorphic.

We remark that if the condition $\lim_{z\to z_0} |z - z_0|^m |f(z)| = 0$ does not hold for some m in the above proposition, then $f(z)$ has what is called an **essential singularity**. One example is $e^{1/z}$ at $z = 0$. While we do not need this, the behavior at essential singularities is so remarkable, we would be remiss not to point it out. For the next two results we assume U is an open neighborhood of z_0 and $f : U - \{z_0\} \to \mathbb{C}$ is analytic.

Theorem 2.2.6. (Weierstrass) *Suppose* f *has an essential singularity at* z_0. *Then on any neighborhood* $V \subseteq U$ *of* z_0, *the values of* f *come arbitrarily close to any complex number, i.e.,* $\{f(z) : z \in V - \{z_0\}\}$ *is dense in* $\mathbb{C}$.

This is a rather surprising result, though it is not too difficult to prove. However, one of the most amazing theorems in complex analysis (or even all of mathematics) is that something *much* stronger is true.

Theorem 2.2.7. (Big Picard). *Suppose* f *has an essential singularity at* z_0. *Then on any neighborhood* $V \subseteq U$ *of* z_0, *the values of* f *range over all complex number with at most one possible exception, i.e.,* $|\hat{\mathbb{C}} - f(V)| = 0$ *or* 1.

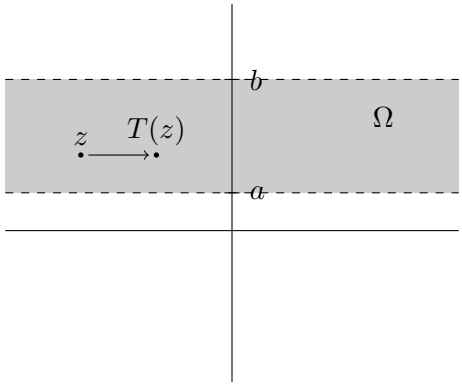
For example, since we know $e^{1/z}$ is never 0 for $z \in \mathbb{C}$, Big Picard tells us that in every neighborhood of 0, $e^{1/z}$ attains every nonzero complex value!

(Note the above theorem is called "Big Picard" because Picard has another beautiful theorem in complex analysis named after him, now called Little Picard, which is a consequence of Big Picard. We will recall Little Picard later.)

2.3 Periodic functions

Let $\omega \in \mathbb{C} - \{0\}$ and Ω a region in $\mathbb{C}$, i.e., Ω is a nonempty connected open subset of $\mathbb{C}$. We assume the map $T(z) = T_\omega(z) = z + \omega$ is an isometry of Ω. Since T preserves distance in $\mathbb{C}$, this simply means $T(\Omega) = \Omega$. For instance, if $\omega = 1$ (or any nonzero real number), we could

take Ω to be a horizontal strip $\{a < \mathrm{Im}(z) < b\}$ for some fixed a, b. This is clearly invariant under T.



(A slightly stranger looking region for Ω could be the region between the curves $y = \sin(x)$ and $y = 1 + \sin(x)$.)

We say $f : \Omega \to \mathbb{C}$ has **period** ω if $f(z + \omega) = f(z)$ for all $z \in \Omega$. Note that if Ω is not preserved by the translation T, then there are no functions with period ω on Ω because $f(z)$ and $g(z) := f(z + \omega)$ would have different domains.

Let Λ be the cyclic group of isometries of Ω generated by T, i.e.,

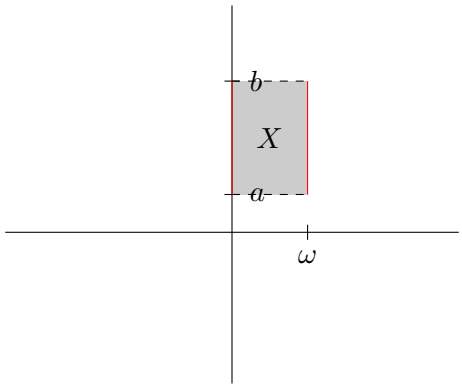
$$\Lambda = \left\{\ldots, T^{-2}, T^{-1}, I = T^0, T, T^2, \ldots\right\} = \{T_{k\omega} : k \in \mathbb{Z}\},$$

i.e., Λ is the group of all translations on Ω by an integer multiple of ω.

Any time we have a group of isometries acting on a metric space, we can consider the quotient, in this case $X = \Omega/\Lambda$. What this means is the following. Define two points of Ω to be Λ-equivalent if they differ by some element $\tau \in \Lambda$, i.e., $z \sim_\Lambda z'$ if $\tau(z) = z'$ for some $\tau \in \Lambda$. Since Λ is a group, $\sim_\Lambda$ is an equivalence relation, and we can let $[z]$ denote the Λ-equivalence class of z.

Then the quotient space $X = \Omega/\Lambda$, as a set, is defined to be the set of Λ-equivalence classes $\{[z] : z \in \Omega\}$ of point of Ω. This naturally inherits a topology from Ω, the quotient topology. Specifically, the open sets of X are of the form $\{[z] : z \in U\}$, where U is an open set of Ω.

Graphically, in the above example where $\omega = 1$ and $\Omega = \{a < \mathrm{Im}(z) < b\}$, we can picture $X = \Omega/\Lambda$ as below.



Here we identify the left and right borders of X. Topologically this is homeomorphic to $S^1 \times (a, b)$. Precisely, we have a "flat" open cylinder of diameter 1 and height $b - a$. The adjective "flat" here refers to the fact that, while X is topologically a cylinder, its geometry is Euclidean, i.e., it has no curvature.

While X is technically not a subset of Ω, let alone a specific subset of Ω as above, it is often convenient identify X with a subset of Ω (at least as a set—geometrically, we have to identify the left and right borders of the shaded region above to get X). This is done through the notion of a fundamental domain.

Definition 2.3.1. *Let $\Omega \subset \mathbb{C}$ be a region and Λ a group of isometries of Ω. We say a closed subset $\mathcal{F}$ of Ω with connected interior is a* **fundamental domain** *for Λ (or Ω/Λ) if*

(i) any $z \in \Omega$ is Λ-equivalent to some point in $\mathcal{F}$;

(ii) no two interior points of $\mathcal{F}$ are Λ-equivalent; and

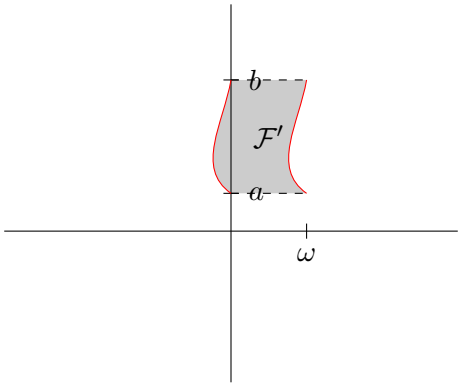
(iii) the boundary of $\mathcal{F}$ is a finite union of smooth curves in Ω.

(Note: different authors impose different conditions on fundamental domains. For some, the fundamental domain would be the interior of what we called the fundamental domain. Others may have different conditions on the shape of the boundary, or require convexity.)

So, in our above example, we can take a fundamental domain $\mathcal{F}$ to be the shaded region in the previous picture, including the boundary, i.e.,

$$\mathcal{F} = \{z \in \mathbb{C} : a < \mathrm{Im}(z) < b, 0 \leq \mathrm{Re}(z) \leq 1\} = [0, 1] \times (a, b).$$

(While this is not closed in $\mathbb{C}$, it is closed in Ω.) One could of course construct a different fundamental domain by translating $\mathcal{F}$ to the left or right. A slightly different fundamental domain $\mathcal{F}'$ is pictured below.



Now we can view a function $f : \Omega \to \mathbb{C}$ with period ω as a function of X since the value of $f(z)$ only depends upon the Λ-equivalence class of z. In particular, in our above example, we can identify continuous functions on Ω with period ω with continuous functions on X, i.e., continuous functions on the fundamental domain $\mathcal{F} = [0, 1] \times (a, b)$ such that $f(iy) = f(1 + iy)$ for $y \in (a, b)$.

Moving back to the case of a general region Ω invariant under $T = T_\omega$, there are two basic approaches to constructing periodic functions on Ω:

(1) Clearly the function $e^{2\pi iz/\omega}$ has period ω. Hence we can use this to construct other functions with period ω.

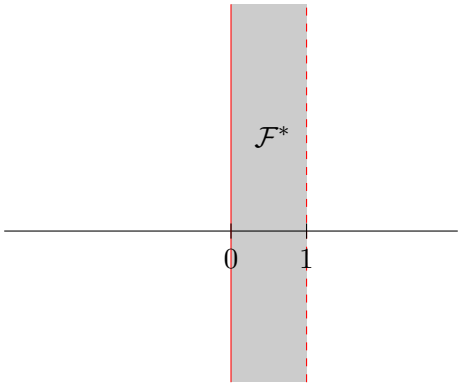
(2) We can start with a function $f(z)$, which decreases sufficiently fast, such as $f(z) = \frac{1}{z^2}$, and *average* f over all elements of Λ, e.g.,

$$\widetilde{f}(z) = \sum_{n\in\mathbb{Z}} f(z + n\omega).$$

The condition on the rate of decay of f will guarantee that $\sum f(z + n\omega)$ converges, and it is clear that $\widetilde{f}(z + \omega) = f(z)$.

For now, let's focus on (1). It turns out that we can write all (analytic) periodic functions in terms of $e^{2\pi iz/\omega}$, but approach (2) will be useful in more complicated situations.

By making a change of variable $z \mapsto z/\omega$, which transforms the domain Ω to $\omega^{-1}\Omega$, we may assume our period $\omega = 1$. For simplicity, let us assume $\Omega = \mathbb{C}$, so $X = \Omega/\Lambda = \mathbb{C}/\mathbb{Z}$. Thus, for a fundamental domain $\mathcal{F}$ for $\mathbb{C}/\mathbb{Z}$ is $\mathcal{F} = \{z \in \mathbb{C} : 0 \le \mathrm{Re}(z) \le 1\}$, and we identify the sides $\mathrm{Re}(z) = 0$ and $\mathrm{Re}(z) = 1$ to get X. In other words, X is a flat cylinder of infinite height, and X is in bijection with the subset $\mathcal{F}^* = \{0 \le \mathrm{Re}(z) < 1\}$ of $\mathbb{C}$.



Consider the image of $\mathcal{F}^*$ under the map $f(z) = e^{2\pi iz}$. Write $z \in \mathcal{F}^*$ as $z = x + iy$ where $0 \le x < 1$. Then

$$f(z) = e^{2\pi iz} = e^{-2\pi y}e^{2\pi ix} = re^{i\theta},$$

where we put $r = e^{-2\pi y}$ and $\theta = 2\pi x \bmod 2\pi$. Viewing $f : z \mapsto (r, \theta)$, it is clear that the image of $\mathcal{F}^*$ under f is $(0, \infty) \times [0, 2\pi)$. Further this map is injective. Thinking back in terms of $f : \mathcal{F}^* \to \mathbb{C}$, we see $e^{2\pi iz}$ is an analytic one-to-one map of $\mathcal{F}^*$ onto $\mathbb{C}^\times = \mathbb{C} - \{0\}$.

Suppose $g : \mathbb{C} \to \hat{\mathbb{C}}$ has period 1. Assuming g is continuous, we have a Fourier expansion

$$g(z) = \sum_{n=-\infty}^{\infty} a_n(y)e^{2\pi inz},$$

where the n-th **Fourier coefficient** $a_n(y)$ is given by

$$a_n(y) = \hat{g}(n) = \int_0^1 g(z)e^{-2\pi inz}dx, \quad (z = x + iy).$$

A priori, $a_n(y)$ is a just function of y.

However if g is meromorphic, then we have a stronger result. Put $\zeta = f(z)$. Then there is a unique meromorphic $G : \mathbb{C}^\times \to \hat{\mathbb{C}}$ such that $g(z) = G(\zeta)$. It is natural to ask when G extends to a meromorphic function of $\mathbb{C}$. By Theorem 2.2.5, this is if and only if $\lim_{\zeta\to 0} G(\zeta)|\zeta|^m = 0$ for some m. Note that $\zeta = e^{2\pi iz} \to 0$ for $z \in \mathcal{F}^*$ if and only if $\mathrm{Im}(z) \to \infty$. Hence we may restate the above condition on $G(\zeta)$ being meromorphic at 0 as $\lim_{\mathrm{Im}(z)\to\infty} g(z)|e^{2\pi imz}| \to 0$ for some m, i.e.,

$$\text{for } \mathrm{Im}(z) \text{ large, there exists } m \text{ such that } |g(z)| < e^{2\pi my}. \tag{2.3.1}$$

Since $\zeta \to 0$ corresponds to $\mathrm{Im}(z) \to \infty$, we say g is **meromorphic at** $i\infty$ if (2.3.1) is satisfied.

Assume now that g is also meromorphic at $i\infty$. Then $G(\zeta)$ has a Laurent series expansion about 0,

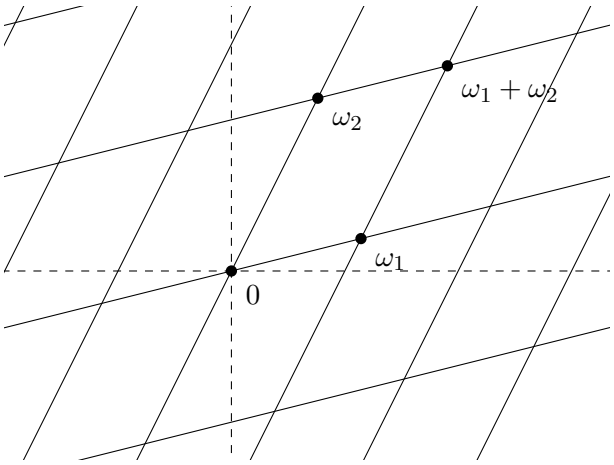
$$g(z) = G(\zeta) = \sum_{n=-m}^{\infty} c_n \zeta^n = \sum_{n=-m}^{\infty} c_n e^{2\pi inz},$$

where m is the order of the pole of G at 0. (We can take $m = 0$ if G does not have a pole at 0.) This must agree with the Fourier expansion above, so we have $a_n(y) = c_n$ for all $n \geq -m$ and $a_n(y) \equiv 0$ for $n < -m$. Hence for meromorphic periodic functions, the Fourier coefficients are constant (independent of y) and all but finitely many of the negative Fourier coefficients vanish.

These facts about the Fourier expansion will be crucial in our study of modular forms.

2.4 Doubly periodic functions

Let ω_1, ω_2 be two complex periods, i.e., two nonzero complex numbers linearly independent over $\mathbb{R}$. Then they generate a **lattice** $\Lambda = \langle \omega_1, \omega_2 \rangle = \{a\omega_1 + b\omega_2 : a, b \in \mathbb{Z}\} \subset \mathbb{C}$. We might picture Λ as "parallelogram grid" follows.



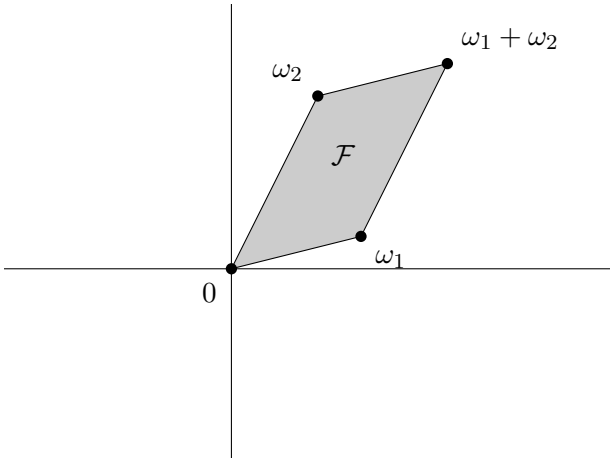
While technically the lattice Λ is only the set of vertices of the above parallelogram grid, it is convenient to draw it as above.

Definition 2.4.1. *A meromorphic function $f : \mathbb{C} \to \hat{\mathbb{C}}$ is* **elliptic** *(or* **doubly periodic***)* **with respect to** Λ *if*

$$f(z+\omega) = f(z) \quad \text{for all } z \in \mathbb{C},\ \omega \in \Lambda.$$

The space of all such functions is denoted $E(\Lambda)$.

Note $f \in E(\Lambda)$ if and only if $f(z+\omega_1) = f(z+\omega_2) = f(z)$. As with singly periodic functions, we can identify doubly periodic functions with functions on a quotient space $\mathbb{C}/\Lambda$. Note this quotient space is a (flat) torus. We can take for a fundamental domain $\mathcal{F}$ the (closed) parallelogram whose vertices are drawn above.



Then $\mathbb{C}/\Lambda$ is identified with the torus obtained by gluing together both opposite pairs of edges of the fundamental parallelogram $\mathcal{F}$.

We remark that elliptic functions were originally studied because they contain, as a special case, the inverse functions of the classical elliptic integrals (integrals involving the square root of a cubic or quartic polynomial, which arise in the problem of finding the arc length of certain elliptical shapes, such as spirals and cycloids). This is also, as you may guess, the root of the modern terminology.

We would like to be able to describe the space of meromorphic functions on $\mathbb{C}/\Lambda$. Earlier, we suggested two methods for constructing periodic functions: (1) write functions in terms of a simple periodic function you know; and (2) average functions over Λ. Since there are no obvious nonconstant doubly periodic functions, here we'll start with approach (2).

In order to get something that converges, we need to average a function that decays sufficiently fast. A first idea might be to try averaging $\frac{1}{z^2}$ over the lattice $\Lambda = \langle 1, i\rangle$. This would be

$$\sum_{\omega\in\Lambda} \frac{1}{(z+\omega)^2} = \sum_{a,b\in\mathbb{Z}} \frac{1}{(z+(a+bi)^2)}. \tag{2.4.1}$$

If this converges, it would have a pole of order 2 at each lattice point, and nowhere else. However, it does not converge (absolutely). To see this, take $z = 0$, and sum up the absolute values of all terms excluding the $\frac{1}{z^2}$ term to get

$$\sum_{(a,b)\in\mathbb{Z}^2-\{(0,0)\}} \frac{1}{|a+bi|^2} = \sum_{a,b} \frac{1}{a^2+b^2} \approx 4\int_1^\infty \int_1^\infty \frac{dx\,dy}{x^2+y^2} \approx \int_0^{2\pi}\int_1^\infty \frac{r\,dr\,d\theta}{r^2} = \infty.$$

(Here of course the approximations are certainly good enough to formally check divergence of the right most integral implies divergence of the sum on the left.)

Since (2.4.1) does not converge absolutely, the next simplest idea would be to consider

$$g(z) = \sum_{\omega \in \Lambda} \frac{1}{(z+\omega)^3}.$$

Indeed this sum converges absolutely, except of course when $z \in \Lambda$, in which case we get a pole of order 3. However, it turns out that one can modify the sum in (2.4.1) to get an elliptic function with only poles of order 2 at each lattice point. Precisely, we define **Weierstrass pe** function (with respect to Λ) to be

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

This can be shown to converge and although not as obvious as (2.4.1), to be doubly periodic (see exercises below). Given this, it is clear this has a pole of order 2 at each $z \in \Lambda$ and no poles elsewhere, so $\wp$ is an elliptic function of order 2. (The **order** of an elliptic function is the sum of the orders of its poles in $\mathbb{C}/\Lambda$.)

Since $\wp$ is analytic on $\mathbb{C} - \Lambda$, it is differentiable, and on checks that its derivative is

$$\wp'(z) = -2g(z) = -2 \sum_{\omega \in \Lambda} \frac{1}{(z+\omega)^3}, \tag{2.4.2}$$

which has order 3.

Exercise 2.4.2. *Show $\wp(z)$ converges absolutely for $z \notin \Lambda$ and uniformly on compact sets in $\mathbb{C} - \Lambda$.*

Exercise 2.4.3. *Verify Equation (2.4.2).*

Exercise 2.4.4. *Use the previous exercise together with the fact that $\wp$ is even to show $\wp$ is elliptic with respect to Λ.*

Now one might ask: are there any (nonconstant) elliptic functions of order 0 or 1? A basic result of complex analysis is the following.

Theorem 2.4.5. (Liouville) *Any bounded entire function is constant.*

If $f \in E(\Lambda)$ has no poles, then it must be holomorphic on $\mathbb{C}$, i.e., entire. Further the only values attained by f are the ones attained by restricting f to the fundamental parallelogram $\mathcal{F}$, so f is also bounded. This shows there are no nonconstant elliptic functions of order 0 ($\iff$ holomorphic).

Furthermore, one can show there are no elliptic functions of order 1. For those of you who remember your complex analysis, the argument goes as follows. The double periodicity condition, together with Cauchy's theorem, implies that the integral around any fundamental parallelogram with no poles on the border must be 0. (We may assume by shifting our fundamental domains that there are no poles on these parallelograms.) Hence by the residue

theorem, the sum of the residues in any parallelogram must be 0. In particular, we cannot have only one simple pole inside such a parallelogram.

The point is that $\wp$ is the simplest (nonconstant) elliptic function. Furthermore, it is not too difficult to show that the space of elliptic functions $E(\Lambda)$ is a field, and it is generated by $\wp$ and $\wp'$, i.e.,

$$E(\Lambda) = \mathbb{C}(\wp, \wp'),$$

i.e., any elliptic function can be expressed as a rational function in $\wp$ and $\wp'$.

2.5 Elliptic functions to elliptic curves

The theory of elliptic functions gave rise to the modern theory of elliptic curves and to modular forms. This is very beautiful and important mathematics, so despite the fact that it will not be so relevant to our treatment of modular forms, I feel obliged to at least summarize these connections.

In this section, x and y do not denote the real and imaginary parts of a complex number z.

Definition 2.5.1. *Let F be a field. An* **elliptic curve** *over F is a nonsingular (smooth) cubic curve in F^2.*

We won't give a precise definition of nonsingular, but by cubic curve we mean a curve defined by a cubic polynomial (in this case, in two variables).

Theorem 2.5.2. *Let F be a field of characteristic zero and E/F be an elliptic curve. Then, up to a change of variables, we may express E in* **Weierstrass form** *as*

$$y^2 = x^3 + ax + b, \quad (a, b \in F).$$

Furthermore, an equation of the above type defines an elliptic curve if and only if the **discriminant** $\Delta = \Delta(E) = -16(4a^3 + 27b^2) \neq 0$.

The nonzero discriminant condition is what forces the equation to be nonsingular. E.g., the equation $y^2 = x^3$ has a cusp at the origin (graph it over $\mathbb{R}$).

Actually, one typically wants to consider *projective* elliptic curves rather than just the affine curves. Suffice it to say that, for a curve in Weierstrass form, the projective elliptic curve can be thought of as the affine curve given above together with a *point at infinity*. It is well known that the points of an elliptic curve form an abelian group, with the point at infinity being the identity element. However to prove it directly from the above definition is not so easy.

Let $\Lambda = \langle \omega_1, \omega_2 \rangle$ be a period lattice in $\mathbb{C}$, and $\wp$ be the associated Weierstrass pe function. Then for all $z \in \mathbb{C} - \Lambda$,

$$\wp'(z)^2 = 4\wp(z)^3 - 60G_4\wp(z) - 140G_6,$$

where $G_k = \sum_{\omega \in \Lambda - \{0\}} \omega^{-k}$. In other words, the map $z \mapsto (\wp(z), \wp'(z))$ maps $\mathbb{C}$ onto (the affine points) of the elliptic curve $E_\Lambda : y^2 = 4x^3 - 60G_4x - 140G_6$ defined over $\mathbb{C}$, assuming

Λ is chosen so that the discriminant of E_λ is nonzero. (Replacing y with $2y$ puts this in Weierstrass form, if one wishes to do that.) By the periodicity of $\wp$ and $\wp'$, this factors through $(\mathbb{C} - \Lambda)/\Lambda$. We extend this map to $\mathbb{C}/\Lambda$ by sending Λ to the point at infinity on E_Λ.

This map gives an analytic isomorphism

$$\mathbb{C}/\Lambda \simeq E_\Lambda$$

from the torus $\mathbb{C}/\Lambda$ to the elliptic curve E_Λ. Furthermore, all elliptic curves over $\mathbb{C}$ (and consequently all elliptic curves over any subfield of $\mathbb{C}$) arise from some such lattice Λ. Consequently all (projective) elliptic curves over $\mathbb{C}$ are topologically tori. Now it is clear that the points on the (projective) elliptic curve E_Λ naturally form an additive group because $\mathbb{C}/\Lambda$ is (with respect to addition on $\mathbb{C}$, taken modulo Λ), and the identity of E_Λ is the point at infinity, which corresponds to the identity Λ of $(\mathbb{C}/\Lambda, +)$ by construction.

Hence elliptic functions (particularly $\wp$ and $\wp'$) provide a link between lattices in $\mathbb{C}$ and complex elliptic curves. We will now sketch how elliptic functions and elliptic curves lead to the theory modular forms.

We say two lattices Λ and Λ' in $\mathbb{C}$ are **equivalent** if $\Lambda' = \lambda\Lambda$ for some $\lambda \in \mathbb{C}^\times$.

Exercise 2.5.3. *Show that if Λ and Λ' are equivalent, the elliptic functions on Λ' are simply the elliptic functions on Λ composed with a simple change of variables.*

Furthermore, Λ and Λ' being equivalent is the same as the elliptic curves E_Λ and $E_{\Lambda'}$ being group isomorphic by an analytic map.

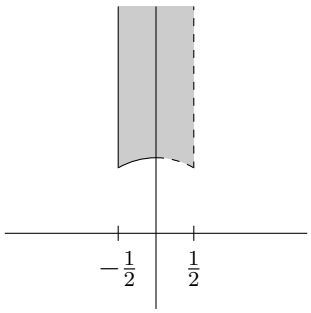
Write $\Lambda = \langle \omega_1, \omega_2 \rangle$. Clearly Λ is equivalent to the lattice $\langle 1, \tau \rangle$ where $\tau = \omega_2/\omega_1$. Since $\langle 1, \tau \rangle = \langle 1, -\tau \rangle$ and $\tau \notin \mathbb{R}$, we may assume $\mathrm{Im}(\tau) > 0$, i.e., τ lies in the **upper half-plane** $\mathfrak{H} = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$.

For $\tau, \tau' \in \mathfrak{H}$, the lattices $\langle 1, \tau \rangle$ and $\langle 1, \tau' \rangle$ are equal if and only if $\tau' \in \langle 1, \tau \rangle$ and $\tau \in \langle 1, \tau' \rangle$, i.e., if and only if $\tau' = a\tau + b$ and $\tau = c\tau' + d$ for some $a, b, c, d \in \mathbb{Z}$. It is easy to see this means $\tau' = \tau + b$ for some $b \in \mathbb{Z}$.

Hence we may assume $\tau \in \mathfrak{H}/\mathbb{Z}$, or more precisely, $-\frac{1}{2} \leq \mathrm{Re}(\tau) < \frac{1}{2}$. This means that $\{\pm 1, \pm\tau\}$ will be the four nonzero lattice points of $\langle 1, \tau \rangle$ closest to the origin. Hence the only way two lattices $\langle 1, \tau \rangle$ and $\langle 1, \tau' \rangle$ can be equivalent is if $\lambda \{\pm 1, \pm\tau\} = \{\pm 1, \pm\tau'\}$. It is not hard to see this means either $\tau' = \tau$ or $\tau' = -\frac{1}{\tau}$.

Exercise 2.5.4. *Let $\tau, \tau' \in \mathfrak{H}$. Fill in the details in the above argument to show that $\langle 1, \tau \rangle$ and $\langle 1, \tau' \rangle$ are equivalent if and only if $\tau' = \tau + b$ or $\tau' = -\frac{1}{\tau} + b$, for some $b \in \mathbb{Z}$.*

What the above shows is that equivalence classes of lattices are parametrized by the set of $\tau \in \mathfrak{H}$ subject to the conditions either $-\frac{1}{2} \leq \mathrm{Re}(\tau) < \frac{1}{2}$ and $|\tau| > 1$ or $-\frac{1}{2} \leq \mathrm{Re}(\tau) \leq 0$ and $|\tau| = 1$.



We will see in the next section that the (closure of the) above region is a fundamental domain for $\Gamma\backslash\mathfrak{H}$, where $\Gamma \simeq \mathrm{PSL}(2,\mathbb{Z})$ is the group of (hyperbolic) isometries of $\mathfrak{H}$ generated by the transformations $\tau \mapsto \tau + 1$ and $\tau \mapsto -\frac{1}{\tau}$.

What does all this have to do with modular forms? Well, if you are studying elliptic curves (or even just lattices), one thing you want to look is invariants. For example, given $\tau \in \mathfrak{H}$, consider the map

$$\Delta : \tau \mapsto \Lambda = \langle 1, \tau \rangle \mapsto E_\Lambda \mapsto \Delta(E_\lambda),$$

sending a point in $\mathfrak{H}$ to the discriminant of the associated elliptic curve. Equivalent (analytically group isomorphic) elliptic curves have essentially the same discriminant, so Δ is a (holomorphic) function on $\mathfrak{H}$ satisfying certain transformation properties under Γ. These properties will make Δ what is called a modular form of weight 12 and level 1.

Chapter 3

The Poincaré upper half-plane

There are two basic kinds of non-Euclidean geometry: spherical and hyperbolic. Spherical geometry is easy enough to imagine. Consider the sphere $S^2 = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}$. Then the geodesics (paths which locally minimize distance) are simply the great circles. In other words, for two points P and Q on S^2, a shortest path in S^2 between them is an arc lying on a great circle connecting P and Q (this great circle will be unique provided $P \neq -Q$), and the distance between P and Q in S^2 is the (Euclidean) length of this arc.

Hyperbolic geometry is a bit harder to picture. Perhaps the easiest way to initially visualize the hyperbolic plane is one sheet of the two-sheeted hyperboloid $x^2 + y^2 - z^2 = 1$, however this is not the easiest model to work with for many purposes. Just like with the Euclidean plane, the hyperbolic plane has translations, rotations and reflections. The two most common models to use, both named after Poincaré, are the Poincaré disc model (due to Beltrami) and the Poincaré upper half-plane model (due to Riemann). The disc model has the advantage of making rotations easy to visualize, and the upper half-plane model has the advantage of making translations easy to visualize. When working with modular forms, one always uses the upper half-plane model.

3.1 The hyperbolic plane

Unless stated otherwise, in this chapter $z, w \in \mathfrak{H}$ and $x, y \in \mathbb{R}_{>0}$.

Definition 3.1.1. *The* **Poincaré upper half-plane**, *or* **hyperbolic plane**, *is the set*

$$\mathfrak{H} = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$$

together with the metric given by the distance function

$$d(z, w) = \cosh^{-1}\left(1 + \frac{|z - w|^2}{2\,\mathrm{Im}(z)\,\mathrm{Im}(w)}\right).$$

Angles in the hyperbolic plane $\mathfrak{H}$ are just taken to be the usual Euclidean angles.

We will often write

$$d(z, w) = \cosh^{-1}\left(1 + \frac{u(z, w)}{2}\right),$$

where

$$u(z,w) = \frac{|z-w|^2}{\mathrm{Im}(z)\,\mathrm{Im}(w)}.$$

While modular forms will be functions on $\mathfrak{H}$ satisfying certain transformation laws under isometry groups, we do not actually need to know too much about the geometry of $\mathfrak{H}$ for our study of modular forms. Thus some basic facts about the geometry of $\mathfrak{H}$ which we do not need, but may help conceptually to be aware of, I will just state without proof. Instead one may refer to basic references on hyperbolic geometry, such as [And05] or [Kat92].

The first such fact is the following.

Lemma 3.1.2. *$d(z,w)$ is a metric on $\mathfrak{H}$.*

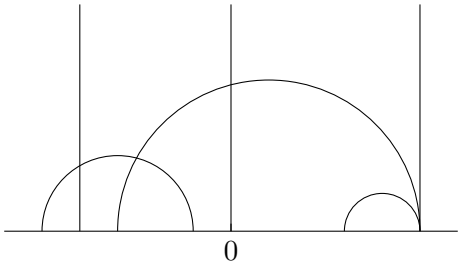
To see why this is plausible, first notice the argument of $\cosh^{-1}$ always lies in $[1,\infty)$. Recall $\cosh(x) = \frac{e^x+e^{-x}}{2}$ so $\cosh(0) = 1$ and cosh is increasing and concave up on $[0,\infty)$. So $\cosh^{-1}(1) = 0$ and $\cosh^{-1}$ is increasing and concave down on $[1,\infty)$ (it grows logarithmically of course).

In particular, (i) $d(z,w) : \mathfrak{H}\times\mathfrak{H} \to [0,\infty)$ with $d(z,w) = 0$ if and only if the argument of $\cosh^{-1}$ is 1, i.e., if and only if $u(z,w) = 0$, i.e., if and only if $z = w$. It is also clear that (ii) $d(z,w) = d(w,z)$ for $z,w \in \mathfrak{H}$ since $u(z,w) = u(w,z)$. Thus to show $d(z,w)$ is a metric, it remains to show (iii) the triangle inequality, $d(z,w) \le d(z,v) + d(v,w)$, holds for all $z,w,v \in \mathfrak{H}$. This is somewhat technical (and more easily proven using a different formulation of the hyperbolic distance), so I'll leave it to you to either work out or look up.

Another fact one should know, though not formally used in our development of modular forms is the following.

Proposition 3.1.3. *The geodesics of $\mathfrak{H}$ are the vertical lines $\{z \in \mathfrak{H} : \mathrm{Re}(z) = x_0\}$ and the semicircles $\{z \in \mathfrak{H} : |z - z_0| = r_0\}$ in $\mathfrak{H}$ which meet $\mathbb{R}$ orthogonally. In particular, given any $z,w \in \mathfrak{H}$, there is a unique geodesic connecting them.*

Here are some drawing of geodesics:



Don't get confused by analogy with the spherical model here. While it's true that the geodesic arc from z to w is the shortest path between them, the distance between them is not the Euclidean arc length of this arc, but rather the hyperbolic arc length given by $(ds)^2 = \frac{(dx)^2+(dy)^2}{y^2}$, where $z = x + iy$. In other words, the hyperbolic distance $d(z,w)$ as defined above does not give you the Euclidean length of geodesic arc from z to w. For instance, the distance from any point $z \in \mathfrak{H}$ to the origin (i.e., $\lim_{y\to 0^+} d(iy,z)$) is infinite. Similarly, the distance from any finite point $z \in \mathfrak{H}$ to $i\infty$ (i.e., $\lim_{y\to\infty} d(iy,z)$) is infinite. We compute the case of $z = i$ below.

Example 3.1.4. *Let $y \in \mathbb{R}_{>0}$. Then*

$$d(i, iy) = \cosh^{-1}\left(1 + \frac{(y-1)^2}{2y}\right)$$

In fact, since $\cosh^{-1}(x) = \ln(x + \sqrt{x+1}\sqrt{x-1})$*, we can write*

$$d(i, iy) = \ln\left(\frac{y^2+1}{2y} + \frac{|y^2-1|}{2y}\right) = |\ln y|.$$

In particular, we see iy and i/y are equidistant from i, and indeed $d(i,0) := \lim_{y\to 0^+} d(i,iy) = \infty = \lim_{y\to\infty} d(i,iy) =: d(i,i\infty)$. Applying the triangle inequality shows in fact $d(z,0) = d(z,i\infty) = \infty$ for any $z \in \mathfrak{H}$.

Exercise 3.1.5. *For $x, y > 0$, compute the hyperbolic distance $d(ix, iy)$.*

Exercise 3.1.6. *For $z = e^{i\theta} \in \mathfrak{H}$, compute the hyperbolic distance $d(i, z)$.*

3.2 Fractional linear transformations

In this section, z will denote an element of $\mathfrak{H}$, $x = \mathrm{Re}(z)$ and $y = \mathrm{Im}(z) > 0$.

One nice thing about the upper half-plane model is that there is a nice way to describe the isometries of $\mathfrak{H}$ in terms of fractional linear transformations (also called Möbius transformations), which you may have encountered in complex analysis.

Recall for a ring R, we may define the 2×2 **special linear group**

$$\mathrm{SL}_2(R) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in R,\ ad - bc = 1 \right\},$$

as well as the corresponding **projective special linear group**

$$\mathrm{PSL}_2(R) := \mathrm{SL}_2(R)/Z,$$

where Z denotes the (central) subgroup of scalar matrices of $\mathrm{SL}_2(R)$, i.e.,

$$Z = \left\{ \begin{pmatrix} a & \\ & a \end{pmatrix} \in \mathrm{SL}_2(R) \right\} = \left\{ \begin{pmatrix} a & \\ & a \end{pmatrix} : a \in R,\ a^2 = 1 \right\}.$$

(If you're not familiar with these, check that these are groups.) The most important cases for us in this course are $R = \mathbb{R}$ and $R = \mathbb{Z}$; in both of these cases $Z = \{\pm I\}$.

Definition 3.2.1. *A* **fractional linear transformation** *of $\mathfrak{H}$ is a map $\mathfrak{H} \to \mathbb{C}$ of the form*

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

where $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R})$.

(More generally, one can consider fractional linear transformations $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}_2(\mathbb{C})$, but we will always mean elements of $\mathrm{SL}_2(\mathbb{R})$ by the term fractional linear transformation.)

First note that, for $c, d \in \mathbb{R}$ not both 0, $z \to \frac{az+b}{cz+d}$ is a rational function of $\mathbb{C}$ with at most one pole (of order 1) at $-d/c \in \mathbb{R}$ if $c \neq 0$, fractional linear transformations are holomorphic maps from $\mathfrak{H}$ to $\mathbb{C}$.

Exercise 3.2.2. *Show that the image of $\mathfrak{H}$ under a linear fractional transformation is contained in $\mathfrak{H}$.*

This allows us to compose fractional linear transformations on $\mathfrak{H}$.

Lemma 3.2.3. *For $\sigma, \tau \in \mathrm{SL}_2(\mathbb{R})$, the fractional linear transformation $\sigma \circ \tau : \mathfrak{H} \to \mathfrak{H}$ given by their composition is equal to the fractional linear transformation $\sigma\tau$ given by their matrix product.*

The proof is just a simple computation, which we relegate to

Exercise 3.2.4. *Prove Lemma 3.2.3.*

Lemma 3.2.5. *Let $\tau \in \mathrm{SL}_2(\mathbb{R})$. The fractional linear transformation $\tau : \mathfrak{H} \to \mathbb{C}$ in fact an analytic automorphism $\tau : \mathfrak{H} \to \mathfrak{H}$, i.e., $\tau : \mathfrak{H} \to \mathfrak{H}$ an analytic bijection whose inverse is also analytic. Further, τ acts trivially on $\mathfrak{H}$ if and only if $\tau = \pm I$.*

Proof. Write $\tau = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then

$$\tau(z) = \frac{ax + aiy + b}{cx + ciy + d} \cdot \frac{cx + d - ciy}{cx + d - ciy} = \frac{(ad + bc)x + bd + acy^2 + iy(ad - bc)}{(cx + d)^2 + (cy)^2}$$

Since $ad - bc = 1$, we have

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} z = \frac{(ad + bc)x + bd + acy^2 + iy}{(cx + d)^2 + (cy)^2} = \frac{(ad + bc)x + bd + acy^2 + iy}{|cz + d|^2}, \tag{3.2.1}$$

which clearly has positive imaginary part, so $\tau(\mathfrak{H}) \subseteq \mathfrak{H}$.

To see this is one-to-one, suppose $\tau(z) = \tau(w)$ for $z, w \in \mathfrak{H}$. Then cross-multiplying denominators and expanding out

$$\frac{az + b}{cz + d} = \frac{aw + b}{cw + d}$$

gives

$$adz + bcw = adw + bcz,$$

which says

$$(ad - bc)z = (ad - bc)w,$$

i.e., $z = w$.

Note the identity of $\mathrm{SL}_2(\mathbb{R})$ acts trivially on $\mathfrak{H}$. By the previous lemma, this means $\tau^{-1} \circ \tau$ acts trivially on $\mathfrak{H}$. In particular, τ maps onto $\mathfrak{H}$. Hence $\tau : \mathfrak{H} \to \mathfrak{H}$ is an analytic automorphism.

In order for τ to be the identity on $\mathfrak{H}$, looking at the imaginary part of $\tau(z)$ shows $(cx+d)^2+(cy)^2=1$ for all $x\in\mathbb{R}$ and $y>0$. This implies $c=0$ and $d=\pm 1$. Then the real part of $\tau(z)$ is simply $adx+bd$. For this to always equal x, we need $b=0$ and $ad=1$. Hence $\tau=\pm I$ are the only elements of $\mathrm{SL}_2(\mathbb{R})$ which act trivially on $\mathfrak{H}$. □

So from now on, we will view fractional linear transformations as analytic automorphisms on $\mathfrak{H}$ (i.e., holomorphic bijections from $\mathfrak{H}$ to $\mathfrak{H}$ whose inverses are also holomorphic).

Because composition respects group multiplication, and the only fractional linear transformations acting trivially are $\pm I$, two elements $\sigma,\tau\in\mathrm{SL}_2(\mathbb{R})$ define the same map if and only if $\sigma=\pm\tau$. This means we can identify the group of fractional linear transformations on $\mathfrak{H}$ with $\mathrm{PSL}_2(\mathbb{R})$. Despite the fact that elements of $\mathrm{PSL}_2(\mathbb{R})$ are technically not 2×2 matrices, we will still write elements of $\mathrm{PSL}_2(\mathbb{R})$ as elements of $\mathrm{SL}_2(\mathbb{R})$ with the convention that $-\tau$ is identified with τ.

Now let's look at a few specific examples of fractional linear transformations. Let

$$\tau_n=\begin{pmatrix}1 & n\\ & 1\end{pmatrix},$$
$$\delta_{m^2}=\begin{pmatrix}m & \\ & m^{-1}\end{pmatrix},\quad\text{and}$$
$$\iota=\begin{pmatrix} & 1\\ -1 & \end{pmatrix}.$$

Then

$$\tau_n(z)=\begin{pmatrix}1 & n\\ & 1\end{pmatrix}z=z+n$$

simply translates $\mathfrak{H}$ to the right by n,

$$\delta_{m^2}(z)=\begin{pmatrix}m & \\ & m^{-1}\end{pmatrix}z=m^2z$$

dilates, or scales, $\mathfrak{H}$ outward radially from the origin by m^2, and

$$\iota(z)=\begin{pmatrix} & 1\\ -1 & \end{pmatrix}z=-\frac{1}{z}$$

inverts $\mathfrak{H}$ about the semicircle $\{z\in\mathfrak{H}:|z|=1\}$. (Writing $z=re^{i\theta}$ may make these latter two transformations easier to visualize. The second then becomes obvious, and we see $\iota(re^{i\theta})=\frac{1}{r}e^{i(\pi-\theta)}$.) Note all these transformations map geodesics to geodesics. Furthermore, because they are holomorphic maps, they are conformal, i.e., they preserve angles.

Lemma 3.2.6. *The group of fractional linear transformations* $\mathrm{PSL}_2(\mathbb{R})$ *acts transitively on* $\mathfrak{H}$*, i.e., for any* $z,w\in\mathfrak{H}$*, there exists* $\tau\in\mathfrak{H}$ *such that* $\tau(w)=z$*.*

Proof. It suffices to show that for any $z\in\mathfrak{H}$, there exists $\tau\in\mathrm{PSL}_2(\mathbb{R})$ such that $\tau(i)=z$. For then there also exists $\sigma\in\mathrm{PSL}_2(\mathbb{R})$ such that $\sigma(i)=w$, so $(\tau\sigma^{-1})(w)=z$ by the previous lemma.

Write $z=x+iy$ with $x,y\in\mathbb{R}$ and let $\tau=\tau_x\delta_y$. Then $\tau(i)=\tau_x(iy)=x+iy=z$. □

A basic fact, though we will not need it, is the following.

Proposition 3.2.7. *The group* $\mathrm{PSL}_2(\mathbb{R})$ *of fractional linear transformations on* $\mathfrak{H}$ *is precisely the group of orientation-preserving isometries of* $\mathfrak{H}$.

(To obtain the orientation-reversing isometries, one considers 2×2-matrices of determinant -1.)

We note that since angles in $\mathfrak{H}$ are simply Euclidean angles, the fact that holomorphic maps are conformal implies fractional linear transformations preserve hyperbolic angles.

While we omit the proof, we remark that it is straightforward (though somewhat tedious) to check that all fractional linear transformations are isometries, i.e., they preserve distance, i.e., $d(z,w) = d(\tau(z), \tau(w))$ for $z, w \in \mathfrak{H}$ and $\tau \in \mathrm{PSL}_2(\mathbb{R})$. However the computations are simpler in the following case.

Exercise 3.2.8. *Let* $\tau \in \mathrm{PSL}_2(\mathbb{R})$ *and* $z \in \mathfrak{H}$. *Show* $d(z,i) = d(\tau(z), \tau(i))$. *(Note it suffices to show* $u(z,i) = u(\tau(z), \tau(i))$.*)*

From now on, we will often write $\tau(z)$ simply as τz for $\tau \in \mathrm{PSL}_2(\mathbb{R})$.

3.3 The modular group

In this course, we will not work with the group of all fractional linear transformations $\mathrm{PSL}_2(\mathbb{R})$, but rather certain nice discrete subgroups Γ. First we will study the most basic discrete subgroup, the **(full) modular group**, $\mathrm{PSL}_2(\mathbb{Z})$.

(Some authors call $\mathrm{SL}_2(\mathbb{Z})$ the full modular group because they prefer to work in terms of matrices, but we prefer to think in terms of fractional linear transformations of $\mathfrak{H}$, so $\mathrm{PSL}_2(\mathbb{Z})$ is more natural to work with. However, there is no difference between special linear groups and projective linear groups in terms of fractional linear transformations, so this should not cause any confusion.)

Proposition 3.3.1. $\mathrm{PSL}_2(\mathbb{Z}) = \langle S, T \rangle$ *where* $S = \begin{pmatrix} & 1 \\ -1 & \end{pmatrix}$ *and* $T = \begin{pmatrix} 1 & 1 \\ & 1 \end{pmatrix}$.

Note that S and T are simply the inversion ι and translation τ_1 defined in the previous sections, however the notation of S and T is standard for these specific matrices.

Proof. Let $\Gamma = \langle S, T \rangle$. It is clear $\Gamma \subseteq \mathrm{PSL}_2(\mathbb{Z})$. Note also that $T \in \Gamma$ implies

$$N = \left\{ \begin{pmatrix} 1 & x \\ & 1 \end{pmatrix} : x \in \mathbb{Z} \right\} \subseteq \Gamma.$$

Similarly

$$\overline{N} = \left\{ \begin{pmatrix} 1 & \\ y & 1 \end{pmatrix} : y \in \mathbb{Z} \right\} \subseteq \Gamma,$$

since N is generated by $\overline{T} = S^{-1}TS = \begin{pmatrix} 1 & \\ -1 & 1 \end{pmatrix} \in \Gamma$.

Now let $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PSL}_2(\mathbb{Z})$. We want to show $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$. First note that

$$S^{-1}\begin{pmatrix} a & b \\ c & d \end{pmatrix} S = \begin{pmatrix} d & -c \\ -b & a \end{pmatrix},$$

we may assume $|a| \leq |d|$. Next observe

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 1 & x \\ & 1 \end{pmatrix} = \begin{pmatrix} a & ax+b \\ c & cx+d \end{pmatrix}.$$

Hence by multiplying $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ on the right by some element of N, we may assume $0 \leq b < |a|$. Similarly, since

$$\begin{pmatrix} 1 & \\ y & 1 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & b \\ ay+c & by+d \end{pmatrix},$$

we may also assume $0 \leq c < |a|$. Hence ad and bc are integers such that $ad - bc = 1$, $|ad| \geq a^2$ and $|bc| < a^2$. So we must have $ad = 1$ and $bc = 0$. In particular, $a = d = \pm 1$, and $0 \leq b < |a|$ and $0 \leq c < |a|$ then means $b = c = 0$, i.e.,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \pm\begin{pmatrix} 1 & 0 \\ & 1 \end{pmatrix} \in \Gamma.$$

□

Remark. One can show in fact $\langle S, T | S^2 = (ST)^3 = 1\rangle$ is a presentation for $\mathrm{PSL}_2(\mathbb{Z})$.

We have already defined the notion of a fundamental domain for Ω/Λ, where $\Omega \subset \mathbb{C}$ and Λ is a group of (Euclidean) isometries of Ω. We define fundamental domains in the hyperbolic plane the same way. For formality's sake, we write out the definition now.

Definition 3.3.2. *Let $\mathcal{F} \subset \mathfrak{H}$ be a closed set with connected interior, and Γ a subgroup of $\mathrm{PSL}_2(\mathbb{Z})$. We say $\mathcal{F}$ is a* **fundamental domain** *for Γ (or $\Gamma\backslash\mathfrak{H}$) if*

(i) any $z \in \mathfrak{H}$ is Γ-equivalent to some point in $\mathcal{F}$;

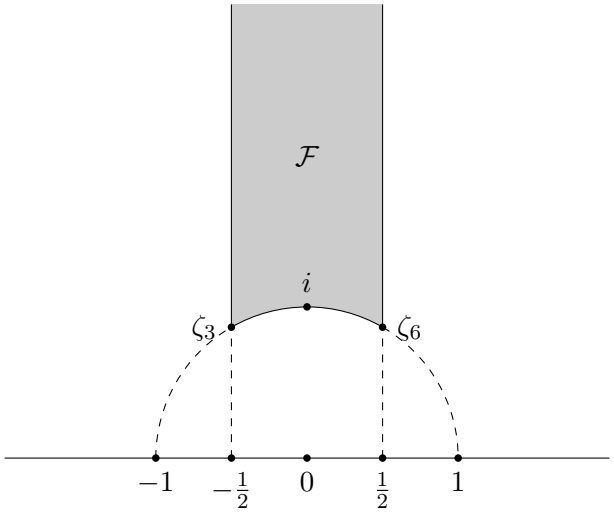
(ii) no two interior points of $\mathcal{F}$ are Γ-equivalent; and

(iii) the boundary $\partial\mathcal{F}$ of $\mathcal{F}$ is a finite union of smooth curves in $\mathfrak{H} \cap \mathcal{F}$.

To clarify for future use, by (ii) we mean that if z, z' lie in the interior $\mathcal{F}^0$ of $\mathcal{F}$ such that $\gamma z = z'$ for $\gamma \in \Gamma$, then $z = z'$. We do not mean that $\gamma z = z'$ implies $\gamma = I$. However, we will see later (Lemma 3.5.1) that our interpretation of (ii) implies $\gamma = I$.

Let

$$\mathcal{F} = \left\{ z \in \mathfrak{H} : |\mathrm{Re}(z)| \leq \frac{1}{2},\ |z| \geq 1 \right\}.$$



(Here $\zeta_k = e^{2\pi i/k}$.)

Proposition 3.3.3. *$\mathcal{F}$ is a fundamental domain for* $\mathrm{PSL}_2(\mathbb{Z})$.

Proof. Condition (iii) of Defintion 3.3.2 ($\partial\mathcal{F}$ is a finite union of smooth curves) is obviously satisfied, so we just need to show the first two conditions.

First we show (i) any point $z \in \mathfrak{H}$ is $\mathrm{PSL}_2(\mathbb{Z})$-equivalent to some point in $\mathcal{F}$, i.e., for any $z \in \mathfrak{H}$, there is some $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PSL}_2(\mathbb{Z})$ such that $\gamma z \in \mathcal{F}$. Observe that by (3.2.1), we have $\mathrm{Im}(\gamma z) = \frac{\mathrm{Im}(z)}{|cz+d|^2}$. Since $c, d \in \mathbb{Z}$ and they are not both 0, $|cz+d|^2$ attains a minimum as γ ranges over $\mathrm{PSL}_2(\mathbb{Z})$. Consequently, we may choose γ such that $\mathrm{Im}(\gamma z)$ is maximal. Replacing γ by $T^n\gamma$ for some $n \in \mathbb{Z}$ shows that we may further assume $|\mathrm{Re}(\gamma z)| \leq \frac{1}{2}$.

It remains to show $|\gamma z| \geq 1$. Suppose not. Let $w = \gamma z$ and write $w = u + iv$. Then

$$Sw = \frac{-u + iv}{u^2 + v^2}.$$

In particular, if $|w| < 1$, then $\mathrm{Im}(Sw) > \mathrm{Im}(w)$, i.e., $\mathrm{Im}(S\gamma z) > \mathrm{Im}(\gamma z)$, contradicting the maximality of $\mathrm{Im}(\gamma z)$. This proves (i).

Now we need to show (ii) no two interior points of $\mathcal{F}$ are $\mathrm{PSL}_2(\mathbb{Z})$-equivalent. Suppose z and w lie in the interior of $\mathcal{F}$, and $\gamma z = w$ for some $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PSL}_2(\mathbb{Z})$. We may assume $\mathrm{Im}(w) \geq \mathrm{Im}(z)$, or else we could switch z and w and replace γ by γ^{-1}. Since $\mathrm{Im}(w) = \frac{\mathrm{Im}(z)}{|cz+d|^2}$, this means $|cz + d| \leq 1$. In particular $|\mathrm{Im}(cz + d)| = |c|\mathrm{Im}(z) < 1$, but $z \in \mathcal{F}$ implies $\mathrm{Im}(z) > \frac{\sqrt{3}}{2}$ so $|c| < 2$.

First suppose $c = 0$. Then $ad - bc = ad = 1$. Multiplying γ by ± 1 if necessary (recall we are working in $\mathrm{PSL}_2(\mathbb{Z})$, so this does not change γ), we may assume $a = d = 1$ so $\gamma = T^n$ for some $n \in Z$. In this case it is clear we cannot have $\gamma z \in \mathcal{F}$.

Now suppose $c = \pm 1$. Multiplying γ by ± 1 if necessary, we may assume $c = 1$. Note

$$1 > |cz + d|^2 = \mathrm{Re}(z + d)^2 + \mathrm{Im}(z)^2 > \mathrm{Re}(z + d)^2 + \frac{3}{4}.$$

Hence $|\mathrm{Re}(z + d)| < \frac{1}{2}$, which implies $d = 0$ since $|\mathrm{Re}(z)| < \frac{1}{2}$. Then the determinant condition implies $b = -1$. So

$$\gamma = \begin{pmatrix} a & -1 \\ 1 & 0 \end{pmatrix},$$

i.e., $\gamma z = a - \frac{1}{z}$, which cannot lie in $\mathcal{F}$. This proves (ii). □

Recall when we defined the notion of a fundamental domain, we said some authors require fundamental domains to be convex. Note that $\mathcal{F}$ is convex in $\mathfrak{H}$ (with respect to the hyperbolic metric), i.e., given any two points in $\mathcal{F}$, one can connect them with a unique geodesic segment, and that segment—either a vertical line segment or a segment of a semicircle centered on the real line—lies entirely in $\mathcal{F}$.

Exercise 3.3.4. *Show two boundary points z, z' of $\mathcal{F}$ are $\mathrm{PSL}_2(\mathbb{Z})$-equivalent if and only if $\mathrm{Im}(z) = \mathrm{Im}(z')$ and $\mathrm{Re}(z) = -\mathrm{Re}(z')$.*

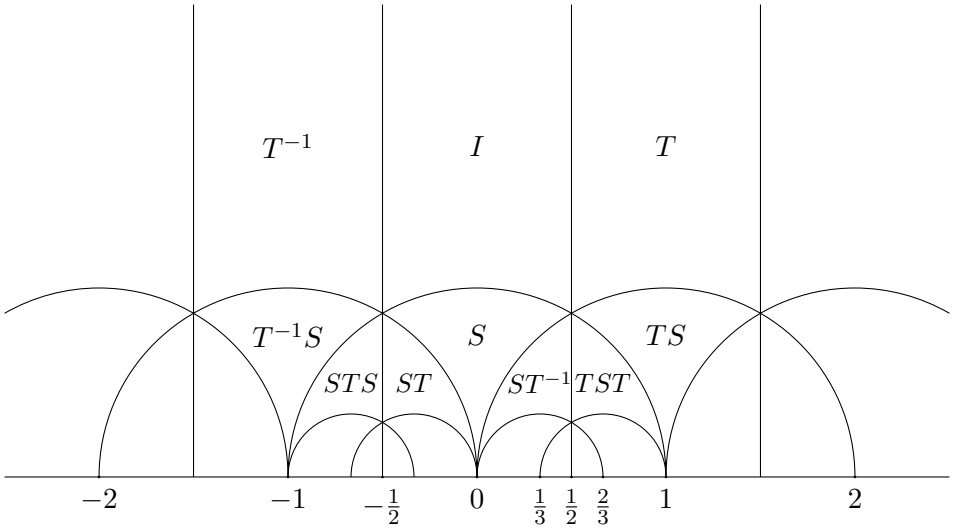
The above exercise shows the quotient space $\mathrm{PSL}_2(\mathbb{Z})\backslash\mathcal{H}$ parametrizes equivalence classes of lattices as described Section 2.5 (cf. figure at end of Section 2.5). (Note, in contrast to Chapter 2, it is standard to write the quotient by the action of $\mathrm{PSL}_2(\mathbb{Z})$ on the left here because we think of matrices as acting on $\mathfrak{H}$ from the left.)

Since the fundamental domain $\mathcal{F}$ we have given above is universally used, it is called the **standard fundamental domain** for $\mathrm{PSL}_2(\mathbb{Z})$. There is a trivial way to cook up other fundamental domains—namely $\gamma\mathcal{F}$ is also a fundamental domain for $\mathrm{PSL}_2(\mathbb{Z})$ for any $\gamma \in \mathrm{PSL}_2(\mathbb{Z})$. This statement is slightly generalized in the following simple exercise.

Exercise 3.3.5. *Let $\Gamma \subseteq \mathrm{PSL}_2(\mathbb{R})$ be discrete and suppose $\mathcal{F}'$ is a fundamental domain for Γ. Suppose $\gamma \in \mathrm{PSL}_2(\mathbb{R})$ such that $\gamma^{-1}\Gamma\gamma = \Gamma$. Show $\gamma\mathcal{F}'$ is also a fundamental domain for Γ.*

***Exercise 3.3.6.** *(i) Show that the set of $\gamma\mathcal{F}$ with $\gamma \in \mathrm{PSL}_2(\mathbb{Z})$ tile $\mathfrak{H}$, i.e., their union covers $\mathfrak{H}$ and their interiors are pairwise disjoint.*

(ii) Show the picture below of the partial tiling of $\mathfrak{H}$ given by $\gamma\mathcal{F}$ is correct, where we labelled the region $\gamma\mathcal{F}$ simply by γ. (Hint: begin by showing S takes the vertical line $x + i\mathbb{R}_{>0}$ to the upper semicircle passing through 0 and $-\frac{1}{x}$.)



In the above diagram, note $STS = T^{-1}ST^{-1}$ and $TST = ST^{-1}S$.

3.4 Congruence subgroups

From our point of view so far, the simplest kinds of (meromorphic) modular forms (modular functions of level one) will be meromorphic functions $f : \mathfrak{H} \to \hat{\mathbb{C}}$ satisfying

$$f(\gamma z) = f(z) \quad \text{for } \gamma \in \mathrm{PSL}_2(\mathbb{Z}). \tag{3.4.1}$$

Note this is a hyperbolic analogue of the definition of elliptic functions. Specifically, fix a lattice $\Lambda \subset \mathbb{C}$ and let Γ_Λ be the group of isometries of $\mathbb{C}$ given by $\gamma_\omega(z) = z + \omega$ where ω ranges over Λ. Then elliptic functions w.r.t. Λ were simply defined to be the (meromorphic) functions on $\mathbb{C}$ satisfying $f(\gamma z) = f(z)$ for $\gamma \in \Gamma_\Lambda$.

So instead of being functions on $\mathbb{C}$ (with Euclidean geometry) invariant under an isometry group Γ_Λ corresponding to a lattice Λ, modular functions will be functions on $\mathfrak{H}$ (with hyperbolic geometry) invariant under a hyperbolic isometry group Γ, with the simplest case being $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$.

If one wants to think of a hyperbolic analogue of the lattice Λ, there is no issue here. Note one can recover the lattice $\Lambda \subset \mathbb{C}$ from the group Γ_Λ simply by looking at the Γ_Λ-translates of the origin in $\mathbb{C}$. In this way, we see any isometry group of $\mathbb{C}$ isomorphic to $\mathbb{Z} \times \mathbb{Z}$ gives a lattice. Similarly, for any (noncyclic) "discrete" isometry group $\Gamma \subset \mathrm{PSL}_2(\mathbb{R})$ of $\mathfrak{H}$, one can think of the corresponding "hyperbolic lattice" as simply the Γ-translates of i. (Note discrete here precisely means the Γ-translates of i form a discrete subset of $\mathfrak{H}$, and one excludes the case of cyclic groups Γ as they will correspond to "1-dimensional" lattices in $\mathfrak{H}$.)

Now the modular functions satisfying (3.4.1), while certainly very interesting, are not sufficient for most number theoretic purposes. There are two ways of generalizing (3.4.1) that will include many import functions arising naturally in number theory. One way is to loosen the invariance by only requiring $f(\gamma z) = \alpha(\gamma, z)f(z)$ for some "weight" $\alpha(\gamma, z)$ (then one calls the resulting functions modular *forms*). The second way is to not require (3.4.1)

hold for all $\gamma \in \mathrm{PSL}_2(\mathbb{Z})$, but merely all $\gamma \in \Gamma$, where Γ is a suitable subgroup of $\mathrm{PSL}_2(\mathbb{Z})$. In fact, as alluded to earlier, one could just require that Γ be a noncyclic discrete subgroup of $\mathrm{PSL}_2(\mathbb{R})$. However the most interesting cases of study for number theory is when Γ is a *congruence subgroup* of $\mathrm{PSL}_2(\mathbb{Z})$.

Definition 3.4.1. *Let $N \in \mathbb{N}$. The* **modular group of level** N *is*

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PSL}_2(\mathbb{Z}) : c \equiv 0 \bmod N \right\}.$$

Note when $N = 1$, we have $\Gamma_0(1) = \mathrm{PSL}_2(\mathbb{Z})$, and we sometimes call this the **full modular group**. It is also clear that

$$\Gamma_0(N) \subseteq \Gamma_0(M) \quad \text{if } M|N.$$

One also has the more refined congruence subgroups

$$\Gamma_1(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PSL}_2(\mathbb{Z}) : \begin{pmatrix} a & b \\ c & d \end{pmatrix} \equiv \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} \bmod N \right\}$$

and the **principal congruence subgroups**

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

All of these subgroups are finite index inside $\mathrm{PSL}_2(\mathbb{Z})$ and have been studied classically in the context of modular forms. Note that

$$\Gamma(N) \subseteq \Gamma_1(N) \subseteq \Gamma_0(N).$$

In general, a **congruence subgroup** of $\mathrm{PSL}_2(\mathbb{Z})$ is a subgroup containing some $\Gamma(N)$. However, the most important congruence subgroups are the modular groups $\Gamma_0(N)$, and we will restrict our focus in this course to them.

Unless explicitly stated otherwise, from now on, Γ will always denote a congruence subgroup of $\mathrm{PSL}_2(\mathbb{Z})$.

In order to understand the coset space $\mathrm{PSL}_2(\mathbb{Z})/\Gamma_0(N)$, it will be helpful to use the "projective line over $\mathbb{Z}/N\mathbb{Z}$." For $(a, c), (a', c') \in \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z}$, we write $(a, c) \sim (a', c')$ if $(a', c') = (\lambda a, \lambda c)$ for some $\lambda \in (\mathbb{Z}/N\mathbb{Z})^\times$. It is clear that $\sim$ is an equivalence relation. We denote the equivalence class of (a, c) by $(a : c)$, and define the **projective line**

$$\mathbb{P}^1(\mathbb{Z}/N\mathbb{Z}) = \{(a : c) | a, c \in \mathbb{Z}/N\mathbb{Z}, \gcd(a, c) = 1\}.$$

Here by $\gcd(a, c) = 1$, we mean that there exist $\widetilde{a}, \widetilde{c} \in \mathbb{Z}$ in the congruence classes $a \bmod N$ and $c \bmod N$ such that $\gcd(\widetilde{a}, \widetilde{c}) = 1$); or, alternatively, there exist $r, s \in \mathbb{Z}/N\mathbb{Z}$ such that $ar + cs \equiv 1 \bmod N$.

Lemma 3.4.2. *The coset space $\mathrm{PSL}_2(\mathbb{Z})/\Gamma_0(N)$ is finite, in bijection with $\mathbb{P}^1(\mathbb{Z}/N\mathbb{Z})$, and a set of coset representatives is given by matrices of the form $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, where $(a, c) \in \mathbb{Z} \times \mathbb{Z}$ such that (i) $(a : c)$ ranges over $\mathbb{P}^1(\mathbb{Z}/N\mathbb{Z})$ and (ii) b, d are chosen such that $ad - bc = 1$.*

Proof. First observe that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} \Gamma_0(N) \iff \begin{pmatrix} d' & -b' \\ -c' & a' \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N).$$

By definition, this holds if and only if the lower left hand coefficient of the product on the right, $a'c - c'a$, is divisible by N. In other words, we have

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} \Gamma_0(N) \iff a'c \equiv ac' \bmod N.$$

Note that a pair of integers (a, c) can occur as a column of a matrix in $\mathrm{PSL}_2(\mathbb{Z})$ if and only if $ad - bc = 1$ is solvable, i.e., if and only if $\gcd(a, c) = 1$.

Hence the space of cosets $\mathrm{PSL}_2(\mathbb{Z})/\Gamma_0(N)$ is in bijection with the set of pairs $(a, c) \in \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z}$ such that $\gcd(a, c) = 1$ under the equivalence that $(a, c) \equiv (a', c') \iff a'c \equiv ac' \bmod N$. It suffices to show that $a'c \equiv ac' \bmod N$ is equivalent to $(a', c') = (\lambda a, \lambda c)$ for some $\lambda \in (\mathbb{Z}/N\mathbb{Z}^\times)$.

Clearly if $(a', c') = (\lambda a, \lambda c)$, then $a'c \equiv ac' \bmod N$. On the other hand, suppose $(a, c), (a', c') \in \mathbb{Z} \times \mathbb{Z}$ such that $\gcd(a, c) = \gcd(a', c') = 1$ and $a'c \equiv ac' \bmod N$. Write $N = p_1^{n_1} \cdots p_k^{n_k}$. By the Chinese Remainder Theorem, $\mathbb{Z}/N\mathbb{Z} \simeq \mathbb{Z}/p_1^{n_1}\mathbb{Z} \times \cdots \times \mathbb{Z}/p_k^{n_k}\mathbb{Z}$. Under this isomorphism, write $a = (a_1, \ldots, a_k)$ where $a_i \in \mathbb{Z}/p_i^{n_i}\mathbb{Z}$, and do similarly for c, a', and c'.

Then $a'c \equiv ac' \bmod N$ means $a_i'c_i \equiv a_ic_i' \bmod p_i^{n_i}$ for each i. By the condition $\gcd(a, c) = 1$, we know p_i cannot divide both a_i and c_i. Assume $p_i \nmid a_i$, and say $\gcd(c_i, p_i^{n_i}) = p_i^{e_i}$. Then $a_i'c_i \equiv a_ic_i' \bmod p_i^{n_i}$ means $\gcd(c_i', p_i^{n_i}) = p_i^{e_i}$ so $p_i \nmid a_i'$. Put $\lambda_i = a_i'a_i^{-1}$. Then $a_i' = \lambda_i a_i$ and $c_i' = \lambda_i c_i$. Consequently $\lambda = (\lambda_1, \ldots, \lambda_k) \in (\mathbb{Z}/N\mathbb{Z})^\times$ such that $(a', c') \equiv (\lambda a, \lambda c) \bmod N$. □

Corollary 3.4.3. $[\mathrm{PSL}_2(\mathbb{Z}) : \Gamma_0(N)] = N \prod_{p|N} \left(1 + \frac{1}{p}\right)$, *where p runs through the distinct prime divisors of N.*

Proof. First suppose $N = p^n$. For any equivalence class $(a : c)$, either a is divisible by p or not. If not, we may rewrite this class as $(\lambda a : \lambda c) = (1 : \lambda c)$ where $\lambda \equiv a^{-1} \bmod N$. There are N choices for $\lambda c \bmod N$, and all inequivalent. If $p|a$ then $p \nmid c$, so by the same argument we may rewrite $(a : c) = (\lambda a : 1)$. Since λa must be a multiple of p, there are N/p choices for λa, and they all give inequivalent classes. Thus the proposition holds when $N = p^n$.

The proof for arbitrary N follows from the prime power case using the Chinese Remainder Theorem (see exercise below). □

Corollary 3.4.4. *Let p be a prime, and $n \in \mathbb{N}$. Then a complete set of coset representatives of $\mathrm{PSL}_2(\mathbb{Z})/\Gamma_0(p^n)$ is given by*

$$\left\{ T^{ip}S = \begin{pmatrix} ip & -1 \\ 1 & 0 \end{pmatrix} : 0 \le i < p^{n-1} \right\} \cup \left\{ ST^jS = \begin{pmatrix} 1 & 0 \\ -j & 1 \end{pmatrix} : 0 \le j < p^n \right\}.$$

Proof. First check T^iS and ST^jS are given by the expressions above. Then observe the proof of Corollary 3.4.3 actually tells us that

$$\mathbb{P}^1(\mathbb{Z}/p^n\mathbb{Z}) = \left\{ (ip : 1) : 0 \le i < p^{n-1} \right\} \cup \left\{ (1 : j) : 0 \le j < p^n \right\}.$$

Hence by Lemma 3.4.2, we know $S = T^0S, T^pS, \ldots T^{p^{n-1}}S, I = ST^0S, STS, \ldots, ST^{p^n-1}S$ forms a set of coset representatives for $\mathrm{PSL}_2(\mathbb{Z})/\Gamma_0(p^n)$. □

Exercise 3.4.5. *Complete the proof of Corollary 3.4.3 in the case where N is not a prime power.*

Exercise 3.4.6. *Find a complete set of coset representatives for* $\mathrm{PSL}_2(\mathbb{Z})/\Gamma_0(15)$.

Exercise 3.4.7. *Find* $[\mathrm{PSL}_2(\mathbb{Z}) : \Gamma_1(N)]$. *(Suggestion: determine the index inside* $\Gamma_0(N)$.*)*

Exercise 3.4.8. *Find* $[\mathrm{PSL}_2(\mathbb{Z}) : \Gamma(N)]$. *(Suggestion: see previous suggestion.)*

Lemma 3.4.9. *Let $\mathcal{F}$ be a fundamental domain for* $\mathrm{PSL}_2(\mathbb{Z})$*, and let Γ be a congruence subgroup of* $\mathrm{PSL}_2(\mathbb{Z})$*. Let $\{\alpha_i\}$ be a complete set of coset representatives for* $\mathrm{PSL}_2(\mathbb{Z})/\Gamma$ *such that $\mathcal{F}' = \bigcup \alpha_i^{-1}\mathcal{F}$ has connected interior. Then $\mathcal{F}'$ is a fundamental domain for Γ.*

Proof. Clearly the boundary of $\mathcal{F}'$ is a finite union of smooth curves, since the same is true of each $\alpha_i\mathcal{F}$.

Let $z \in \mathfrak{H}$. We show z is Γ-equivalent to a point in $\mathcal{F}'$. First, there is some $\tau \in \mathrm{PSL}_2(\mathbb{Z})$ such that $\tau z \in \mathcal{F}$. Write $\tau = \alpha_i\gamma$ for some α_i where $\gamma \in \Gamma$. Then $\alpha_i\gamma z \in \mathcal{F}$, so $\gamma z \in \alpha_i^{-1}\mathcal{F} \subseteq \mathcal{F}'$.

Now we want to show no two points z and z' in the interior $\mathcal{F}'^0$ of $\mathcal{F}'$ can be $\Gamma_0(N)$ equivalent. The basic idea is that the following. Suppose $z' = \gamma z$ for $\gamma \in \Gamma$, and write $z = \alpha_i^{-1}w$, $z' = \alpha_j^{-1}w'$ where $w, w' \in \mathcal{F}$. Hence $w' = \alpha_i\gamma\alpha_j^{-1}w$, i.e., w' and w are equivalent in $\mathrm{PSL}_2(\mathbb{Z})$. If w, w' are actually interior points of $\mathcal{F}$, this can only happen if $w' = w$, and one can use this to conclude $z' = z$. However it need not be that w, w' are interior points (e.g., examine the diagram in the example below), so instead we make the following argument to reduce to this case.

Let $\gamma \in \Gamma$ and suppose $U = \mathcal{F}'^0 \cap \gamma\mathcal{F}'^0$ is nonempty. For some α_i, $\alpha_i U \cap \mathcal{F}^0$ is nonempty. Say $w \in \alpha_i U \cap \mathcal{F}^0$. Then $z = \alpha_i^{-1}w \in \mathcal{F}'^0$ such that $z' = \gamma z \in \mathcal{F}'^0$. There exists α_j such that $w' = \alpha_j z' \in \mathcal{F}$. Thus we have $w' = \alpha_j\gamma\alpha_i^{-1}w$, i.e., w is $\mathrm{PSL}_2(\mathbb{Z})$-equivalent to w'. It is a simple exercise (below) that no interior point of $\mathcal{F}$ is $\mathrm{PSL}_2(\mathbb{Z})$-equivalent to a boundary point, hence $w' \in \mathcal{F}^0$.

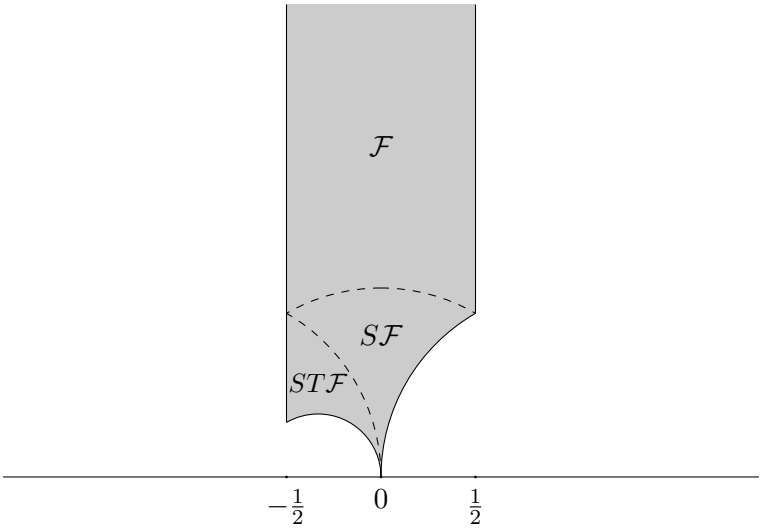
Since no two interior points of $\mathcal{F}$ are $\mathrm{PSL}_2(\mathbb{Z})$-equivalent, we have $w' = w$, i.e., w is a fixed point of $\alpha_j\gamma\alpha_i^{-1}$. However no interior points of $\mathcal{F}$ are fixed by a nontrivial element of $\mathrm{PSL}_2(\mathbb{Z})$ (either justify this to yourself, or see Lemma 3.5.1 below), whence $\alpha_j\gamma\alpha_i^{-1} = I$. This implies $\alpha_j\Gamma = \alpha_i\Gamma$, i.e., $i = j$, which then implies $\gamma = I$ so $z' = z$. □

Exercise 3.4.10. *Let Γ be a congruence subgroup of* $\mathrm{PSL}_2(\mathbb{Z})$ *and $\mathcal{F}$ a fundamental domain for Γ. Show that no interior point of $\mathcal{F}$ is Γ-equivalent to a boundary point.*

Example 3.4.11. *Consider* $\Gamma_0(2)$ *and let $\mathcal{F}$ be the standard fundamental domain for* $\mathrm{PSL}_2(\mathbb{Z})$*. By Corollary 3.4.4, a set of representatives for* $\mathrm{PSL}_2(\mathbb{Z})/\Gamma_0(2)$ *is I, S, STS. Hence by the above lemma, $\mathcal{F} \cup S\mathcal{F} \cup ST^{-1}S\mathcal{F}$ would be a fundamental domain for $\Gamma_0(2)$ if it had connected interior, but it does not (cf. diagram for Exercise 3.3.6). Instead, if we replace the coset representative $STS = T^{-1}ST^{-1}$ with $T^{-1}S$, and then apply the above lemma, we see*

$$\mathcal{F}' = \mathcal{F} \cup S\mathcal{F} \cup ST\mathcal{F},$$

pictured below, is a fundamental domain for $\Gamma_0(2)$.



Note that $\mathcal{F}'$ is convex in $\mathfrak{H}$.

***Exercise 3.4.12.** *Find a fundamental domain for* $\Gamma_0(4)$ *which contains the above fundamental domain for* $\Gamma_0(2)$.

3.5 Cusps and elliptic points

As we said in the previous section, in some sense the simplest kinds of modular forms (modular functions) will be (meromorphic) functions on the upper half-plane $\mathfrak{H}$ which are invariant under transformations in $\mathrm{PSL}_2(\mathbb{Z})$ or some congruence subgroup Γ. This means we can think of them as functions on the quotient space $X = \Gamma\backslash\mathfrak{H}$, or in other words, functions on a fundamental domain $\mathcal{F}$ for Γ which satisfy certain conditions on the boundary. (The space X is simply $\mathcal{F}$ with certain boundary points identified.)

There are 2 special kinds of points for X. First, let us discuss elliptic points. We say $z \in \mathcal{F}$ (or X) is an **elliptic point** if there exists $\gamma \in \Gamma - \{I\}$ such that $\gamma z = z$. In other words, the elliptic points for $\mathcal{F}$ are the points fixed by some (non-identity) transformation in Γ. We have already seen these play a role in finding a fundamental domain (in the proof of Lemma 3.4.9). The $\gamma \in \Gamma - \{I\}$ which fix some point of X (or equivalently $\mathcal{F}$) are called the **elliptic elements** of Γ.

For example if $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$ and $\mathcal{F}$ is the standard fundamental domain, then i is fixed by the transformation S. We will proceed to determine all elliptic points and elliptic elements for $\mathrm{PSL}_2(\mathbb{Z})$. First we observe the following general result.

Lemma 3.5.1. *Suppose* $\mathcal{F}$ *is a fundamental domain for* Γ *and* z *is an elliptic point of* $\mathcal{F}$. *Then* z *lies on the boundary* $\partial\mathcal{F}$ *of* $\mathcal{F}$.

Proof. Suppose z lies in the interior $\mathcal{F}^0$ of $\mathcal{F}$ and $\gamma \in \Gamma$ such that $\gamma z = z$. Then there exists a neighborhood U of z in $\mathcal{F}^0$ such that $\gamma U \subseteq \mathcal{F}^0$. Consequently, any $w \in U$ is Γ-equivalent with γw. However, by the definition of fundamental domain, this means $\gamma w = w$ for all $w \in U$, i.e., γ must act trivially on U. Since two holomorphic functions agreeing on an open set agree everywhere, this gives $\gamma = I$, i.e., z is not elliptic. □

Lemma 3.5.2. *The only elliptic points for the standard fundamental domain of* $\mathrm{PSL}_2(\mathbb{Z})$ *are* i, ζ_3 *and* ζ_6. *The elliptic subgroup fixing* i *is* $\{I, S\}$. *The subgroup fixing* ζ_3 *is* $\left\{I, ST, T^{-1}S\right\}$. *The subgroup fixing* ζ_6 *is* $\left\{I, ST^{-1}, TS\right\}$.

Observe by the diagram in Exercise 3.3.6, we can visually see the only possible elliptic elements of $\mathrm{PSL}_2(\mathbb{Z})$ are S, ST, ST^{-1}, STS, TST, $T^{-1}S$, TS, T^{-1} and T. Further T and T^{-1} clearly fix no points of $\mathfrak{H}$.

***Exercise 3.5.3.** *Prove the previous lemma (either using Exer 3.3.6 or not).*

Lemma 3.5.4. *Let* Γ *be a finite index subgroup of* $\mathrm{PSL}_2(\mathbb{Z})$ *and* $\mathcal{F}$ *a fundamental domain. Then any elliptic point of* $\mathcal{F}$ *must be of the form* $\gamma z \in \partial\mathcal{F}$ *where* $\gamma \in \mathrm{PSL}_2(\mathbb{Z})$ *and* $z \in \{i, \zeta_3, \zeta_6\}$. *i.e., it is a* $\mathrm{PSL}_2(\mathbb{Z})$*-translate of* i, ζ_3 *and* ζ_6 *lying on the boundary of* $\mathcal{F}$.

Further, such a γz *will be elliptic if and only if* $\gamma C \gamma^{-1} \subset \Gamma$ *where* C *is the stabilizer of* z *in* $\mathrm{PSL}_2(\mathbb{Z})$. *In this case,* $\gamma C \gamma^{-1}$ *is the stabilizer subgroup of* γz *in* Γ.

In particular, there are only finitely many elliptic points and elliptic elements for Γ.

Proof. Suppose $w \in \partial\mathcal{F}$ has nontrivial stabilizer C in Γ. Let $\mathcal{F}_0$ be the standard fundamental domain for $\mathrm{PSL}_2(\mathbb{Z})$. There is some $\tau \in \mathrm{PSL}_2(\mathbb{Z})$ such that $z = \tau w \in \mathcal{F}_0$. Consequently, $\tau C \tau^{-1}$ is a subgroup of $\mathrm{PSL}_2(\mathbb{Z})$ stabilizing z, i.e., z is an elliptic point in $\mathrm{PSL}_2(\mathbb{Z})$

On the other hand, let $z \in \{i, \zeta_3, \zeta_6\}$ and C be the subgroup (of order 2 or 3) which stabilizes z in $\mathrm{PSL}_2(\mathbb{Z})$. If $\gamma z \in \partial\mathcal{F}$ for some $\gamma \in \mathrm{PSL}_2(\mathbb{Z})$, then the stabilizer of γz in $\mathrm{PSL}_2(\mathbb{Z})$ is $\gamma C \gamma^{-1}$. Hence γz is elliptic for Γ if and only if $\gamma C \gamma^{-1} \subset \Gamma$. □

If z is an elliptic point for Γ, we say it is **elliptic of order** n if the stabilizing subgroup of z in Γ has order n. From the above lemma, we see that any elliptic point must have order 2 or 3 in $\Gamma \subseteq \mathrm{PSL}_2(\mathbb{Z})$

Exercise 3.5.5. *Determine the elliptic points for* $\Gamma_0(2)$ *and* $\Gamma_0(4)$ *(Use the fundamental domains from 3.4.11 and Exercise 3.4.12.*

Exercise 3.5.6. *Let* p *be a prime.*

(i) Show that the number of elliptic points of order 2 for $\Gamma_0(p)$ *is*

$$\epsilon_2(p) = 1 + \left(\frac{-1}{p}\right) = \begin{cases} 1 & p = 2 \\ 2 & p \equiv 1 \bmod 4 \\ 0 & p \equiv 3 \bmod 4. \end{cases}$$

(ii) Show that the number of elliptic points of order 3 for $\Gamma_0(p)$ *is*

$$\epsilon_3(p) = 1 + \left(\frac{-3}{p}\right) = \begin{cases} 1 & p = 3 \\ 2 & p \equiv 1 \bmod 3 \\ 0 & p \equiv 2 \bmod 3. \end{cases}$$

(iii) Write a table which determines the total number of elliptic points for $\Gamma_0(p)$ *depending only on the values of* p mod 12.

More generally (e.g., [DS05, Corollary 3.7.2]), the number of elliptic points of order 2 (resp. 3) for $\Gamma_0(N)$ is $\prod_{p|N} \epsilon_2(p)$ (resp. $\prod_{p|N} \epsilon_2(p)$).

We remark that elliptic points in $\mathcal{F}$ will be precisely the non-smooth points of the Riemann surface $X = \Gamma\backslash\mathfrak{H}$. To get some idea of the geometry that goes on, consider the case of $\mathcal{F}$ being the standard fundamental domain for $\mathrm{PSL}_2(\mathbb{Z})$. Around a non-elliptic point, one needs to go around 2π to make a closed loop. On the other hand, to make a closed loop around i, one only needs to go π radians. Similarly, to make closed loops around ζ_3 and ζ_6, one needs to go around $2\pi/3$ radians. In fact, the subgroup fixing any elliptic point z is cyclic of order m, and this subgroup is generated by an elliptic element which can be viewed as a hyperbolic rotation around z by $2\pi/m$.

While elliptic points do play a role in the theory of modular forms, more important for us will be the notion of the *cusps* of Γ. In fact, we need to know what cusps are to even give a definition of modular forms.

Thinking of the Riemann surface, $X = \Gamma\backslash\mathfrak{H}$ goes off to infinity in certain places. For example, think of $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$ and the standard fundamental domain $\mathcal{F}$. If we look at the definition of distance in $\mathfrak{H}$, it is clear the width of the fundamental domain $d\left(-\frac{1}{2} + iy, \frac{1}{2} + iy\right) \to 0$ as $y \to \infty$. (Alternatively, think about the fundamental domain $S\mathcal{F}$.) Hence if we think about some truncated fundamental domain, say $\mathcal{F}_{100} = \{z \in \mathcal{F} : \mathrm{Im}(z) > 100\}$, it looks like an half-infinite cylinder whose diameter is shrinking to an infinitesimal amount as $\mathrm{Im}(z) \to \infty$. This behavior at infinity is called a cusp. If we look at the fundamental domain for $\Gamma_0(2)$ in Example 3.4.11, it appears to have two cusps, one at $i\infty$ and one at 0. (Remember, distance to the real line is infinite in $\mathfrak{H}$.)

We will not be overly concerned with the geometry of X at its cusps (geometrically, they all look the same), but mainly in determining what they in terms of fundamental domains.

One way to think about what the cusps for X should be is to think about how to compactify it. Let us think about the case of $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$ with standard fundamental domain $\mathcal{F}$. By our remarks about the geometry of $\mathcal{F}$ with large imaginary part, it makes sense to look at the one-point compactification $\overline{\mathcal{F}} = \mathcal{F} \cup \{i\infty\}$ of $\mathcal{F}$. In other words, we look at $\overline{X} = \mathrm{PSL}_2(\mathbb{Z})\backslash\mathfrak{H} \cup \{i\infty\}$.

It would be nice to think of $\overline{X}$ itself as a quotient space, i.e., $X = \mathrm{PSL}_2(\mathbb{Z})\backslash\overline{\mathfrak{H}}$ where $\overline{\mathfrak{H}}$ is obtained from $\mathfrak{H}$ by adding possible cusps. Precisely, we define the **extended upper half-plane** $\overline{\mathfrak{H}} = \mathfrak{H} \cup \{i\infty\} \cup \mathbb{Q}$.

Now we extend the action of $\mathrm{PSL}_2(\mathbb{Z})$ to $\overline{\mathfrak{H}}$ as follows:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} i\infty = \lim_{y\to\infty} \frac{aiy + b}{ciy + d} = \begin{cases} \frac{a}{c} & c \neq 0 \\ i\infty & c = 0, \end{cases}$$

and

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \frac{p}{q} = \lim_{y\to 0^+} \frac{a\left(\frac{p}{q} + iy\right) + b}{c\left(\frac{p}{q} + iy\right) + d} = \begin{cases} \frac{ap+bq}{cp+dq} & cp + dq \neq 0 \\ i\infty & cp + dq = 0, \end{cases}$$

where $\frac{p}{q} \in \mathbb{Q}$. (Note some authors call $i\infty$ by ∞, though we feel the notation $i\infty$ is pictorially more suggestive. In fact, the natural compactifcation of $\mathfrak{H}$—most easily seen from the Poincaré disc model—is to simply take the one-point compactification of $\mathfrak{H} \cup \mathbb{R}$, so even if we think of the symbol ∞ meaning $\lim_{x\to\infty} = +\infty$, we still have $\infty = i\infty$.) That this is indeed an action (i.e., that the action associates with the group multiplication in $\mathrm{PSL}_2(\mathbb{Z})$) follows formally from the action on $\mathfrak{H}$.

If one wants to specify the topology on $\overline{\mathfrak{H}}$, we can start with defining a basis of open neighborhoods of $i\infty$ in $\overline{\mathfrak{H}}$ to be $\overline{U}_y = \{i\infty\} \cup \{z \in \mathfrak{H} : \mathrm{Im}(z) > y\}$ where $y > 0$. Then for $x \in \mathbb{Q}$, take $\gamma_x \in \mathrm{PSL}_2(\mathbb{Z})$ such that $\gamma i\infty = x$. A basis of open neighborhoods around x in $\overline{H}$ is then given by $\gamma_x \overline{U}_y$ where $y > 0$.

Exercise 3.5.7. *Show the open neighborhood $\gamma_x \overline{U}_y$ is a Euclidean circle in $\mathfrak{H} \cup \mathbb{R}$ which is tangent to $\mathbb{R}$ at x.*

Definition 3.5.8. *Let $\Gamma \subseteq \mathrm{PSL}_2(\mathbb{Z})$. The* **cusps** *of Γ (or $\Gamma\backslash\mathfrak{H}$) are the set of Γ-equivalence classes of $\{i\infty\} \cup \mathbb{Q}$.*

Example 3.5.9. *From the definition, we see any element of $\mathbb{Q}$ is $\mathrm{PSL}_2(\mathbb{Z})$-equivalent to $i\infty$. Hence, $\mathrm{PSL}_2(\mathbb{Z})$ has 1 cusp.*

Example 3.5.10. *Let $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(2)$. Then $c \equiv 0 \bmod 2$, so $a \equiv d \equiv 1 \bmod 2$. Hence $i\infty$ is $\Gamma_0(2)$ equivalent to a (nonzero) rational number $\frac{a}{c}$ in reduced form if and only if c is even. Similarly, 0 is $\Gamma_0(2)$ equivalent to a nonzero rational number $\frac{b}{d}$ in reduced form if and only if d is odd. Thus there are two cusps for $\Gamma_0(2)$, represented by 0 and $i\infty$. This agrees with the picture of the fundamental domain in Example 3.4.11.*

Lemma 3.5.11. *For any prime p, $\Gamma_0(p)$ has precisely 2 cusps, represented by 0 and $i\infty$.*

Proof. Simply replace 2 by p in the previous example. □

Exercise 3.5.12. *Draw a fundamental domain for $\Gamma_0(3)$ which has 0 and $i\infty$ as limit points.*

Lemma 3.5.13. *For any $\Gamma_0(N)$ the number of cusps is finite.*

This statement is true more generally for congruence subgroups, but we will restrict our proof to the case of modular groups.

Proof. Consider any rational number $\frac{p}{q}$ in reduced form. Because we only want to show the number of $\Gamma_0(N)$-equivalence classes of $\{i\infty\} \cup \mathbb{Q}$ is finite, we may assume $\frac{p}{q}$ is not $\Gamma_0(N)$-equivalent to $i\infty$.

There exist $c', d \in \mathbb{Z}$ relatively prime such that $c'pN + dq = \gcd(pN, q)$, which equals $\gcd(N, q)$ since $\gcd(p, q) = 1$. Further $c'pN + dq = \gcd(N, q)$ implies $\gcd(c, d) = 1$ where $c = c'N$. Thus there exist a, b such that $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N)$. Hence

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \frac{p}{q} = \frac{ap + bq}{cp + dq} = \frac{ap + bq}{\gcd(N, q)}.$$

Therefore replacing $\frac{p}{q}$ by something $\Gamma_0(N)$-equivalent, we may assume $q|N$.

Further, since $T \in \Gamma_0(N)$, we see

$$T^b \frac{p}{q} = \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} \frac{p}{q} = \frac{p+bq}{q}$$

is also $\Gamma_0(N)$-equivalent to $\frac{p}{q}$. Hence we can also assume $0 \leq p < q$. This gives only a finite number of possible non-equivalent $\frac{p}{q}$. □

One can refine the argument above to precisely count the number of cusps for $\Gamma_0(N)$.

Exercise 3.5.14. *Show that the number of cusps for $\Gamma_0(N)$ is given by*

$$\sum_{d|N} \phi(\gcd(d, N/d))$$

where ϕ is the Euler phi function.

***Exercise 3.5.15.** *Compute the cusps for $\Gamma_0(4)$.*

Chapter 4

Modular Forms

4.1 Modular curves and functions

Definition 4.1.1. *The* **modular curves of level** N, $Y_0(N)$ *and* $X_0(N)$, *are defined to be*

$$Y_0(N) = \Gamma_0(N)\backslash\mathfrak{H}$$

and

$$X_0(N) = \Gamma_0(N)\backslash\overline{\mathfrak{H}}.$$

In particular, $Y_0(1) = \mathrm{PSL}_2(\mathbb{Z})\backslash\mathfrak{H}$ and $X_0(1) = \mathrm{PSL}_2(\mathbb{Z})\backslash\overline{\mathfrak{H}}$.

Note $X_0(N)$ is the compactification of $Y_0(N)$ obtained by adding the cusps for $\Gamma_0(N)$, as described in the last chapter.

Exercise 4.1.2. *Show* $X_0(N)$, *with the quotient topology, is compact, using the topology on* $\overline{\mathfrak{H}}$ *described Section 3.5.*

Before we go further, let us explain the terminology a bit. In Section 2.5, we saw that $Y_0(1)$ parameterized lattices in $\mathbb{C}$, equivalently, isomorphism classes of (generalized) elliptic curves. (Some lattices would give a curve with discriminant 0, and these degenerate curves are called generalized elliptic curves). One can also define a generalized elliptic curve for the cusp in $X_0(1)$, and thus view $X_0(1)$ as parametrizing the space of (generalized) elliptic curves over $\mathbb{C}$, up to analytic isomorphism. Similarly, $X_0(N)$ parametrizes isomorphism classes of pairs (E, C) where E is a (generalized) elliptic curve over $\mathbb{C}$ and C is a cyclic subgroup of order N.

In algebraic geometry, a *moduli space* is a geometric object (curve, surface, manifold, etc.) whose points parametrize a class of other geometric objects, in this case elliptic curves with a distinguished cyclic subgroup. The Riemann surface $X_0(N)$ can also be viewed as an algebraic curve over $\mathbb{C}$, hence the term modular curve. Realizing $X_0(N)$ as a moduli space is fundamental in the theory of elliptic curves, though we will not focus on this aspect in this course.

For those familiar with Riemann surfaces, or complex manifolds, we can state the definition of a modular function very simply.

Definition 4.1.3. (Riemann surface version) *A* **modular function of level** N *is a meromorphic function* $f : X_0(N) \to \hat{\mathbb{C}}$.

For both readers familiar with Riemann surfaces and those not, let us translate this into a statement about functions on $\mathfrak{H}$.

First, think about f restricted to $Y_0(N) = \Gamma_0(N)\backslash\mathfrak{H}$. Being a meromorphic function on $Y_0(N)$ means we can lift f to a function $f : \mathfrak{H} \to \hat{\mathbb{C}}$ such that

(i) $f(\gamma z) = f(z)$ for all $\gamma \in \Gamma_0(N)$, and
(ii) f is meromorphic on $\mathfrak{H}$.

Now we need to explain what it means for f to be meromorphic at the cusps of $X_0(N)$. We first consider the notion of meromorphy at the cusp $\{i\infty\}$. It may be helpful at this point to recall the discussion at the end of Section 2.3 (though our notation here is different—f and F instead of g and G, and q instead of ζ).

Since $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in \Gamma_0(N)$, we know $f(Tz) = f(z+1) = f(z)$, i.e., f is periodic of period 1. Consequently, this means $f(z)$ has a Fourier expansion

$$f(z) = \sum_{n\in\mathbb{Z}} a_n e^{2\pi i n z}$$

where the n-th Fourier coefficient a_n is given by

$$a_n = \int_0^1 f(z)e^{-2\pi i n z}dx, \quad (z = x + iy)$$

valid for all $z \in \mathfrak{H}$. (Our argument at the end of Section 2.3 will justify that a_n does not depend upon y.) Put $q = e^{2\pi i z}$. Thus the Fourier expansion becomes

$$f(z) = \sum_{n\in\mathbb{Z}} a_n q^n. \tag{4.1.1}$$

This is also called the q**-expansion** for f. Note that the map $z = x+iy \mapsto q = e^{-2\pi y}e^{2\pi i x}$ is an analytic bijection from the vertical strip $\left\{z \in \mathfrak{H} : -\frac{1}{2} \le \mathrm{Im}(z) < \frac{1}{2}\right\}$ in $\mathfrak{H}$ to the punctured disc $D^\times = \{z \in \mathbb{C} : 0 < |z| < 1\}$. Therefore we may think of $f(z)$, a priori a periodic function on $\mathfrak{H}$, instead as a function $F(q)$ on $D^\times$.

If $z = iy$, then $q = e^{-2\pi y}$, so as $z \to i\infty$, $q \to 0$. Hence $f(z)$ being meromorphic at the cusp $z = i\infty$ precisely means that $F(q)$ is meromorphic at $q = 0$. If $F(q)$ is meromorphic at $q = 0$, then it has a Laurent expansion, which must equal the Fourier expansion (4.1.1). In other words $F(q)$ is meromorphic at $q = 0$ if and only if the Fourier coefficients a_n of f are independent of y and

(iii-a) the Fourier coefficients $a_n = 0$ for all but finitely many $n < 0$.

By (2.3.1), this is equivalent to the condition that $|f(z)|$ grows at most exponentially as $y \to \infty$, i.e.,

(iii-a') for $y \gg 0$, there exists m such that $|f(z)| < e^{my}$.

Now that we know what it means for $f(z)$ to be meromorphic at the cusp $i\infty$, we can use this to say what it means to be meromorphic at any cusp. Namely consider a cusp for $\Gamma_0(N)$ represented by $z_0 \in \mathbb{Q}$, and let $\tau \in \mathrm{PSL}_2(\mathbb{Z})$ be an element which maps $i\infty$ to z_0. Consider the function

$$f|_\tau(z) = f(\tau z).$$

Since τ is an isometry of $\mathfrak{H}$, $f|_\tau$ is also meromorphic on $\mathfrak{H}$. Thus we can view $|_\tau$ as an operator on meromorphic functions of $\mathfrak{H}$, called the *slash operator*.

Lemma 4.1.4. *Suppose $f(\gamma z) = f(z)$ for $\gamma \in \Gamma_0(N)$. Let $\tau \in \mathrm{PSL}_2(\mathbb{Z})$. Then $f|_\tau(z+N) = f|_\tau(z)$ for all $z \in \mathfrak{H}$.*

Proof. First note that $f|_\tau$ is left-invariant under $\tau^{-1}\Gamma_0(N)\tau$: for $\gamma \in \Gamma_0(N)$, we have

$$f|_\tau(\tau^{-1}\gamma\tau z) = f(\gamma\tau z) = f(\tau z) = f|_\tau(z).$$

On the other hand, note that the principal congruence subgroup

$$\Gamma(N) = \{\gamma \in \mathrm{PSL}_2(\mathbb{Z}) : \gamma \equiv I \bmod N\} \subseteq \Gamma_0(N)$$

is a normal subgroup of $\mathrm{PSL}_2(\mathbb{Z})$. In particular $\Gamma(N) \subseteq \tau^{-1}\Gamma_0(N)\tau$, so $f|_\tau$ is invariant under $T^N \in \Gamma(N)$. This proves the claim. □

This lemma means that $f|_\tau$ is also periodic with period N. (It may actually have smaller period if a smaller power of T lies in $\tau^{-1}\Gamma_0(N)\tau$.) Consequently, $f|_\tau$ has a Fourier expansion of the form

$$f|_\tau(z) = \sum_{n\in\mathbb{Z}} a_{\tau,n} q_N^n, \tag{4.1.2}$$

where

$$q_N = e^{2\pi i z/N}.$$

(If, for instance, $f|_\tau$ actually has period 1, we can write $f|_\tau(z) = \sum c_n q^n$, i.e., the $a_{\tau,n} = 0$ unless $N|n$.) What $f(z)$ to be meromorphic at the cusp z_0 means is precisely that $f|_\tau$ is meromorphic at $i\infty$. Now our discussion about what it meant for $f(z)$ to be meromorphic at $i\infty$ goes through for functions with period N just as well, and we see $f(z)$ is meromorphic at the cusp z_0 if and only if

(iii-b) the Fourier coefficients $a_{\tau,n} = 0$ for all but finitely many $n < 0$;

or equivalently,

(iii-b') for $y \gg 0$, there exists m such that $|f|_\tau(z)| < e^{my}$.

Observe that (4.1.2) can be used to give another series expansion for f. Namely, since $f(z) = f|_\tau(\tau^{-1}z)$, we can write

$$f(z) = \sum_{n\in\mathbb{Z}} a_{\tau,n} q_\tau^n, \quad q_\tau = e^{2\pi i \tau^{-1} z/N}. \tag{4.1.3}$$

This expansion, again valid for all $z \in \mathfrak{H}$, is called a **Fourier** (or $q-$) **expansion at** z_0 (with respect to τ) for $f(z)$. Consequently, we sometimes refer to the usual Fourier expansion (4.1.1) as the Fourier expansion at $i\infty$.

A priori, the Fourier expansion (4.1.3) at z_0, and consequently the conditions (iii-b) and (iii-b'), depend upon the choice of $\tau \in \mathrm{PSL}_2(\mathbb{Z})$ we used to send $i\infty$ to z_0. We would like to say all of these are independent of the choice of τ—and in fact the choice of z_0 representing the given cusp—and this is essentially what the following lemma tells us.

Lemma 4.1.5. *Suppose a meromorphic $f : \mathfrak{H} \to \hat{\mathbb{C}}$ satisfies $f(\gamma z) = f(z)$ for all $\gamma \in \Gamma_0(N)$. Let $z_0, z_0' \in \{i\infty\} \cup \mathbb{Q}$ represent the same cusp for $\Gamma_0(N)$, i.e., $z_0' = \gamma z_0$ for some $\gamma \in \Gamma_0(N)$. Now suppose $\tau, \tau' \in \mathrm{PSL}_2(\mathbb{Z})$ such that $\tau \cdot i\infty = z_0$ and $\tau' \cdot i\infty = z_0'$. Then $f|_\tau(z) = f|_{\tau'}(z+j)$ for some $j \in \mathbb{Z}$.*

Consequently, the Fourier coefficients with respect to τ and τ' are related by

$$a_{\tau',n} = a_{\tau,n} e^{2\pi i j/N}$$

for all n; in particular $|a_{\tau',n}| = |a_{\tau,n}|$ for all n.

Proof. First observe that $\gamma\tau$ is an element of $\mathrm{PSL}_2(\mathbb{Z})$ which sends $i\infty$ to z_0'. Therefore

$$(\tau')^{-1}\gamma\tau = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PSL}_2(\mathbb{Z})$$

which preserves $i\infty$. By the definition of our action of $\mathrm{PSL}_2(\mathbb{Z})$ on $\{i\infty\} \cup \mathbb{Q}$, this means $c = 0$, whence $a = c = \pm 1$ so $(\tau')^{-1}\gamma\tau = T^j$ for some $j \in \mathbb{Z}$, i.e., $\tau' = \gamma\tau T^{-j}$. Consequently

$$f|_{\tau'}(z) = f(\tau' z) = f(\gamma\tau T^{-j} z) = f(\tau T^{-j} z) = f|_\tau(T^{-j} z) = f|_\tau(z - j).$$

This gives the first part of the lemma.

Since we may also the last equation as $f|_{\tau'} = f|_{\tau T^{-j}}$, the Fourier expansion at z_0' (w.r.t. τ') for $f(\gamma z)$ is

$$f(z) = f(\gamma z) = \sum_{n \in \mathbb{Z}} a_{\tau',n} e^{2\pi i T^j \tau^{-1}\gamma^{-1}\gamma z/N} = \sum_{n \in \mathbb{Z}} a_{\tau',n} e^{2\pi i T^j \tau^{-1} z/N} = \sum_{n \in \mathbb{Z}} a_{\tau',n} e^{2\pi i j/N} e^{2\pi i \tau^{-1} z/N}.$$

Comparing this with the Fourier expansion at z_0 (w.r.t. τ) for $f(z)$ gives the second assertion. □

In other words, $f(z)$ may technically have more than one Fourier expansion at a given cusp, but any Fourier expansion is obtained from a given one by replacing the parameter $q_\tau = e^{2\pi i z/N}$ with a parameter of the form $q_{\gamma\tau T^{-j}} = e^{2\pi i(\tau\gamma^{-1} z + j)}$ for $j \in \mathbb{Z}$ and $\gamma \in \Gamma_0(N)$ and multiplying all the Fourier coefficients $a_{\tau,n}$ by a fixed root of unity $e^{2\pi i j/N}$.

Now we can recast Definition 4.1.3 in the language it is typically presented in most modular forms texts.

Definition 4.1.6. (Upper half-plane version) *We say a function $f : \mathfrak{H} \to \hat{\mathbb{C}}$ is a* **modular function of level** *N if*

(i) $f(\gamma z) = f(z)$ for all $\gamma \in \Gamma_0(N)$;

(ii) f is meromorphic on $\mathfrak{H}$; and

(iii) f is meromorphic at each cusp of $\Gamma_0(N)$; i.e., for each $\tau \in \mathrm{PSL}_2(\mathbb{Z})$, the Fourier coefficients $a_{\tau,n}$ as defined in (4.1.3) satisfy $a_{\tau,n} = 0$ for all but finitely many $n < 0$.

In practice one only needs to check the condition on negative Fourier coefficients $a_{\tau,n}$ as τ ranges over a finite set of elements of $\mathrm{PSL}_2(\mathbb{Z})$ such that $\tau \in i\infty$ runs over all cusps of $\Gamma_0(N)$. Further, our discussion above of course implies that (iii) is equivalent to

(iii') for each $\tau \in \mathrm{PSL}_2(\mathbb{Z})$, the function $f|_\tau$ is of **moderate growth** in $y = \mathrm{Im}(z)$, i.e., for y large, there exists m such that $|f|_\tau(z)| < e^{my}$.

Exercise 4.1.7. *Suppose f is a modular function of level 1 with Fourier expansion $f(z) = \sum_{-m}^{\infty} a_n q^n$. Show that for any $\tau \in \mathrm{PSL}_2(\mathbb{Z})$, the Fourier coefficients with respect to τ are the usual Fourier coefficients, i.e., $a_{\tau,n} = a_n$ for all n.*

Note that our above discussion goes through if we replace $\Gamma_0(N)$ by an arbitrary congruence subgroup Γ, as $\Gamma \supseteq \Gamma(N)$ for some N. Thus for any congruence subgroup Γ of $\mathrm{PSL}_2(\mathbb{Z})$, we can say $f : \mathfrak{H} \to \hat{\mathbb{C}}$ is a **modular function for** Γ by simply replacing $\Gamma_0(N)$ by Γ in the definition above.

Exercise 4.1.8. *Suppose f is a modular function for a congruence subgroup $\Gamma(N)$.*

(i) For $\tau \in \mathrm{PSL}_2(\mathbb{Z})$, show $f|_\tau$ is a modular function for $\tau^{-1}\Gamma(N)\tau$.

(ii) Deduce that $|_\tau : f \mapsto f|_\tau$ operates on the space of modular functions for $\Gamma(N)$.

(iii) Show $f|_\tau = f|_{\tau'}$ if $\tau' \in \Gamma\tau$. Hence at most $|\Gamma(N)\backslash\mathrm{PSL}_2(\mathbb{Z})|$ of the operators $|_\tau$ can be distinct.

(iv) Show $(f|_\tau)|_{\tau'} = |_{\tau\tau'}$, hence these operators give a right action of the quotient group $\Gamma(N)\backslash\mathrm{PSL}_2(\mathbb{Z})$ on the space of modular functions for $\Gamma(N)$.

There are a couple obvious things we can say now about modular functions.

Example 4.1.9. *Any constant function on $\mathfrak{H}$ is a modular function for any congruence subgroup Γ.*

Example 4.1.10. *Suppose Γ, Γ' are two congruence subgroups such that $\Gamma' \subseteq \Gamma$. If f is a modular function for Γ, then f is a modular function for Γ'. In particular, if f is a modular function of level N and $N|M$, then f is also a modular function of level M.*

Now of course we would like to know if some interesting (non-constant) modular functions exist. It turns out they are not trivial to construct. For example, let us try to naively construct a modular function for $\mathrm{PSL}_2(\mathbb{Z})$, i.e., a modular function of level 1. Let $g(z) = \frac{1}{z^n}$ for $n \in \mathbb{Z}$. Then we can try to average over the $\mathrm{PSL}_2(\mathbb{Z})$-translates

$$f(z) = \sum_{\gamma \in \mathrm{PSL}_2(\mathbb{Z})} g(\gamma z) = \frac{1}{(\gamma z)^n}.$$

We can see this diverges just by considering γ of the form ST^j. Such γ contribute

$$\sum_{j \in \mathbb{Z}} \frac{1}{(ST^j z)^n} = \sum \frac{1}{(S(z+j))^n} = \sum \frac{1}{(-1/(z+j))^n} = \sum_j (-1)^n (z+j)$$

to the average, but this already diverges.

An alternative approach might be the following. Since $\mathrm{PSL}_2(\mathbb{Z})$ is generated by S ($z \mapsto -1/z$) and T ($z \mapsto z+1$), we just need to find a function invariant under both of them. We

know $e^{2\pi iz}$ is invariant under T, so we could average this over the group $\langle S\rangle = \{I, S\}$. This would give us

$$e^{2\pi iz} + e^{2\pi iSz} = e^{2\pi iz} + e^{-2\pi i/z}.$$

Unfortunately, because of the second term, this is no longer invariant under T. (And if one tries to average over enough of $\mathrm{PSL}_2(\mathbb{Z})$ to make the sum formally $\mathrm{PSL}_2(\mathbb{Z})$-invariant, then it will diverge as in the first attempt.)

The easiest way to construct nontrivial modular functions will be to use modular forms. We will discuss this in the next section. First, we will present one result about nontrivial modular functions—they cannot be holomorphic!

Recall the following result of complex analysis.

Theorem 4.1.11. (Open Mapping Theorem) *Let $U \subseteq \mathbb{C}$ be an open set, and $f : U \to \mathbb{C}$ be holomorphic and nonconstant. Then f is an open map, i.e., f maps open sets to open sets.*

Corollary 4.1.12. *Let X be a compact Riemann surface, and $f : X \to \mathbb{C}$ holomorphic. Then f is constant.*

Proof. What $f : X \to \mathbb{C}$ holomorphic means is the following. Around any point $p \in X$, there is an open set U such that there is an analytic isomorphism $\iota U \to D$ with the open unit disc $D \subseteq \mathbb{C}$. For each such p, the map $f \circ \iota^{-1} : D \to \mathbb{C}$ is holomorphic at p.

Consequently, the Open Mapping Theorem says if f is not constant, $f(X)$ is an open subset of $\mathbb{C}$. However, since f is continuous and X is compact, $f(X)$ is also compact. But there are no open compact subsets of $\mathbb{C}$. □

Since $X_0(N)$ is compact (as is $\Gamma\backslash\overline{\mathfrak{H}}$ for any congruence subgroup Γ), there are no nonconstant holomorphic modular functions of level N (or for any congruence subgroup Γ). In fact, since the torus $\mathbb{C}/\Lambda$ is also a compact Riemann surface for a lattice Λ, the above corollary also tells us there are no nonconstant holomorphic elliptic functions with respect to Λ.

4.2 Eisenstein series

Even though we couldn't construct modular functions by an averaging process like we could for elliptic functions, the theory of elliptic functions does provide a subtle hint for constructing modular functions. Recall that for a lattice $\Lambda \subset \mathbb{C}$, the field of elliptic functions with respect to Λ was generate by the associate $\wp$-function and its derivative $\wp'$. We constructed $\wp'$ by an simple averaging technique, but the construction of $\wp$ was more complicated.

Consequently, one might ask if the derivatives of modular functions are simpler to construct than modular functions themselves. Suppose f is a modular function for Γ. Let $\gamma \in \Gamma$ and write $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then

$$\frac{d}{dz} f(z) = \frac{d}{dz} f(\gamma z) = f'(\gamma z) \cdot \frac{d}{dz}(\gamma z) = f'(\gamma z) \cdot \frac{d}{dz} \frac{az+b}{cz+d} = \frac{1}{(cz+d)^2} f'(\gamma z).$$

In other words, f' is no longer exactly invariant by Γ, but satisfies the transformation law

$$f'(\gamma z) = (cz+d)^2 f'(z).$$

Similarly if $g(z) = f^{(k)}(z)$, we have

$$g(\gamma z) = (cz+d)^{2k} g(z).$$

As I have suggested, it is easier to construct functions with these transformation properties than those which are strictly invariant under some Γ.

Definition 4.2.1. *Let $f : \mathfrak{H} \to \hat{\mathbb{C}}$ be meromorphic and Γ a congruence subgroup of $\mathrm{PSL}_2(\mathbb{Z})$. Let $k \in 2\mathbb{N} \cup \{0\}$. If*

$$f(\gamma z) = (cz+d)^k f(z), \quad \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma, \quad z \in \mathfrak{H},$$

then f is **weakly modular** *(or a* **weak modular form***) of* **weight** *k for Γ. If $\Gamma = \Gamma_0(N)$ we say f is weakly modular of weight k and* **level** *N.*

Note the factor $cz+d$ is not quite determined uniquely by a $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PSL}_2(\mathbb{Z})$, but only up to ± 1. However since k is assumed to be even, this creates no ambiguity in the definition above.

The adjective "weak" here refers to the fact that there is no condition at the cusps, like we had for modular functions.

(One can define weak modular forms, and modular forms, of odd weight k, just by replacing $\mathrm{PSL}_2(\mathbb{Z})$ with $\mathrm{SL}_2(\mathbb{Z})$. However for the same reason, they will not exist unless $\Gamma \subseteq \mathrm{SL}_2(\mathbb{Z})$ is a sufficiently small congruence subgroup such that $\gamma \in \Gamma \implies -\gamma \notin \Gamma$, or equivalently $-I \notin \Gamma$. This rules out the case of modular groups $\Gamma_0(N)$ viewed as subgroups of $\mathrm{SL}_2(\mathbb{Z})$, but not $\Gamma_1(N)$ or $\Gamma(N)$. Since our main focus in this course is modular forms for $\Gamma_0(N)$, we will not have much reason to consider these forms of odd weight.)

Lemma 4.2.2. *Suppose f (resp. g) is weakly modular of weight k (resp. l) for Γ.*

(i) If $k = l$ and $c \in \mathbb{C}$, then $cf + g$ is weakly modular of weight $k = l$ for Γ. Hence weakly modular functions of weight k form a complex vector space.

(ii) $f \cdot g$ is weakly modular of weight $k + l$ for Γ.

(iii) $\frac{f}{g}$ is weakly modular of weight $k - l$ for Γ provided $g \not\equiv 0$.

The proof is immediate.

Consequently, if we can construct two different weak modular forms of weight k for some k, then their quotient will be a nontrivial weak modular form of weight 0, which is a modular function provided the condition of meromorphic at the cusps is satisfied. We will come back to the condition at the cusps later, and use this to define meromorphic and holomorphic modular forms. For now, we will show how to construct weak modular forms of even weight $k > 2$.

The idea is basically to take a weighted average of some function $g(z) : \mathfrak{H} \to \hat{\mathbb{C}}$ over a congruence subgroup Γ. To explain this, for $\gamma \in \mathrm{PSL}_2(\mathbb{Z})$, we define the **automorphy factor** $j(\gamma, z) : \mathrm{PSL}_2(\mathbb{Z}) \times \mathfrak{H} \to \hat{\mathbb{C}}$ to be

$$j(\gamma, z) = cz + d, \quad \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}. \tag{4.2.1}$$

(As we remarked after Definition 4.2.1, $j(\gamma, z)$ is only defined up to ± 1, but as we will raise this to an even power, this causes no ambiguity.) For k even, consider the average

$$f(z) = \sum_{\gamma \in \Gamma} \frac{1}{(cz+d)^k} g(\gamma z) = \sum_{\gamma \in \Gamma} j(\gamma, z)^{-k} g(\gamma z).$$

Then we formally have

$$f(\gamma_0 z) = \sum_{\gamma \in \Gamma} j(\gamma, \gamma_0 z)^{-k} g(\gamma \gamma_0 z).$$

Now we will need the following property of the automorphy factor.

Lemma 4.2.3. *For $\gamma, \gamma' \in \mathrm{PSL}_2(\mathbb{Z})$, we have*

$$j(\gamma\gamma', z) = \pm j(\gamma', z) j(\gamma, \gamma' z).$$

Proof. Observe that (3.2.1) implies

$$\mathrm{Im}(\gamma z) = \frac{\mathrm{Im}(z)}{|j(\gamma, z)|^2}.$$

Hence

$$\frac{\mathrm{Im}(z)}{|j(\gamma\gamma', z)|^2} = \mathrm{Im}(\gamma\gamma' z) = \frac{\mathrm{Im}(\gamma' z)}{|j(\gamma, \gamma' z)|^2} = \frac{\mathrm{Im}(z)}{|j(\gamma, \gamma' z) j(\gamma', z)|^2}.$$

This shows the lemma is true up to taking absolute values.

Thus the lemma follows from the following claim: *Suppose f and g are meromorphic functions with $|f(z)|^2 = |g(z)|^2$. Then $f(z) = \zeta g(z)$ where $|\zeta| = 1$.* We may assume $g \not\equiv 0$. The hypothesis just says $f\overline{f} = g\overline{g}$, i.e., $f/g = \overline{g}/\overline{f}$. Hence the image of the meromorphic function f/g is contained in the set of $z \in \mathbb{C}$ such that $z = 1/\overline{z}$, i.e., $|z|^2 = 1$. By the Open Mapping Theorem (restrict to an open set where f/g is holomorphic), this is impossible unless f/g is constant. Then the absolute value condition implies $f(z) = \zeta g(z)$ where $|\zeta| = 1$.

This means

$$j(\gamma\gamma', z) = \zeta j(\gamma', z) j(\gamma, \gamma' z)$$

with $|\zeta = 1|$. Then the fact that $j(\gamma, z)$ is of the form $cz + d$ where $c, d \in \mathbb{Z}$ (so e.g., $j(\gamma, z) \in \mathbb{Z}$ for $z \in \mathbb{Z}$) implies that $\zeta = \pm 1$. □

By this lemma, we have

$$f(\gamma_0 z) = j(\gamma_0, z)^k \sum_{\gamma \in \Gamma} j(\gamma\gamma_0, z)^{-k} g(\gamma\gamma_0 z) = j(\gamma_0, z)^k \sum_{\gamma \in \Gamma} j(\gamma, z)^{-k} g(\gamma z) = j(\gamma_0, z)^k f(z),$$

where we replaced γ by $\gamma\gamma_0^{-1}$ in the middle step. In other words, provided $f(z)$ converges and is meromorphic, $f(z)$ is weakly modular of weight k.

However, as alluded to in the last section, averaging over all of Γ will be too much for $f(z)$ to converge. We also briefly toyed with a second idea—start with a periodic function and average over just what we need to yield the weak modularity condition for Γ. Let's restrict to the case $\Gamma = \Gamma_0(N)$ so $T \in \Gamma$. (Arbitrary $\Gamma \supseteq \Gamma(N)$ can be treated similarly since $T^N \in \Gamma(N)$.) Let P be the subgroup of $\mathrm{PSL}_2(\mathbb{Z})$ generated by T, i.e.,

$$P = \langle T \rangle = \left\{ \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix} : n \in \mathbb{Z} \right\} \subseteq \Gamma_0(N).$$

Note for $\gamma \in P$, $j(\gamma, z) = 1$, so any periodic function $g(z)$ with period 1 satisfies the weak modularity condition for T. Thus averaging over $P\backslash\Gamma_0(N)$ should, at least formally, yield something which is weakly modular.

First let's describe this coset space.

Lemma 4.2.4. *A complete set of representatives for $P\backslash\Gamma_0(N)$ is parametrized by the set*

$$\left\{(c,d) \in \mathbb{Z}^2 : c \equiv 0 \bmod N,\ \gcd(c,d) = 1\right\} / \pm 1,$$

where a coset representative with parameter (c,d) can be taken uniquely in the form $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N)$ with $0 \le b < |d|$, or $\begin{pmatrix} a & b \\ c & d \end{pmatrix} = S$ if $(c,d) = \pm(1,0)$ and $N = 1$.

Proof. Let $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N)$, so $ad - bc = 1$ and $c \equiv 0 \bmod N$. We can replace γ with the element

$$\begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a + cn & b + dn \\ c & d \end{pmatrix}$$

which lies in the same right P-coset of $\Gamma_0(N)$ as γ. First suppose $d = 0$. Then $-bc = 1$ so we may assume $b = -c = 1$, which can only happen if $N = 1$. Choosing $n = -a$ shows $\gamma \in P \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$.

Now suppose $d \neq 0$. Then we may (uniquely) choose n such that $0 \le b + dn < |d|$. Consequently we may assume $0 \le b < |d|$. On the other hand, the determinant condition says $a = \frac{1+bc}{d} \in \mathbb{Z}$, i.e., $bc \equiv -1 \bmod d$. This determines b, and consequently a, uniquely. □

So one idea might be to average the function $e^{2\pi i z}$ over this coset space, but again there are some issues with convergence. However, there's a simpler periodic function we could use—a constant function! In the $k = 0$ case, this obviously diverges, but if $k \ge 4$, it will converge.

Definition 4.2.5. *Let $k \ge 4$ be even. The* **(normalized) Eisenstein series of weight k and level N** *is defined to be*

$$E_{k,N}(z) = \sum_{\gamma \in P\backslash\Gamma_0(N)} j(\gamma, z)^{-k} = \frac{1}{2} \sum_{\substack{(c,d) \in \mathbb{Z}^2 \\ \gcd(c,d)=1 \\ c \equiv 0 \bmod N}} \frac{1}{(cz+d)^k}. \tag{4.2.2}$$

If $N = 1$ we often write E_k for $E_{k,N}$.

Note the second expression for $E_{k,N}$ in the definition follows (formally) from the above lemma, and the factor of $\frac{1}{2}$ comes from the fact that we include both (c, d) and $(-c, -d)$ in the sum.

Proposition 4.2.6. *The Eisenstein series $E_{k,N}$ converges absolutely and uniformly on compact subsets of $\mathfrak{H}$. Therefore, $E_{k,N}$ is a weak modular form of weight k and level N which is holomorphic on $\mathfrak{H}$.*

Proof. The discussion above shows that $E_{k,N}$ transforms formally under $\Gamma_0(N)$ as it should. Specifically, for $\gamma_0 \in \Gamma_0(N)$, we have

$$E_{k,N}(\gamma_0 z) = \sum_{\gamma \in P\backslash\Gamma_0(N)} j(\gamma, \gamma_0 z)^{-k} = j(\gamma_0, z)^k \sum_{\gamma \in P\backslash\Gamma_0(N)} j(\gamma\gamma_0, z)^{-k}.$$

Since right multiplying $P\backslash\Gamma_0(N)$ by γ_0^{-1} simply permutes the cosets $P\backslash\Gamma_0(N)$, we may replace γ by $\gamma\gamma_0^{-1}$ in the sum to see

$$E_{k,N}(\gamma_0 z) = j(\gamma_0, z)^k E_{k,N}(z). \tag{4.2.3}$$

Now let's deal with convergence. Since the series $E_{k,N}$ for $N > 1$ is a sum over a strictly smaller set of terms than the series for $E_k = E_{k,1}$, it suffices to prove absolute convergence for $E_k(z)$.

First suppose z lies in the standard fundamental domain $\mathcal{F}$ for $\mathrm{PSL}_2(\mathbb{Z})$. Then

$$|cz + d|^2 = (cz + d)(c\overline{z} + d) = c^2|z|^2 + 2cd\mathrm{Re}(z) + d^2.$$

Since $z \in \mathcal{F}$, $|z| \geq 1$ and $\mathrm{Re}(z) \leq \frac{1}{2}$ so

$$|cz + d|^2 \geq c^2 - |cd| + d^2 = \begin{cases} |c\zeta_3 + d|^2 & cd > 0 \\ |c\zeta_3 - d|^2 & cd < 0. \end{cases}$$

Hence for $z \in \mathcal{F}$,

$$|E_k(z)| \leq \sum_{\substack{c,d \geq 0 \\ \gcd(c,d)=1}} \frac{1}{|c\zeta_3 + d|^k} + \sum_{\substack{c \geq 0, d \leq 0 \\ \gcd(c,d)=1}} \frac{1}{|c\zeta_3 - d|^k} = 2 \sum_{\substack{c,d \geq 0 \\ \gcd(c,d)=1}} \frac{1}{|c\zeta_3 + d|^k}.$$

Note that $\{c\zeta_3 + d : c, d \in \mathbb{Z}\}$ form the lattice in $\mathbb{C}$ generated by 1 and ζ_3. Now we want to solve the lattice point problem of estimating how many $(c, d) \in \mathbb{Z}^2$ there are such that $c\zeta_3 + d$ lies inside the open disc D_r of radius r centered at the origin, i.e., bound the number of (c, d) such that

$$|c\zeta_3 + d| < r.$$

Observe that if $|c| \geq 2r$, then $|\mathrm{Im}(c\zeta_3 + d)| = |\mathrm{Im}(c\zeta_3)| = |c|\frac{\sqrt{3}}{2} > r$, i.e., $c\zeta_3 + d \notin D_r$. Similarly, looking at real parts show if $|c| < 2r$ then we must have $|d| < 2r$ if $c\zeta_3 + d \in D_r$.

I.e., the number of (c,d) such that $|c\zeta_3+d| \in D_r$ is at most $4r^2$. In particular, the number of (c,d) such that $|c\zeta_3+d| \in D_{r+1}-D_r$ is at most $4(r+1)^2$, and we can bound

$$|E_k(z)| \leq \sum_{\substack{(c,d)\in\mathbb{Z}^2 \\ (c,d)\neq(0,0)}} \frac{1}{|c\zeta_3+d|^k} \leq \sum_{\substack{(c,d)\in\mathbb{Z}^2 \\ r\leq|c\zeta_3+d|<r+1}} \frac{1}{r^k} \leq \sum_{r=1}^{\infty} \frac{4(r+1)^2}{r^k},$$

which converges for $k>3$. This establishes absolute convergence on $\mathcal{F}$.

Then the formal identity (4.2.3) implies we have absolute convergence at any point in $\mathfrak{H}$. Namely we can write any point of $\mathfrak{H}$ as γz for some $\gamma \in \mathrm{PSL}_2(\mathbb{Z})$ and $z \in \mathcal{F}$. Thus (4.2.3) says $|E_k(\gamma z)| = |j(\gamma,z)^k E_k(z)|$.

Since the bound above in $\mathcal{F}$ is independent of z, i.e., $E_k(z)$ is a bounded function on $\mathcal{F}$, it also implies uniform convergence of $E_k(z)$ and $E_{k,N}(z)$ on $\mathcal{F}$. For $E_k(z)$ this implies uniform convergence on compact sets of $\mathfrak{H}$ by the transformation property of $E_k(z)$ since for a fixed γ, $j(\gamma,z)$ is uniformly bounded on compact sets. Because the tail of a series for $E_{k,N}(z)$ can be bounded by the tail of a series for $E_k(z)$, this also implies $E_{k,N}(z)$ is uniformly bounded on compact subsets.

The theory of normal families or normal convergence in complex analysis says that if a series of analytic functions converges absolutely and uniformly on compact subsets, the limit function is also analytic. This finishes the proposition. □

For the future, let us record one more thing we got out of the above proof.

Corollary 4.2.7. *$E_{k,N}(z)$ is bounded on the standard fundamental domain for $\mathrm{PSL}_2(\mathbb{Z})$. In particular, $E_{k,N}(z)$ is bounded as $z \to i\infty$.*

Proof. The first statement comes directly from the proof. On the other hand, if $\mathrm{Im} z > 1$, then $T^j z \in \mathcal{F}$ for some j. Therefore $E_{k,N}(z) = j(T^j,z)^{-k} E_{k,N}(T^j z) = E_{k,N}(T^j z)$ is also bounded. □

This says $E_k(z)$ (as well as $E_{k,N}(z)$) should be "holomorphic at $i\infty$." Once we define holomorphy at the cusps in the next section, we will see these Eisenstein series are holomorphic at all cusps, making them holomorphic modular forms.

One interesting and elementary aspect of Eisenstein series is their Fourier expansion. Since $E_{k,N}(z+1) = E_{k,N}(z)$, it has a Fourier expansion

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

Here the sum starts from $n=0$ since $E_{k,N}(z)$ is holomorphic at $i\infty$. The idea is to use the following.

Lemma 4.2.8. (Lipschitz' formula) *Let $k \geq 2$ and $z \in \mathfrak{H}$. Then*

$$\sum_{n\in\mathbb{Z}} \frac{1}{(z+n)^k} = C_k \sum_{n=1}^{\infty} n^{k-1} q^n$$

where

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

Proof. Recall the product formula for sine:

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

Taking the logarithmic derivative of this gives

$$\pi \cot \pi z = \frac{1}{z} + \sum_{n=1}^{\infty} \left(\frac{1}{z+n} + \frac{1}{z-n}\right). \tag{4.2.4}$$

(Writing the cotangent formula this way as opposed to $\sum_{n\in\mathbb{Z}} \frac{1}{z+n}$ gives a series which is absolutely convergent.) On the other hand, since $\sin(\pi z) = \mathrm{Im}(\pi z) = \frac{e^{\pi iz} - e^{-\pi iz}}{2i}$ and $\cos(\pi z) = \mathrm{Re}(\pi z) = \frac{e^{\pi iz} + e^{-\pi iz}}{2}$, we have

$$\pi \cot \pi z = \pi i \frac{e^{\pi iz} + e^{-\pi iz}}{e^{\pi iz} - e^{-\pi iz}} = \pi i \frac{e^{2\pi iz} + 1}{e^{2\pi iz} - 1} = \pi i \frac{q+1}{q-1} = \pi i - \frac{2\pi i}{1-q} = \pi i - 2\pi i \sum_{n=0}^{\infty} q^n.$$

Taking the $(k-1)$-st derivative of each of these expressions for $\pi \cot \pi z$ gives the lemma. □

Comparing

$$E_{k,N}(z) = \frac{1}{2} \sum_{\substack{(c,d)\in\mathbb{Z}^2 \\ \gcd(c,d)=1 \\ c\equiv 0 \bmod N}} \frac{1}{(cz+d)^k}$$

with Lipschitz' formula, it would appear easier to compute the Fourier expansion for $E_{k,N}(z)$ without the condition that $\gcd(c,d) = 1$ in the above sum.

For simplicity, let's first consider the case of full level, i.e., $N = 1$. Define

$$\begin{aligned} G_k(z) &= \sum_{(0,0)\neq(c,d)\in\mathbb{Z}^2} \frac{1}{(cz+d)^k} \\ &= \sum_{n=1}^{\infty} \sum_{\substack{(c,d)\in\mathbb{Z}^2 \\ \gcd(c,d)=n}} \frac{1}{(cz+d)^k} \\ &= \sum_{n=1}^{\infty} \frac{1}{n^k} \sum_{\substack{(c,d)\in\mathbb{Z}^2 \\ \gcd(c,d)=1}} \frac{1}{(cz+d)^k} = 2\zeta(k) E_k(z), \end{aligned} \tag{4.2.5}$$

where $\zeta(s) = \sum \frac{1}{n^s}$ for $\mathrm{Re}(s) > 1$ is the Riemann zeta function. Since $G_k(z)$ is just a scalar multiple of $E_k(z)$, we also call $G_k(z)$ an (unnormalized) Eisenstein series. (In fact, in most classical treatments, one defines $G_k(z)$ before $E_k(z)$.)

Now applying the Lipschitz formula, we have

$$\begin{aligned} G_k(z) &= \sum_{(0,0)\neq(c,d)\in\mathbb{Z}^2} \frac{1}{(cz+d)^k} \\ &= \sum_{d\neq 0} \frac{1}{d^k} + 2\sum_{c=1}^{\infty}\sum_{d\in\mathbb{Z}} \frac{1}{(cz+d)^k} \\ &= 2\zeta(k) + 2C_k \sum_{c=1}^{\infty}\sum_{n=1}^{\infty} n^{k-1} e^{2\pi i c z n} \\ &= 2\zeta(k) + 2C_k \sum_{c=1}^{\infty}\sum_{n=1}^{\infty} n^{k-1} q^{cn}. \end{aligned}$$

Then what is the coefficient of a given q^m? There is a contribution from $n^{k-1}q^{cn}$ whenever $cn = m$, hence it is $2C_k\sigma_{k-1}(m)$ where $\sigma_k(m)$ is the classical divisor function

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

This establishes the following Fourier expansion.

Proposition 4.2.9. *Let $k \geq 4$ be even. Then we have the following Fourier expansion for the Eisenstein series of weight k.*

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

In terms of the normalized Eisenstein series, we can write this as

$$E_k(z) = 1 - \frac{2k}{B_k} \sum_{n=1}^{\infty} \sigma_{k-1}(n) q^n,$$

where B_k is the k-th Bernoulli number.

We recall the *Bernoulli numbers* can be defined as the coefficients in the expansion

$$\frac{x}{e^x - 1} = \sum_{k=0}^{\infty} B_k \frac{x^k}{k!}.$$

Alternatively we can define them recursively by $B_0 = 1$ and

$$B_k = -\sum_{j=0}^{k-1} \binom{k}{j} \frac{B_j}{k-j+1}.$$

Exercise 4.2.10. *Check these two definitions of B_k are equivalent.*

Proof. We already derived the expansion for $G_k(z)$. Dividing by $2\zeta(k)$ and using Euler's formula

$$\zeta(k) = -\frac{C_k B_k}{2k}$$

gives the expansion for $E_k(z)$. □

The fact that the Fourier coefficients of Eisenstein series are divisor functions—objects of study in elementary number theory—is our first real evidence that modular forms and functions are important in number theory. Later, we can use the theory of modular forms to prove results about these divisor functions, as well as relate them to problems in quadratic forms, as we discussed in the introduction.

At this point we can say why the E_k are called *normalized* Eisenstein series—they are normalized so that their leading Fourier coefficient is 1.[1] Since $z = i\infty$ corresponds to $q = 0$, the 0-th Fourier coefficient gives us the "value" of a periodic function at $i\infty$, i.e., $E_k(i\infty) = 1$.

Example 4.2.11.

$$E_4(z) = 1 + 240\sum_{n=1}^{\infty}\sigma_3(n)q^n.$$

$$E_6(z) = 1 - 504\sum_{n=1}^{\infty}\sigma_5(n)q^n.$$

$$E_8(z) = 1 + 480\sum_{n=1}^{\infty}\sigma_7(n)q^n.$$

Exercise 4.2.12. *From the first few examples, it looks like the Fourier coefficients for E_k might always be integers. This is not true (though they are clearly rational from Proposition 4.2.9). Find the first even k such that $\frac{2k}{B_k} \notin \mathbb{Z}$.*

To come back to the original question we asked as motivation—are derivatives of modular functions easier to construct directly than modular functions themselves?—we've only constructed holomorphic things which behave like derivatives of modular functions. Since non-constant modular functions are non-holomorphic, their derivatives are non-holomorphic and cannot be things like these Eisenstein series. However, we can still use these Eisenstein series (or more generally holomorphic modular forms, which we define below) to construct modular forms by taking quotients! (Of course, all this was just motivation for the funny transformation law in the definition of weak modular forms—we're really interested in modular forms in this course, not modular functions.) Here is an example that is important in the theory of elliptic curves.

***Exercise 4.2.13.** *We define the* **j-invariant** *to be*

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

Use the Fourier expansions for E_k to show j is a nonconstant modular function for $\mathrm{PSL}_2(\mathbb{Z})$.

[1] One can instead normalize so that $a_1 = 1$. This will have the benefit that the n-th Fourier coefficient is just $\sigma_{k-1}(n)$, so the a_n's with $n \geq 1$ are multiplicative. We will consider this other normalization in (6.0.1) and again in Chapter 8.

Since isomorphism classes of (generalized) elliptic curves are parametrized by $X_0(1) = \mathrm{PSL}_2(\mathbb{Z})\backslash\mathfrak{H}$, this gives a (nontrivial) invariant for elliptic curves which varies smoothly in the parameters defining the curve.

Remark. While it is obvious that E_k cannot converge if $k = 0$, it is natural to ask what happens if $k = 2$. The sum does not converge absolutely, but if we specify the order of summation, we can get a convergent function.

Exercise 4.2.14. *Define*

$$E_2(z) = \frac{1}{2\zeta(2)} \sum_{c=-\infty}^{\infty} \left(\sum_{d=-\infty}^{\infty}{}' \frac{1}{(cz+d)^2} \right),$$

where the prime on the inner sum means we omit the term $d = 0$ when $c = 0$.

(a) Using Lipshitz' formula on the inner sum first, show the resulting expression for $E_2(z)$ converges absolutely.

(b) Show

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

However $E_2(z)$ is not quite modular. It can be shown (e.g., [Kob93, p. 113]) that

$$E_2(-1/z) = z^2 E_2(z) + \frac{12}{2\pi i z}.$$

Now let us consider the case of higher level. The naive generalization to level N would be to consider

$$G_{k,N}(z) = \sum_{\substack{(0,0)\neq(c,d)\in\mathbb{Z}^2 \\ c\equiv 0 \bmod N}} \frac{1}{(cz+d)^k} = \sum_{n=1}^{\infty} \sum_{\substack{(c,d)\in\mathbb{Z}^2 \\ c\equiv 0 \bmod N \\ \gcd(c,d)=n}} \frac{1}{(cz+d)^k}, \tag{4.2.6}$$

but we can't write this in terms of $E_{k,N}(z)$ as simply as we did when $N = 1$. The reason is that if we rewrite the inner sum on the right as $\gcd(c, d) = 1$ and factor out a n^{-k}, we break the condition $c \equiv 0 \bmod N$ if $\gcd(n, N) \neq 1$.

Let's see what happens when $N = p$ is prime. Then

$$\begin{aligned} G_{k,p}(z) &= \sum_{\substack{(0,0)\neq(c,d)\in\mathbb{Z}^2 \\ c\equiv 0 \bmod p}} \frac{1}{(cz+d)^k} \\ &= \sum_{\substack{(c,d)\in\mathbb{Z}^2 \\ c\equiv 0 \bmod p \\ \gcd(d,p)=1}} \frac{1}{(cz+d)^k} + \sum_{\substack{(0,0)\neq(c,d)\in\mathbb{Z}^2 \\ c\equiv d\equiv 0 \bmod p}} \frac{1}{(cz+d)^k} \\ &= G^*_{k,p}(z) + \frac{1}{p^k} G_k(z), \end{aligned}$$

where

$$\begin{aligned} G^*_{k,p}(z) &= \sum_{\substack{(c,d)\in\mathbb{Z}^2 \\ c\equiv 0 \bmod p \\ \gcd(d,p)=1}} \frac{1}{(cz+d)^k} \\ &= \sum_{\substack{n>0 \\ \gcd(n,p)=1}} \sum_{\substack{(c,d)\in\mathbb{Z}^2 \\ c\equiv 0 \bmod p \\ \gcd(c,d)=n}} \frac{1}{(cz+d)^k} \\ &= \sum_{\substack{n>0 \\ \gcd(n,p)=1}} \frac{1}{n^k} \sum_{\substack{(c,d)\in\mathbb{Z}^2 \\ c\equiv 0 \bmod p \\ \gcd(c,d)=1}} \frac{1}{(cz+d)^k} \\ &= 2L(k,1_p)E_{k,p}(z) \end{aligned}$$

and $L(s,1_p)$ is the Dirichlet series associated to the trivial character mod p, i.e., for $\mathrm{Re}(s) > 1$,

$$L(s,1_p) = \sum_{\substack{n>0 \\ \gcd(n,p)=1}} \frac{1}{n^s} = \sum_{n=1}^{\infty} \frac{1}{n^s} - \sum_{n=1}^{\infty} \frac{1}{(pn)^s} = \left(1 - \frac{1}{p^s}\right)\zeta(s).$$

In summary, we have

$$E_{k,p}(z) = \left(1 - \frac{1}{p^k}\right)^{-1} \frac{1}{2\zeta(k)} G^*_{k,p}(z)$$

and

$$G^*_{k,p}(z) = G_{k,p}(z) - \frac{1}{p^k} G_k(z).$$

Since we already know the Fourier expansion for $G_k(z)$, it suffices to determine the Fourier expansion for $G_{k,p}(z)$. This is similar to the expansion for $G_k(z)$:

$$\begin{aligned} G_{k,p}(z) &= \sum_{\substack{(0,0)\neq(c,d)\in\mathbb{Z}^2 \\ c\equiv 0 \bmod p}} \frac{1}{(cz+d)^k} \\ &= \sum_{d\neq 0} \frac{1}{d^k} + 2\sum_{c=1}^{\infty} \sum_{d\in\mathbb{Z}} \frac{1}{(cpz+d)^k} \\ &= 2\zeta(k) + 2C_k \sum_{c=1}^{\infty} \sum_{n=1}^{\infty} n^{k-1} e^{2\pi i cpzn} \\ &= 2\zeta(k) + 2C_k \sum_{c=1}^{\infty} \sum_{n=1}^{\infty} n^{k-1} q^{cpn} \\ &= 2\zeta(k) + 2C_k \sum_{n=1}^{\infty} \sigma_{k-1}(n) q^{pn} = G_k(pz), \end{aligned}$$

Here the last equality follows from Proposition 4.2.9 together with the observation that if $f(z)$ has Fourier expansion $\sum a_n q^n$, then $f(pz)$ has Fourier expansion $a_n q^{pn}$.

Hence the above relations between $E_{k,p}$, $G^*_{k,p}$, $G_{k,p}$, G_k and E_k allow us to write $E_{k,p}$ in terms of E_k:

$$E_{k,p}(z)\left(1-\frac{1}{p^k}\right)^{-1}\left(E_k(pz)-\frac{1}{p^k}E_k(z)\right). \tag{4.2.7}$$

The above calculations also give the following Fourier expansions:

Proposition 4.2.15. *Let $N=p$ be prime, and $k\geq 4$ even. Then*

$$G^*_{k,p}(z)=2\left(1-\frac{1}{p^k}\right)\zeta(k)+2C_k\sum_{n=1}^{\infty}\left(\sigma_{k-1}(n/p)-\frac{\sigma_{k-1}(n)}{p^k}\right)q^n$$

and

$$E_{k,p}(z)=1-\frac{2k}{B_k}\left(1-\frac{1}{p^k}\right)^{-1}\sum_{n=1}^{\infty}\left(\sigma_{k-1}(n/p)-\frac{\sigma_{k-1}(n)}{p^k}\right)q^n.$$

(Here $\sigma_{k-1}(n/p)=0$ if $n/p\notin\mathbb{Z}$.)

***Exercise 4.2.16.** *Work out the Fourier expansion for $E_{k,N}(z)$ when $N=4$ and $k\geq 4$ even. Specifically, show*

$$G_{k,4}(z)=G^*_{k,4}(z)+\frac{1}{2^k}G_{k,2}(z)$$

where

$$G^*_{k,4}(z)=\sum_{\substack{c\equiv 0 \bmod 4\\ d \text{ odd}}}\frac{1}{(cz+d)^k}=2\left(1-\frac{1}{2^k}\right)\zeta(k)E_{k,4}(z).$$

Deduce that

$$E_{k,4}(z)=1-\frac{2k}{B_k}\left(1-\frac{1}{2^k}\right)^{-1}\sum_{n=1}^{\infty}\left(\sigma_{k-1}(n/4)-\frac{1}{2^k}\sigma_{k-1}(n/2)\right)q^n. \tag{4.2.8}$$

Comparing Fourier expansions, conclude

$$E_{k,4}(z)=E_{k,2}(2z).$$

To treat Eisenstein series of arbitrary level N, we need to break up the inner sum in the right of (4.2.6) into various pieces depending upon $\gcd(n,N)$. The trick is to use the Möbius function $\mu(n)$, i.e., $\mu(n)=-1$ (resp. 1) if n is squarefree and has an odd (resp. even) number of prime factors, and $\mu(n)=0$ if it is not squarefree. This is an important tool in the study of multiplicative functions[2] because of **Möbius inversion**: if f and g are multiplicative functions,

$$g(n)=\sum_{d|n}f(d)$$

implies

$$f(n)=\sum_{d|n}\mu(d)g(n/d).$$

[2]Recall $f:\mathbb{N}\to\mathbb{C}$ is *multiplicative* if $f(mn)=f(m)f(n)$ whenever $\gcd(m,n)=1$.

(This is like an arithmetic Fourier transform.)

For the sake of time and simplicity, we will not work out the Fourier expansions for arbitrary N, but simply state them and refer to [Sch74] (see also [Boy01], [DS05], [Kob93]) for details. For $k \geq 3$, let

$$\sigma_{k,N}(n) = \sum_{d|n} \frac{\mu(N/\gcd(d,N))}{\phi(N/\gcd(d,N))} d^k.$$

Then

$$E_{k,N}(z) = 1 - \frac{2k\phi(N)}{N^k B_k} \prod_{p|N} \left(1 - \frac{1}{p^k}\right)^{-1} \sum_{n=1}^{\infty} \sigma_{k-1,N}(n) q^n. \tag{4.2.9}$$

Exercise 4.2.17. *Check (4.2.9) agrees with Proposition 4.2.15 when N is prime.*

Exercise 4.2.18. *Using the identity $E_2(-1/z) = z^2 E_2(z) + \frac{12}{2\pi i z}$ mentioned in Exercise 4.2.14, show that for $N > 1$ prime, $E_{2,N}(z) := E_2(z) - N E_2(Nz)$ is a weak modular form of weight 2 and level N. Determine the Fourier expansion of $E_{2,N}(z)$. (It is not necessary that N be prime here, but we will define $E_{2,N}$ differently when N is not prime in Chapter 8.)*

We remark that for level $N > 1$ (or general Γ), one can construct a different Eisenstein series for each cusp of $\Gamma_0(N)$ (all holomorphic if $k \geq 4$). We'll discuss this some in the next chapter, but here is what it means in prime level. Recall from Lemma 3.5.11 that $\Gamma_0(p)$ has two cusps. This means means there should be a 2-dimensional space of Eisenstein series associated to $\Gamma_0(p)$. This space is generated by E_k and $E_{k,p}$ (or by (4.2.7), $E_k(z)$ and $E_k(pz)$)—cf. Example 4.3.10 below.

4.3 Modular forms

It is finally time for us to get to the full definition of modular forms. What we need to do is refine the notion of weak modularity to include a condition of meromorphy or holomorphy at the cusps.

First let's think about the Eisenstein series E_k ($k \geq 4$ even) of full level defined in the previous section, which we said should be modular forms. In this case, the relevant group of transformations is $\Gamma_0(1) = \mathrm{PSL}_2(\mathbb{Z})$ and there is only one cusp for $X_0(1)$. In Corollary 4.2.7, we saw that E_k is bounded as $z \to i\infty$, which means that E_k should be "holomorphic at $i\infty$." Or, thinking in terms of Fourier expansions, we saw $E_k(z) = 1 - \frac{2k}{B_k} \sum_{n=1}^{\infty} \sigma_{k-1}(n) q^n$, so E_k can be thought of as having the value 1 at $z = i\infty$ ($\leftrightarrow q = 0$).

On the other hand, any rational number also represents the cusp for $X_0(1)$. What happens to $E_k(z)$ as z tends to an element of $\mathbb{Q}$? The weak modularity condition actually forces $E_k(z)$ to be unbounded as $z \in \mathfrak{H}$ approaches a rational number z_0. To see this, say $\gamma \cdot i\infty = z_0$ and write $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then

$$\lim_{z \to z_0} E_k(z) = \lim_{z \to i\infty} E_k(\gamma z) = \lim_{z \to i\infty} (cz + d)^k E_k(z) = (c \cdot i\infty + d)^k \cdot 1 = \infty,$$

since one cannot have $c \neq 0$ if $z_0 \in \mathbb{Q}$. In other words, while E_k is "holomorphic" at $i\infty$, the weak modularity property (when $k \neq 0$) forces it to have a "pole" (which should be of order k) at all $z_0 \in \mathbb{Q}$.

While it perhaps makes sense to say that $E_k(z)$ is meromorphic at the cusp for $\mathrm{PSL}_2(\mathbb{Z})$ (it has at most a pole of finite order at each element of $\{i\infty\} \cup \mathbb{Q}$), it is not clear whether we should think of it as being holomorphic at the cusp. In fact, from what I said above, one might be inclined not to. However, $E_k(z)$ does have a Fourier expansion with no negative terms, and it is as close as possible to being holomorphic at the cusp given the weak modularity transformation property.

Bearing this in mind, one defines holomorphy at a cusp as follows. Similar to the $|_\gamma$ operator we defined in Section 4.1, we can deal with the automorphy factor by considering the *weight k slash operator* $|_{\gamma,k}$, defined for $\gamma \in \mathrm{PSL}_2(\mathbb{Z})$ and $k \in 2\mathbb{Z}_{\geq 0}$, which sends any $f : \mathfrak{H} \to \hat{\mathbb{C}}$ to

$$f|_{\gamma,k}(z) = j(\gamma, z)^{-k} f(\gamma z) = (cz + d)^{-k} f(\gamma z), \tag{4.3.1}$$

where $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Note $|_{\gamma,0}$ is precisely the operator $|_\gamma$ from Section 4.1. Observe that the weight k weak modular transformation property, $f(\gamma z) = j(\gamma, z) f(z) = (cz + d)^k f(z)$ for $\gamma \in \Gamma$, is equivalent to the statement

$$f|_{\gamma,k}(z) = f(z), \quad \gamma \in \Gamma. \tag{4.3.2}$$

In other words, a meromorphic $f : \mathfrak{H} \to \hat{\mathbb{C}}$ is a weak modular form of weight k if and only if (4.3.2) holds.

Consequently, while $E_k(z)$ is not holomorphic at any $z_0 \in \mathbb{Q}$, it is true that $E_k|_{\gamma,k}(z)$ is holomorphic at $i\infty$ for each $\gamma \in \mathrm{PSL}_2(\mathbb{Z})$. We will call this being holomorphic at the cusp. Or, more generally, we make the following definition.

Recall if f is a meromorphic, periodic function on $\mathfrak{H}$ with a period $N \in \mathbb{R}$, f has a Fourier expansion $f(z) = \sum_{n \in \mathbb{Z}} a_n q_N^n$ where $q_N = e^{2\pi i z/N}$. In Section 4.1, we said such an f is meromorphic at at $i\infty$ if $a_n = 0$ for all but finitely many $n < 0$, or alternatively, there exists m such that $|f(z)| < e^{my}$ for $y = \mathrm{Im}(z) \gg 0$. Similarly, we say such an f is **holomorphic at** $i\infty$ if $a_n = 0$ for all $n < 0$, or equivalently, if $f(z)$ is bounded for $y \gg 0$.

Definition 4.3.1. *Let $f : \mathfrak{H} \to \hat{\mathbb{C}}$ be a weak modular form of weight k for $\Gamma \subseteq \mathrm{PSL}_2(\mathbb{Z})$. We say f is* **meromorphic** *(resp.* **holomorphic***)* **at the cusps** *if $f|_{\tau,k}$ is meromorphic (resp. holomorphic) at $i\infty$ for each $\tau \in \mathrm{PSL}_2(\mathbb{Z})$.*

To see this definition makes sense, we need to know that $f|_{\tau,k}$ is also periodic with real period.

Exercise 4.3.2. *Let $\Gamma \supseteq \Gamma(N)$ be a congruence subgroup of $\mathrm{PSL}_2(\mathbb{Z})$ and suppose f is weakly modular of weight k for Γ. Show that $f|_{\tau,k}(z + N) = f|_{\tau,k}(z)$ for all $\tau \in \mathrm{PSL}_2(\mathbb{Z})$.*

We would also like to know that to check meromorphy/holomorphy at the cusps, it suffices to check it for a single representative of each cusp. This follows from the following generalization of Lemma 4.1.5.

For $\tau \in \mathrm{PSL}_2(\mathbb{Z})$ and f a weak modular form of weight k for $\Gamma \supseteq \Gamma(N)$, write the Fourier expansion (at $i\infty$) of $f|_{\tau,k}$ as

$$f|_{\tau,k}(z) = \sum_{n\in\mathbb{Z}} a_{\tau,n} q_N^n, \quad q_N = e^{2\pi i z/N}.$$

We call the $a_{\tau,n}$'s the **Fourier coefficients for f with respect to τ**. This is perhaps a slight misuse of notation, as, unlike in the case of modular functions, this does not actually give us a series expansion for $f(z)$ in $q_\tau = e^{2\pi i \tau^{-1} z/N}$, but rather just an expansion for $f|_{\tau,k}(z)$, or if one wishes, $j(\tau,z)^{-k} f(z)$. Indeed, $f(\tau z)$ will not generally have a real period, so there will not typically be a series expansion for f in terms of powers of q_τ.

Exercise 4.3.3. *Suppose a meromorphic $f : \mathfrak{H} \to \hat{\mathbb{C}}$ is weakly modular of weight k for a congruence subgroup $\Gamma \supseteq \Gamma(N)$. Let $z_0, z_0' \in \{i\infty\} \cup \mathbb{Q}$ represent the same cusp for Γ, i.e., $z_0' = \gamma z_0$ for some $\gamma \in \Gamma$. Now suppose $\tau, \tau' \in \mathrm{PSL}_2(\mathbb{Z})$ such that $\tau z_0 = \tau' z_0 = i\infty$.*

(i) Show $f|_{\tau,k}(z) = f|_{\tau',k}(z+j)$ for some $j \in \mathbb{Z}$.

(ii) Deduce the Fourier coefficients with respect to τ and τ' are related by

$$a_{\tau',n} = a_{\tau,n} e^{2\pi i j/N}$$

for all n; in particular $|a_{\tau',n}| = |a_{\tau,n}|$ for all n.

Both of these exercises are really no more than verifying the proofs of Lemma 4.1.4 and Lemma 4.1.5 go through in the setting $k > 0$.

Definition 4.3.4. *Let $f : \mathfrak{H} \to \hat{\mathbb{C}}$ be meromorphic, k even, and Γ a congruence subgroup of $\mathrm{PSL}_2(\mathbb{Z})$. We say f is a* **meromorphic modular form of weight k for Γ** *if*

(i) f is weakly modular of weight k for Γ, i.e., $f|_{\tau,k}(z) = f(z)$ for all $\tau \in \Gamma$; and

(ii) f is meromorphic at the cusps.

Definition 4.3.5. *Let $f : \mathfrak{H} \to \hat{\mathbb{C}}$ be meromorphic, k even, and Γ a congruence subgroup of $\mathrm{PSL}_2(\mathbb{Z})$. We say f is a* **(holomorphic) modular form of weight k for Γ** *if*

(i) f is weakly modular of weight k for Γ, i.e., $f|_{\tau,k}(z) = f(z)$ for all $\tau \in \Gamma$; and

(ii) f is holomorphic at the cusps.

The space of modular forms of weight k for Γ is denoted $M_k(\Gamma)$.

In other words, if we just say "modular form" we mean holomorphic modular form. This is (fairly) standard terminology (an older term you may sometimes see is "entire modular form"), though terminology for meromorphic modular forms varies and is less standardized. For example, they are called modular functions (even though they are not functions on $\Gamma\backslash\mathfrak{H}$!) by [Kob93] and automorphic forms (even though this does not agree with the standard usage of the term!) by [DS05].

As we have mentioned earlier, the modular forms we will be most interested in are modular forms for the modular groups $\Gamma_0(N)$. Unless otherwise specified, we will assume the weight $k \geq 2$ is even, so that nonzero modular forms exist on $\Gamma_0(N)$, and even non-constant ones provided $N > 1$ if $k = 2$ (see below).

Definition 4.3.6. *If f is a modular form of weight k for $\Gamma_0(N)$, we say f is a modular form of weight k and* **level** *N. The space of such forms is denoted $M_k(N)$.*

In other words, the notation $M_k(N)$ just means $M_k(\Gamma_0(N))$.

Example 4.3.7. *For Γ a congruence subgroup, any constant function lies in $M_0(\Gamma)$.*

Note the level of a modular form is not unique—the above example says a constant function is level N for any N. More generally, if $M|N$, then $\Gamma_0(M) \supset \Gamma_0(N)$ so it is clear from the definition that $M_k(M) \subset M_k(N)$. (In fact, we will see later that there are other embeddings besides the obvious one.) Yet more generally, if $\Gamma \subset \Gamma'$, then $M_k(\Gamma') \subset M_k(\Gamma)$.

Proposition 4.3.8. *For $k \geq 4$ even, the Eisenstein series $E_{k,N}$ is a modular form of weight k and level N.*

We remark this is also true for $k = 2$ and $N > 1$.

Proof. By Proposition 4.2.6, we know $E_{k,N}$ is weakly modular of weight k and level N, and holomorphic on $\mathfrak{H}$. Hence we just need to show $E_{k,N}$ is holomorphic at the cusps. (In fact, we already know from Section 4.2 that it is holomorphic at the cusp given by $i\infty$, and that $E_k(i\infty) = E_{k,p}(i\infty) = 1$.)

Let $\tau = \begin{pmatrix} r & s \\ t & u \end{pmatrix} \in \mathrm{PSL}_2(\mathbb{Z})$ and consider

$$\begin{aligned} E_{k,N}|_{\tau,k}(z) &= \frac{1}{2}\frac{1}{(tz+u)^k} \sum_{\substack{(c,d)\in\mathbb{Z}^2 \\ \gcd(c,d)=1 \\ c\equiv 0 \bmod N}} \frac{1}{\left(c\frac{rz+s}{tz+u}+d\right)^k} \\ &= \frac{1}{2} \sum_{\substack{(c,d)\in\mathbb{Z}^2 \\ \gcd(c,d)=1 \\ c\equiv 0 \bmod N}} \frac{1}{(c(rz+s)+d(tz+u))^k} \\ &= \frac{1}{2} \sum_{\substack{(c,d)\in\mathbb{Z}^2 \\ \gcd(c,d)=1 \\ c\equiv 0 \bmod N}} \frac{1}{((cr+dt)z+(cs+du))^k} \\ &= \frac{1}{2} \sum_{\substack{(c,d)\in\mathbb{Z}^2 \\ \gcd(c,d)=1 \\ c\equiv 0 \bmod N}} \frac{1}{(c'z+d')^k}, \end{aligned}$$

where in the last sum we have put $c' = cr + dt$ and $d' = cs + du$. Note we can view ${}^T\tau = \begin{pmatrix} r & t \\ s & u \end{pmatrix}$ as an element of $\mathrm{GL}_2(\mathbb{Z})$ acting on $\mathbb{Z} \times \mathbb{Z}$, and ${}^T\tau \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} c' \\ d' \end{pmatrix}$. Hence the final sum runs over a certain subset of pairs $(c', d') \in \mathbb{Z} \times \mathbb{Z} - \{(0,0)\}$. Thus

$$|E_{k,N}|_{\tau,k}(z)| \leq \frac{1}{2} \sum_{(c',d')\in\mathbb{Z}^2-\{(0,0)\}} \frac{1}{|c'z+d'|^k}.$$

However, we already showed this sum on the right is bounded in the proof of Proposition 4.2.6. □

Proposition 4.3.9. *For Γ a congruence subgroup and $k \geq 0$ even, $M_k(\Gamma)$ is a complex vector space.*

Proof. We show $M_k(\Gamma)$ is a subspace of the space of holomorphic functions on $\mathfrak{H}$.

Suppose $f \in M_k(\Gamma)$ and $c \in \mathbb{C}$. Then it is clear that $cf \in M_k(\Gamma)$. Similarly, if $f, g \in M_k(\Gamma)$, then $f + g$ clearly satisfies the weak modularity condition (cf. Lemma 4.2.2, and $f|_{\tau,k}$, $g|_{\tau,k}$ being bounded at $i\infty$ for each $\tau \in \mathrm{PSL}_2(\mathbb{Z})$ implies $(f+g)|_{\tau,k} = f|_{\tau,k} + g|_{\tau,k}$ is also bounded at $i\infty$, i.e., $f + g$ is also holomorphic at the cusps. □

Example 4.3.10. *From Proposition 4.3.8 and the observation that $M_k(1) \subset M_k(p)$, we have $E_k(z), E_{k,p}(z) \in M_k(p)$. Hence by (4.2.7) we also have $E_k(pz) \in M_k(p)$. The subspace of $M_k(p)$ generated by two of these Eisenstein series is called the* Eisenstein subspace *of $M_k(p)$. We will see later that for k small the Eisenstein subspace makes up all of $M_k(p)$ but for k large there will be other forms in $M_k(p)$ due to the existence of cusp forms.*

There are two main reasons we work with holomorphic forms, rather than meromorphic forms. The first is that restricting to holomorphic forms, the spaces $M_k(\Gamma)$ are *finite-dimensional* vector spaces. (This is also why one needs the condition of being holomorphic at the cusps.) One of the main applications of modular forms, say to the theory of quadratic forms as presented in the introduction, is to construct a modular form in two different ways and show that the constructions are equal. Knowing that $M_k(\Gamma)$ is finite dimensional means that one can check two modular forms are identical by checking that a finite number of their Fourier coefficients agree. We will explore these ideas some in the rest of this chapter, and explain this precisely in the next chapter.

One simple but important way to construct modular forms is the following.

Lemma 4.3.11. *Let Γ be a congruence subgroup, and $f \in M_k(\Gamma)$, $g \in M_l(\Gamma)$. Then $fg \in M_{k+l}(\Gamma)$.*

Proof. We already observed that $f + g$ is weakly modular of weight $k + l$ in Lemma 4.2.2. Then, as in the previous proof, if $f|\tau, k$ and $g|_{\tau,l}$ are bounded at $i\infty$, we see $(fg)|_{\tau,k+l} = (f|_{\tau,k})\,(g|_{\tau,l})$ is also bounded at $i\infty$, i.e., fg is also holomorphic at the cusps. □

The second reason to restrict our study to holomorphic modular forms is that meromorphic modular forms can be simply constructed as quotients of holomorphic modular forms (cf. exercise below), and, at least on the full modular group (cf. [FB09, Exercise VI.3.5]), all meromorphic modular forms arise this way. (If you find a reference which discusses this for other congruence subgroups, or has the complete proof for $\mathrm{PSL}_2(\mathbb{Z})$, please let me know.) If one restricts to meromorphic modular forms of weight 0 for Γ, i.e., modular functions on $\Gamma\backslash\overline{\mathfrak{H}}$, then basic function theory for compact Riemann surfaces implies any modular function is a quotient of two elements of $M_k(\Gamma)$ for some k. We will prove something more precise in the case of $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$ later.

Exercise 4.3.12. *Let Γ be a congruence subgroup, $f \in M_k(\Gamma)$ and $g \in M_l(\Gamma)$. Show $\frac{f}{g}$ is a meromorphic modular form of weight $k - l$ provided $g \not\equiv 0$.*

Now let's look at an example of the above lemma.

Example 4.3.13. *Consider the Eisenstein series*

$$E_4(z) = 1 + 240 \sum_{n=1}^{\infty} \sigma_3(n) q^n \in M_4(1).$$

Then

$$\begin{aligned} E_4^2(z) &= 1 + 480 \sum_{m=1}^{\infty} \sigma_3(m) q^m + (240)^2 \sum_{m,n=1}^{\infty} \sigma_3(m)\sigma_3(n) q^{m+n} \\ &= 1 + 480 \sum_{n=1}^{\infty} \left(\sigma_3(n) + 120 \sum_{m=1}^{n-1} \sigma_3(m)\sigma_3(n-m) \right) q^n \in M_8(1). \end{aligned}$$

On the other hand,

$$E_8(z) = 1 + 480 \sum_{n=1}^{\infty} \sigma_7(n) q^n \in M_8(1).$$

Computing the first few Fourier coefficients of E_4^2,

$$E_4(z)^2 = 1 + 480 \left(q + 129q^2 + 2188q^3 + \cdots \right),$$

we see the first several Fourier coefficients of E_4^2 match with those of E_8. This suggests

$$E_4^2 = E_8,$$

which implies the sum-of-divisors identity

$$\sigma_7(n) = \sigma_3(n) + 120 \sum_{m=1}^{n-1} \sigma_3(m)\sigma_3(n-m).$$

In particular, this would mean

$$\sigma_7(n) \equiv \sigma_3(n) \bmod 120.$$

In the next chapter, we will see that $M_8(1)$ is 1 dimensional, so just knowing the 0-th Fourier coefficients (the "constant terms") match is enough to deduce $E_4^2 = E_8$. While one can prove this and similar sum-of-divisors identities with just elementary number theory, it is not at all simple.

4.4 Theta series

The motivation we presented in the introduction for studying modular forms is the application to the theory of quadratic forms, though we have also seen that they arise in connection with the function theory of the modular curves $X_0(N)$. (The reader may want to reread the introduction at this point.) The passage from quadratic forms to modular forms comes via *theta series*.

Jacobi seems to have begun the study theta functions in connection with elliptic functions. Specifically, in 1829, he obtained an expression for the Weierstrass $\wp$-functions in

terms of theta functions. This has become important in mathematical physics. However, our interest in theta functions will be in relation to modular and quadratic forms.

Recall from the introduction, **Jacobi's theta function**

$$\vartheta(z) = \sum_{n=-\infty}^{\infty} q^{n^2}, \tag{4.4.1}$$

where, as usual, $q = e^{2\pi i z}$. Thinking of this as a power series in the parameter q, it is clear this series converges absolutely and uniformly on compact sets in $\mathfrak{H}$, hence it is a meromorphic function on $\mathfrak{H}$ with period 1 (w.r.t. z). Then

$$\vartheta^k(z) = \prod_{i=1}^{k} \left(\sum_{n_i=-\infty}^{\infty} q^{n_i^2} \right) = \sum_{n_1,\ldots,n_k \in \mathbb{Z}} q^{n_1^2+n_2^2+\cdots+n_k^2} = \sum_{n \geq 0} r_k(n) q^n, \tag{4.4.2}$$

where $r_k(n)$ is the number of ways to write n as a sum of k squares, i.e.,

$$r_k(n) = \#\left\{ (x_1, \ldots, x_k) \in \mathbb{Z}^k : x_1^2 + x_2^2 + \cdots + x_k^2 = n \right\}.$$

In other words, the Fourier coefficients of ϑ^k tell us the number of ways n can be written as a sum of k squares. We would like to say this is a modular form.

Proposition 4.4.1. *Jacobi's theta function satisfies the transformation laws*

$$\vartheta(z+1) = \vartheta(z), \quad \vartheta\left(\frac{-1}{4z}\right) = \sqrt{\frac{2z}{i}}\vartheta(z). \tag{4.4.3}$$

Proof. The first equation is obvious from the definition as q is invariant under $z \mapsto z+1$.

The idea to prove the second equation is to use Poisson summation, which says for any Schwartz function (i.e., rapidly decreasing function) $f : \mathbb{R} \to \mathbb{R}$, we have

$$\sum_{n \in \mathbb{Z}} f(n) = \sum_{n \in \mathbb{Z}} \hat{f}(n),$$

where

$$\hat{f}(t) = \int_{-\infty}^{\infty} e^{2\pi i s t} f(s) ds$$

is the Fourier transform of f. (See any introduction to Fourier analysis for a proof of Poisson summation, which is a simple consequence of Fourier inversion.)

Note that we can write (for fixed $z \in \mathfrak{H}$)

$$\vartheta(z) = \sum_{n \in \mathbb{Z}} f(n)$$

where

$$f(t) = e^{2\pi i z t^2}.$$

Now we will assume $z = iy$ with $y > 0$, so that $f(t) = e^{-2\pi y t^2}$ is Schwartz.

We compute

$$\hat{f}(t) = \int_{-\infty}^{\infty} e^{2\pi i(st+s^2z)} ds.$$

Completing the square, we see

$$s^2z + st = z\left(s + \frac{t}{2z}\right)^2 - \frac{t^2}{4z}$$

Let $u = (s + t/2z)$. Then

$$\hat{f}(t) = e^{-\pi i t^2/2z} \int_{-\infty}^{\infty} e^{2\pi i z u^2} du = e^{\pi t^2/2y} \int_{-\infty}^{\infty} e^{-2\pi y u^2} du = \sqrt{\frac{1}{2y}} e^{\pi t^2/2y}.$$

Here the last equality follows from the substitution $v = \sqrt{2y}u$, which reduces the problem to the well known Gaussian integral

$$\int_{-\infty}^{\infty} e^{-\pi v^2} dv = 1.$$

Now applying Poisson summation proves $\vartheta\left(\frac{-1}{4z}\right) = \sqrt{\frac{2z}{i}}\vartheta(z)$ when $z = iy$. However, since both sides are holomorphic functions, knowing they agree on a set with an accumulation point implies they agree everywhere. □

Lemma 4.4.2. *$\Gamma_0(4)$ is generated by T and $\widetilde{T} = \begin{pmatrix} 1 & 0 \\ 4 & 1 \end{pmatrix}$.*

Proof. Suppose $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(4)$. Note

$$\gamma T^k = \begin{pmatrix} a & b+ka \\ c & d+kc \end{pmatrix}.$$

Since $\gcd(a, c) = 1$ but $c \equiv 0 \bmod 4$, we know a is odd. In particular $a \neq 0$, so by replacing γ by some γT^k, we may assume $|b| \leq \frac{|a|}{2}$. Further, a odd means $|b| \neq \frac{|a|}{2}$, i.e., $|b| < \frac{|a|}{2}$.

On the other hand,

$$\gamma \widetilde{T}^k = \begin{pmatrix} a+4kb & b \\ c+4kd & d \end{pmatrix}.$$

If $|b| \neq 0$, note $|a + 4kb| < 2|b|$ for some $k \in \mathbb{Z}$, so we upon replacing γ by $\gamma\widetilde{T}^k$, we may assume $|a| < 2b$.

Continuing in this manner of replacing γ by elements of the form γT^k and $\gamma\widetilde{T}^k$ alternately, we eventually reduce to the situation $b = 0$ by Fermat descent. (Observe in the above replacements, when we reduce $|b|$ we do not change a, and when we reduce $|a|$ we do not change b, so this process terminates in a finite number of steps.)

But if $b = 0$, by the determinant condition and the definition of $\Gamma_0(4)$, we know that (up to ± 1) γ must be of the form

$$\begin{pmatrix} 1 & 0 \\ 4k & 1 \end{pmatrix} = \widetilde{T}^k.$$

□

The following exercise is a simple consequence of Lemma 4.2.3.

Exercise 4.4.3. *Suppose* $\Gamma = \langle \gamma_1, \dots, \gamma_r \rangle$ *and let* $f : \mathfrak{H} \to \hat{\mathbb{C}}$ *be meromorphic. Show that if*

$$f(\gamma_i z) = j(\gamma_i, z)^k f(z)$$

for $i = 1, \dots, r$*, then* f *is weakly modular of weight* k *for* Γ*.*

Proposition 4.4.4. *Let* $k \in \mathbb{N}$ *such that* $k \equiv 0 \bmod 4$*. Then* $\vartheta^k \in M_{k/2}(4)$*.*

Proof. We know ϑ^k is holomorphic on $\mathfrak{H}$ because ϑ is. By the above lemma and exercise, to show ϑ^k is weakly modular of weight $\frac{k}{2}$ for $\Gamma_0(4)$, it suffices to show

$$\vartheta^k(z+1) = \vartheta^k(Tz) = j(T,z)^{k/2}\vartheta^k(z) = \vartheta^k(z)$$

and

$$\vartheta^k\left(\frac{z}{4z+1}\right) = \vartheta^k(\widetilde{T}z) = j(\widetilde{T},z)^{k/2}\vartheta^k(z) = (4z+1)^{k/2}\vartheta^k(z).$$

The former is clear, since $\vartheta(z+1) = \vartheta(z)$. To see the second, put $w = \frac{z}{4z+1}$ and observe the second part of (4.4.3) implies

$$\begin{aligned}
\vartheta^k\left(\frac{z}{4z+1}\right) = \vartheta^k(w) &= \left(\frac{i}{2w}\right)^{k/2} \vartheta^k\left(\frac{-1}{4w}\right) \\
&= \left(\frac{i}{2w}\right)^{k/2} \vartheta^k\left(-1 - \frac{1}{4z}\right) \\
&= \left(\frac{i}{2w}\right)^{k/2} \vartheta^k\left(\frac{-1}{4z}\right) \\
&= \left(\frac{i}{2w}\right)^{k/2} \left(\frac{2z}{i}\right)^{k/2} \vartheta^k(z) \\
&= (4z+1)^{k/2}\vartheta^k(z).
\end{aligned}$$

Thus it remains to show ϑ^k is holomorphic at the cusps. By Corollary 3.4.4, we can take for a set of representatives of $\mathrm{PSL}_2(\mathbb{Z})/\Gamma_0(4)$ the elements

$$I, \quad S, \quad T^2S = \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix}, \quad \text{and} \quad ST^jS = \begin{pmatrix} 1 & 0 \\ -j & 1 \end{pmatrix} \quad \text{for } 1 \le j \le 3.$$

We know ϑ is holomorphic at $i\infty$ by its Fourier expansion, so we need to show $f|_{\gamma,k/2}$ is holomorphic at $i\infty$ for $\gamma = S, STS$ and T^2S. Note by (4.4.3) we have

$$\vartheta^k|_{S,k/2}(z) = \left(\frac{1}{z}\right)^{k/2} \vartheta^k\left(\frac{-1}{z}\right) = \left(\frac{1}{z}\right)^{k/2} \left(\frac{z}{2i}\right)^{k/2} \vartheta^k\left(\frac{z}{4}\right) = \left(\frac{1}{2i}\right)^{k/2} \vartheta^k\left(\frac{z}{4}\right),$$

which is holomorphic at $i\infty$ since $\vartheta^k(z)$ is. We leave the cases of $\gamma = ST^jS$ and $\gamma = T^2S$ as an exercise. □

Exercise 4.4.5. *Complete the proof of the above proposition by showing that, for $k \equiv 0 \bmod 4$ with $k > 0$, $\vartheta^k|_{\gamma,k/2}$ is holomorphic at all cusps. (You can either check that it is holomorphic at $i\infty$ for $\gamma = T^2S$ and each $\gamma = ST^jS$, $1 \le j \le 3$, or check holomorphy at each individual cusp.)*

One might like to say that ϑ^k is a modular form of odd weight when $k \equiv 2 \bmod 4$, and a modular form of *half-integral weight* when k is odd. These notions can be made precise, but the transformation group can no longer be $\Gamma_0(4)$ (see parenthetical remarks after Definition 4.2.1). We will not treat the theory of modular forms of odd or half-integral weights in this course (see, e.g., [Kob93]), but leave the case of ϑ^k when $k \equiv 2 \bmod 4$ as an exercise.

Exercise 4.4.6. *Let Γ be a finite index subgroup of $\mathrm{SL}_2(\mathbb{Z})$. Then for $\gamma \in \Gamma$, $j(\gamma, z)$ is well defined (not just defined up to ± 1).*

Let $k \in \mathbb{Z}$. We say a holomorphic function $f : \mathfrak{H} \to \mathbb{C}$ is a modular form of weight k for Γ, if $f|_{\gamma,k}(z) = f(z)$ for all $\gamma \in \Gamma$ and $z \in \mathfrak{H}$, and $f|_{\tau,k}(z)$ is holomorphic at $i\infty$ for all $\tau \in \mathrm{SL}_2(\mathbb{Z})$.

(a) Suppose $-I \in \Gamma$. Show there are no nonzero modular forms of odd weight for Γ.

(b) Consider

$$\Gamma = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \equiv \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} \bmod 4 \right\} \subset \mathrm{SL}_2(\mathbb{Z}),$$

i.e., $\Gamma = \Gamma_1(4)$ viewed as a subgroup of $\mathrm{SL}_2(\mathbb{Z})$ (not $\mathrm{PSL}_2(\mathbb{Z})$). For any $k \ge 2$ even, show ϑ^k is a modular form of weight $k/2$ for Γ.

Example 4.4.7. *From the above proposition, we know*

$$\vartheta^8(z) = 1 + 16q + 112q^2 + 448q^3 + \cdots \in M_4(\Gamma_0(4)).$$

(There is 1 way to write 0 as a sum of 4 squares, 16 ways to write $1 = (\pm 1)^2 + 0^2 + 0^2 + \cdots + 0^2$ as a sum of 8 squares counting symmetries, and so on.)

From Exercise 4.2.16, we know

$$E_{4,4}(z) = 1 + 16 \sum_{n=1}^{\infty} \left(16\sigma_3(n/4) - \sigma_3(n/2)\right) q^n = 1 - 16q^2 + 112q^4 - \cdots \in M_4(4).$$

We also have

$$E_{4,2}(z) = 1 + 16 \sum_{n=1}^{\infty} \left(16\sigma_3(n/2) - \sigma_3(n)\right) q^n = 1 - 16q + 112q^2 - 448q^3 + \cdots \in M_4(2) \subseteq M_4(4)$$

and

$$E_4(z) = 1 + 240 \sum_{n=1}^{\infty} \sigma_3(n) q^n = 1 + 240 \left(q + 9q^2 + 28q^3 + \cdots\right) \in M_4(1) \subseteq M_4(4).$$

Assuming these three Eisenstein series give a basis for $M_4(4)$ (they do, as we will see in the next chapter), so a little linear algebra on the first three coefficients of the Fourier expansions shows

$$\vartheta^8(z) = \frac{1}{16} E_4(z) - \frac{1}{16} E_{4,2}(z) + E_{4,4}(z).$$

Comparing Fourier expansions shows

$$r_8(n) = 16\left(\sigma_3(n) - 2\sigma_3(n/2) + 16\sigma_3(n/4)\right).$$

Besides completing a proof of the above result in the next chapter, we will study $r_{2k}(n)$ for other values of k later in the course.

Another classical example of a theta series comes from asking about representing numbers as sums of triangular numbers. Recall the n-th **triangular number** is $\frac{n(n+1)}{2}$, which is the number of entries in the first n rows of Pascal's triangle. Let $\delta_k(n)$ be the number of ways n is a sum of k triangular numbers, i.e.,

$$\delta_k(n) = \#\left\{(x_1, \cdots, x_k) \in \mathbb{Z}_{\geq 0}^k : \frac{x_1(x_1+1)}{2} + \frac{x_2(x_2+1)}{2} + \cdots + \frac{x_k(x_k+1)}{2} = n.\right\}$$

Let

$$\psi(z) = q^{1/8}\sum_{n=0}^{\infty} q^{\frac{n(n+1)}{2}}.$$

Then

$$\psi^k(z) = q^{k/8}\sum_{n=0}^{\infty}\delta_k(n)q^n.$$

Proposition 4.4.8. *Suppose $k \equiv 0 \bmod 8$. Then $\psi^k(z) \in M_{k/2}(4)$.*

In the interest of time, we will not prove this now, but may give a proof at a later time. However, the point for now is it allows one to do the following exercise.

***Exercise 4.4.9.** *(i) Compute $\delta_8(n)$ for $n = 0, 1, 2, 3$.*

(ii) Assuming E_4, $E_{4,2}$ and $E_{4,4}$ is a basis for $M_4(4)$, determine $a, b, c \in \mathbb{Q}$ such that

$$\psi^8(z) = aE_4 + bE_{4,2} + cE_{4,4}.$$

(iii) Using (ii), show $\delta_8(n) = \sigma_3^\sharp(n+1)$ where

$$\sigma_m^\sharp(n) = \sum_{\substack{d|n \\ n/d \text{ odd}}} d^m.$$

We remark that for a general positive definite quadratic form Q in r variables over $\mathbb{Z}$, one can associate the theta series

$$\Theta_Q(z) = \sum_{n=0}^{\infty} r_Q(n)q^n,$$

where $r_Q(n)$ denotes the number of solutions to $Q(x_1, \ldots, x_r) = n$ in $\mathbb{Z}^r$. One can write

$$Q(x_1, \ldots, x_r) = \frac{1}{2}\begin{pmatrix} x_1 & x_2 & \cdots & x_r \end{pmatrix} A \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_r \end{pmatrix},$$

where $A \in M_{r\times r}(\mathbb{Z})$ is symmetric with $\det A > 0$, and the diagonal entries of A are even. Let $N \in \mathbb{N}$ be minimal such that $NA^{-1} \in M_{r\times r}(\mathbb{Z})$ with even diagonal entries. Then one can show $\Theta_Q(z) \in M_{k/2}(\Gamma_1(N))$.

4.5 η and Δ

About half a century after Jacobi's began the theory of theta functions, Dedekind introduced a similar kind of function, which has played an important role in the theory of modular forms.

Definition 4.5.1. *The* **Dedekind eta function** $\eta : \mathfrak{H} \to \mathbb{C}$ *is given by*

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

The factor $q^{1/24}$ may seem curious at first, as with the factor $q^{1/8}$ in the definition of ψ in the previous section, but it is needed to get the desired transformation laws. Namely, we have

Proposition 4.5.2. *The function η is a holomorphic function on $\mathfrak{H}$ satisfying the transformation properties*

$$\eta(z+1) = \eta(z), \quad \eta\left(\frac{-1}{z}\right) = \sqrt{\frac{z}{i}}\eta(z).$$

Proof. To see that it is holomorphic, it suffices to show the product converges absolutely and uniformly on compacts sets in $\mathfrak{H}$, which is equivalent to doing the same for

$$\log \eta(z) = \log q^{1/24} + \sum_{n=1}^{\infty} \log(1 - q^n) = \frac{\pi i z}{12} + \sum_{n=1}^{\infty} \log(1 - q^n).$$

Recall $z \in \mathfrak{H}$ corresponds to $q = e^{2\pi i x} e^{-2\pi y}$ in the region $0 < |q| < 1$. Using the Taylor expansion for $\log(1-x) = -\sum_{j=1}^{\infty} \frac{x^j}{j}$, we see if $x = |q|$ then

$$\left| \sum_{n=1}^{\infty} \log(1 - q^n) \right| \leq \sum_{n=1}^{\infty} \sum_{j=1}^{\infty} x^{nj} j = \sum_{n=1}^{\infty} \sigma_{-1}(n) x^n$$

where $\sigma_{-1}(n)$ is the sum of the reciprocals of divisors of n. In particular, this is less than $\sum_{n=1}^{\infty} n x^n$, which we know converges (say by the ratio test) absolutely with a uniform bound for x lying in a compact subset of $(0, 1)$. This establishes holomorphy.

The transformation $\eta(z+1) = \eta(z)$ is evident since q is invariant under $z \mapsto z + 1$.

There are many ways to obtain the second identity, and we follow [Els06], which gives a proof suggested by Petersson. Since both sides are analytic, it suffices to prove logarithmic derivative of both sides of the desired identity are equal, i.e.,

$$\frac{1}{z^2} \frac{\eta'(-1/z)}{\eta(-1/z)} = \frac{d}{dz} \log \eta\left(\frac{-1}{z}\right) = \frac{d}{dz} \log \sqrt{\frac{z}{i}}\eta(z) = \frac{\sqrt{\frac{z}{i}}\eta'(z) - \frac{i}{2}\sqrt{\frac{i}{z}}\eta(z)}{\sqrt{\frac{z}{i}}\eta(z)} = \frac{\eta'(z)}{\eta(z)} + \frac{1}{2z}.$$

Let $\tau \in \mathfrak{H}$ and consider the lattice $\Lambda = \langle 1, \tau \rangle$. One defines the associated Weierstrass zeta function

$$\zeta(s; \Lambda) = \frac{1}{s} + \sum_{\substack{(m,n) \in \mathbb{Z}^2 \\ (m,n) \neq (0,0)}} \left(\frac{1}{s + m\tau + n} - \frac{1}{m\tau + n} + \frac{s}{(m\tau + n)^2} \right)$$

Using Euler's expression (4.2.4) for cotangent, and a similar one for $\csc^2$, we have

$$\zeta(s;\Lambda) = \frac{\pi^2}{3}s + \pi\cot\pi s + \sum_{m\neq 0}\left(\pi\cot\pi(m\tau+s) - \pi\cot\pi m\tau + \pi^2 s\csc^2\pi m\tau\right).$$

Consequently, we have

$$\phi_1(\tau) := \zeta\left(\frac{1}{2};\Lambda\right) = \left(\frac{\pi^2}{6} + \pi^2\sum_{m=1}^{\infty}\csc^2\pi m\tau\right)$$

and

$$\phi_2(\tau) := \zeta\left(\frac{\tau}{2};\Lambda\right) = \tau\phi_1(\tau) + \frac{1}{2}.$$

On the other hand, the definition of $\zeta(z;\Lambda)$ implies

$$\phi_2(\tau) = \frac{1}{\tau}\phi_1\left(\frac{-1}{\tau}\right).$$

Comparing the previous two equations yields

$$\frac{1}{\tau}\phi_1\left(\frac{-1}{\tau}\right) = \tau\phi_1(\tau) + \frac{1}{2}.$$

Note our expression for ϕ_1 combined with the Fourier expansion $\csc^2\pi z = -4\sum_{n=1}^{\infty} nq^n$ gives

$$\phi_1(z) = -2\pi i\frac{\eta'(z)}{\eta(z)}$$

(cf. Exercise below). But now the previous transformation property for ϕ_1 gives us precisely the desired identity for $\frac{\eta'(z)}{\eta(z)}$. □

Exercise 4.5.3. *(a) Determine the q-expansion for* $\frac{\eta'(z)}{\eta(z)}$.

(b) Check $\phi_1(z) = -2\pi i\frac{\eta'(z)}{\eta(z)}$ *as asserted in the proof above.*

We remark that had we proved the functional equation for the (pseudo-)Eisenstein series E_2 stated in Exercise 4.2.14, one can derive the transformation law for η in a more straightforward manner from that. See, e.g., [Kob93] or [DS05].

Just like ϑ and ψ, one can relate η to quadratic forms. Specifically, one can ask about representing an integer n as a sum of k pentagonal numbers. Here the n-th **pentagonal number** is $\frac{3n^2-n}{2}$ for $n \geq 0$, and this has a geometric interpretation just like square and triangular numbers do (draw a few). Then Euler's pentagonal number theorem (Euler had previously considered the formal product $\prod_{n=1}^{\infty}(1-x^n)$, not actually η as a function of $\mathfrak{H}$) states

$$\eta(z) = q^{1/24}\sum_{n=-\infty}^{\infty}(-1)^n q^{\frac{3n^2-n}{2}}.$$

Note this is not exactly the analogue of ψ for pentagonal numbers, due to both the factor of $(-1)^n$ and the appearance of the "negative pentagonal numbers" $\frac{3n^2-n}{2}$ for $n < 0$.

However, what is more interesting, and Euler's motivation for the pentagonal number theorem, is that

$$\frac{1}{\eta(z)} = q^{-1/24}\sum_{n=0}^{\infty} p(n)q^n, \tag{4.5.1}$$

where $p(n)$ denotes the partition function. Recall a partition of n is a way of writing n as a sum of positive integers (order does not matter), e.g., the partitions of 3 are $3 = 2+1 = 1+1+1$. Then $p(n)$ is the number of partitions of n, so, e.g., $p(3) = 3$ (we count 3 itself as the trivial partition of 3). Here is a table of some partition numbers

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	$\cdots$	100
$p(n)$	1	2	3	5	7	11	15	22	30	42	56	77	101	135	176	231	$\cdots$	190569292

The relation (4.5.1) follows immediately from the product expansion (our definition) for η, and the following exercise.

Exercise 4.5.4. *Show the formal identity*

$$\prod_{n=1}^{\infty}\frac{1}{1-x^n} = \sum_{n=0}^{\infty} p(n)x^n,$$

where $p(0) = 1$.

A classical problem in number theory and combinatorics was to find a simple expression for $p(n)$. (Partition numbers come up often in combinatorial problems.) There is no (known) simple (finite rational) closed form expression for $p(n)$. As you can see from the table, the numbers $p(n)$ start off growing quite slowly, but then appear to have exponential growth (an asymptotic was given by Hardy and Ramanujan). However, by Euler's relation (4.5.1), one can use η and the theory of modular forms to get many results about the partition function. In fact, just in January 2011, Brunier and Ono gave a *finite algebraic* expression for $p(n)$ in terms of η, E_2 and certain binary quadratic forms of discriminant $1-24n$.

So what precisely is the relation with modular forms? Just like with ϑ and ψ, you might expect certain powers of η to be modular forms. From the $q^{1/24}$ factor appearing, it seems clear that such an exponent would need to be a multiple of 24. (This is the same reason we needed to consider ψ^k where $k \equiv 0 \bmod 8$.)

Proposition 4.5.5. *Let* $k \equiv 0 \bmod 24$. *Then* $\eta^k \in M_{k/2}(1)$.

In fact, since these forms are of level 1, the proof is easier than that for ϑ.

***Exercise 4.5.6.** *Prove Proposition 4.5.5.*

There are a couple of things to remark here on apparent differences with powers of ϑ and ψ.

First, we are getting modular forms on the full modular group, as opposed to $\Gamma_0(4)$ in the case of powers of ϑ and ψ. What this means is that η has slightly more refined symmetry than ϑ and ψ. In fact one can write ϑ and ψ in terms of η:

$$\vartheta(z) = \frac{\eta(2z)^5}{\eta(z)^2\eta(4z)^2}, \quad \psi(z) = \frac{\eta^2(2z)}{\eta(z)}$$

One reason why η is quite special, is that all modular forms of "small level" can be generated by "eta quotients." This is useful in giving product expansions for modular forms.

Second, one can try to write η^{24} a linear combination of Eisenstein series on $M_{12}(1)$. However there is only one Eisenstein series here, namely E_{12} and just looking at the constant term (i.e., the value when $z = i\infty$ or $q = 0$) shows η^{24} is not a constant multiple of E_{12}. (When we worked on $M_4(4)$ we could use Eisenstein series from lower levels in addition to $E_{4,4}$, but there is no lower level to work with now.) In other words, the Fourier coefficients for η^{24} do not satisfy such a simple expression in terms of divisor functions as those for ϑ^8 or ψ^8. However, this issue not so much from any inherent difference between η versus ϑ and ψ, but rather from the difference in the exponents of the functions, or put another way, the weights of the modular forms.

Except in the case of low weights k, the space $M_k(\Gamma)$ will not be generated (as a vector space) by only Eisenstein series, but also require *cusp forms*. The Fourier coefficients of these cusp forms, are on one hand more mysterious than the Fourier coefficients of Eisenstein series (not easily expressible in terms of elementary functions like $\sigma_k(n)$), but on the other hand are much better behaved (asymptotically they satisfy better bounds and are in a sense "evenly distributed".)

Definition 4.5.7. *Let $f \in M_k(\Gamma)$. We say f is a* **cusp form** *(of weight k for Γ) if f* vanishes at the cusps*, i.e., if, for each $\tau \in \mathrm{PSL}_2(\mathbb{Z})$, $f|_{\tau,k}(z) \to 0$ as $\mathrm{Im}(z) \to \infty$. The space of cusp forms of weight k for Γ is denoted by $S_k(\Gamma)$, or simply $S_k(N)$ if $\Gamma = \Gamma_0(N)$.*

Note the S in $S_k(\Gamma)$ stands for Spitzenform, which is German for cusp form. The French do not like this notation. (Cusp form is *forme parabolique* in French, but no one uses $P_k(\Gamma)$.)

Lemma 4.5.8. *Let $f \in M_k(\Gamma)$. Then f vanishes at the cusps if and only if $f(z) \to 0$ as $z \to z_0$ for any $z_0 \in \mathbb{Q} \cup \{i\infty\}$.*

Contrast this with the case of Eisenstein series, where holomorphic at the cusps does not mean that $E_{k,N}(z)$ is bounded as z tends to a rational number.

Proof. First suppose f vanishes at the cusps. Let $z_0 \in \mathbb{Q} \cup \{i\infty\}$, and $\tau = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PSL}_2(\mathbb{Z})$ such that $\tau i\infty = z_0$. Since $f|_{\tau,k}(z+N) = f|_{\tau,k}(z)$ for some N, we have a Fourier expansion

$$f|_{\tau,k}(z) = \sum_{n \geq 0} a_{\tau,n} q_N^n, \quad q_N = e^{2\pi i z/N}.$$

The fact that this tends to 0 as $z \to i\infty$ implies $a_{\tau,0} = 0$. Consequently $f|_{\tau,k}(z) = q_N h(z)$ where $h(z) = \sum_{n \geq 0} a_{\tau,n+1} q_N^n$ is a holomorphic function. On the other hand

$$f(\tau z) = (cz+d)^k f|_{\tau,k}(z) = (cz+d)^k q_N h(z) \to 0$$

as $z \to i\infty$ because $q_N = e^{2\pi i x/N} e^{-2\pi y/N} \to 0$ faster than $(cz+d)^k \to \infty$. This implies $f(z) \to 0$ as $z \to \tau i\infty = z_0$.

Conversely, suppose $f(z) \to 0$ as $z \to z_0$ for all $z_0 \in \mathbb{Q} \cup \{i\infty\}$. Take $\tau \in \mathrm{PSL}_2(\mathbb{Z})$. Then

$$f|_{\tau,k}(z) = (cz+d)^{-k} f(\tau z) \to 0 \text{ as } z \to i\infty$$

since both $(cz+d)^{-k} \to 0$ and $f(\tau z) \to 0$ as $z \to i\infty$. □

As with holomorphy at the cusps, to check the condition of vanishing at the cusps given above, it suffices to check the vanishing condition at each cusp, rather than each element of $\mathbb{Q} \cup \{i\infty\}$.

Lemma 4.5.9. *Let $f \in M_k(\Gamma)$. Let $\{z_1, \dots z_r\} \in \mathbb{Q} \cup \{i\infty\}$ be a set of representatives for the cusps for Γ. If $f(z) \to 0$ as $z \to z_i$ for $1 \le i \le r$, then f vanishes at the cusps.*

Proof. Let $z_0 \in \mathbb{Q} \cup \{i\infty\}$ and write $z_i = \gamma z_0$ for some $1 \le i \le r$ and $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$. Since $f(z) \to 0$ as $z \to z_i$, we have $f(\gamma z) \to 0$ as $z \to z_0$. Therefore, we also have

$$f(z) = f|_{\gamma,k}(z) = (cz + d)^{-k} f(\gamma z) \to 0$$

as $z \to z_0$ since here $(cz + d)^{-k} \to 0$. □

Definition 4.5.10. *We define the* **discriminant modular form** $\Delta := \eta^{24} \in M_{12}(1)$. *The* **Ramanujan τ-function** *defined by the Fourier expansion of Δ so that*

$$\Delta(z) = \sum_{n=1}^{\infty} \tau(n) q^n.$$

It is clear from the definition of η that η^{24} has no constant ($n = 0$) term in the Fourier expansion, so Δ is holomorphic at $i\infty$, which is the only cusp for $\mathrm{PSL}_2(\mathbb{Z})$. Thus Δ is a cusp form, i.e., $\Delta \in S_{12}(1)$.

The reason for the terminology of Δ is that is gives the discriminant for elliptic curves. Namely if E is the elliptic curve corresponding to the lattice $\langle 1, z \rangle$, and $\Delta(E)$ is the appropriately normalized discriminant, then $\Delta(E) = \Delta(z)$.

In 1916, Ramanujan computed the first 30 values of $\tau(n)$ and noticed various arithmetical curiosities. Here are the first several values

n	1	2	3	4	5	6	7	8	9	10	11	12
$\tau(n)$	1	-24	252	-1472	4830	-6048	-16744	84480	-113643	-115920	534612	-370944

The most important property Ramaujan numerically observed is that τ is multiplicative, i.e., $\tau(mn) = \tau(m)\tau(n)$ whenever $\gcd(m, n) = 1$. This alone means that τ must have some interesting arithmetical content. He also observed (again numerically) that $\tau(p^2) = \tau(p)^2 - p^{11}$. Both these statements were proved by Mordell in 1917, and this was generalized by Hecke. We will prove these results later using the theory of Hecke operators.

Ramanujan also conjectured that $|\tau(p)| \le 2p^5\sqrt{p}$. Compare this with the p-th Fourier coefficient for E_{12} which is, up to a constant, $\sigma_{11}(p) = 1 + p^{11}$. In other words, the Fourier coefficients for the cusp form Δ of weight 12 should grow at most like the square root of the growth of the Eisenstein series of weight 12. This was proved by Deligne in 1974 as a consequence of his Fields-medal–winning proof of the Weil conjectures. While we obviously can't prove this in our course, there is a weaker bound due to Hecke which is easy to prove that we will treat later. Hecke's bound states $|\tau(p)| = O(p^6)$, i.e., $|\tau(p)| \le Cp^6$ for some constant C. In fact, the proof so simple this bound is often called the trivial bound, and any improvement in the exponent was considered substantial progress.

Ramanujan also noticed some remarkable congruences, such as the following.

Example 4.5.11. *We will see in the next chapter that* $\dim M_{12}(1) = 2$. *Observe that*

$$E_{12}(z) = 1 + \frac{65520}{691}\sum_{n=1}^{\infty}\sigma_{11}(n)q^n = 1 + \frac{65520}{691}q + \cdots$$

and

$$E_4(z)^3 = \left(1 + 240\sum_{n=1}^{\infty}\sigma_3(n)q^n\right)^3 = 1 + 720q + \cdots.$$

The fact that $\dim M_{12}(1) = 2$ *implies any cusp form (any form with zero constant term) must be a multiple of* Δ, *so comparing coefficients of* q, *we see*

$$691E_{12}(z) - 691E_4(z)^3 = -432000q + \cdots = -432000\Delta.$$

Since the coefficients of $E_4(z)^3$ *are integers, taking this mod 691 gives*

$$566\sum_{n=1}^{\infty}\sigma_{11}(n)q^n \equiv 691E_{12}(z) \equiv 566\Delta \equiv \sum_{n=1}^{\infty}\tau(n)q^n \bmod 691,$$

i.e.,

$$\tau(n) \equiv \sigma_{11}(n) \bmod 691.$$

Note, while we have expressed Δ as an algebraic (polynomial) combination of Eisenstein series, our comments before the definition of cusp form were stating that not all modular forms will be *linear* combinations of Eisenstein series. Roughly, it is the cusp forms which will not be. In fact, once we have the notion of an inner product on $M_k(\Gamma)$, we will see that the space of cusp forms is orthogonal to the space generated (linearly) by Eisenstein series.

While we could, from the previous example, write $\tau(n)$ as a polynomial expression in σ_3 and σ_{11} akin to Example 4.3.13 (in fact, we will see later that one can write $\tau(n)$—and more generally the Fourier coefficients of any modular form of full level—as a polynomial expression in σ_3 and σ_5), this is qualitatively different than being able to write the $r_8(n)$ and $\delta_8(n)$ as linear expressions in σ_3 as in Example 4.4.7 and Exercise 4.4.9. This is reflected in the very different behavior of Fourier coefficients for forms which are linear combinations of Eisenstein series and cusp forms.

Chapter 5

Dimensions of spaces of modular forms

As indicated in the introduction and the last chapter, we would like to know that a given space of modular forms $M_k(\Gamma)$ is finite dimensional. Then, for example, we may be able to find a basis for the space in terms of Eisenstein series, and use this to study questions about quadratic forms, as Example 4.4.7 and Exercise 4.4.9. In order to do this, we will need to know the dimension of the relevant space $M_k(\Gamma)$.

In this chapter, we will first handle the case of full level, proving finite dimensionality and giving a simple formula for the dimension of $M_k(1)$. Then, as in [Kil08], we will prove Sturm's bound, which will allow us to deduce finite dimensionality for $M_k(\Gamma)$. One can prove dimension formulas for $M_k(N)$ (and more generally $M_k(\Gamma)$), but in the interest of time we will not prove them in general, but merely state them for reference. However, it will be a consequence of Sturm's bound that one can, for instance, rigorously verify the results of Example 4.4.7 and Exercise 4.4.9 without actually checking the dimension of $M_4(4)$.

5.1 Dimensions for full level

To study the dimension of $M_k(1)$, we will need the residue theorem from complex analysis, which we now recall. If f is meromorphic with a pole at s, then the **residue** of f at s is $\text{Res}_{z=s} f(z) = a_{-1}$ where $f(z) = \sum a_n (z-s)^n$, i.e., the residue is the coefficient of the $\frac{1}{z-s}$ term in the Laurent expansion of f at s.

Theorem 5.1.1. (Cauchy's Residue Theorem) *Let U be an open set in $\mathbb{C}$ whose boundary is a simple closed curve C. Let $f : U \to \hat{\mathbb{C}}$ be meromorphic and let $\{s_j\}$ denote the set of singularities of f. If f extends to a holomorphic function at each $z \in C$, then*

$$\int_C f(z)dz = 2\pi i \sum_j \text{Res}_{z=s_j} f(z)$$

By convention, the integral around a closed curve will be in the counterclockwise direction. The condition that f must be holomorphic on C essentially means that there should be no poles along our path of integration.

Example 5.1.2. *Suppose f is holomorphic on a region U. In this case there are no residues because there are no poles, so $\int_C f(z)dz = 0$ for any simple closed curve $C \subset U$.*

Example 5.1.3. *Consider $f(z) = \frac{1}{z^k}$, and let C be any simple closed curve containing the origin. Here there is just one pole, at $z = 0$, and $\mathrm{Res}_{z=0} f(z) = 1$ if $k = 1$ and 0 otherwise. Thus*

$$\int_C \frac{1}{z^k} dz = \begin{cases} 2\pi i & k = 1 \\ 0 & \text{else.} \end{cases}$$

Meromorphic functions are controlled by their zeroes and poles. For a meromorphic function f on U, we define for any point $p \in U$, the **order** of f of f at p to be $v_p(f) = m$ where the Laurent (or Taylor) expansion of f at p is $f(z) = \sum_{n=m}^{\infty} a_n(z-p)^n$ with $a_m \neq 0$, i.e., the order tells you where the Laurent/Taylor series expansion at p starts. In other words, if $v_p(f) = m > 0$ means f has zero of order m at p, $v_p(f) = 0$ means $f(p) \neq 0$, and $v_p(f) = m < 0$ means f has a pole of order $-m$ at p.

Aside: in algebraic geometry, one associates to f a *divisor* $\mathrm{div}(f)$ defined to be a *formal* sum

$$\mathrm{div}(f) = \sum_{p \in U} v_p(f)p.$$

This tells you where the zeroes and the poles are and what their order is, and if f is a rational function, then this sum is finite. For example, if we consider the meromorphic function

$$f(z) = 12086 \frac{(z-1)^3(z-4)}{z^2(z+2i)}$$

on $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, then

$$\mathrm{div}(f) = 3 \cdot 1 + 1 \cdot 4 - 2 \cdot 0 - 1 \cdot (-2i) - 1 \cdot \infty.$$

In other words f has zeroes of order 3 and 1 at $z = 1$ and $z = 4$, poles of order 2, 1 and 1 at $z = 0$, $2i$ and ∞. I.e., $v_1(f) = 3$, $v_4(f) = 1$, $v_0(f) = -2$, $v_{-2i}(f) = -1$ and $v_\infty(f) = -1$ (and $v_p(f) = 0$ for any other $p \in \hat{\mathbb{C}}$. Note replacing f with a nonzero constant multiple does not change the divisor, and it is not hard to see that $\mathrm{div}(f)$ determines f up to a constant.

Observe that for any rational function f,

$$\sum_{p \in \hat{\mathbb{C}}} v_p(f) = 0. \tag{5.1.1}$$

(If $f = \frac{q}{r}$ where q and r are polynomials, $\mathrm{div}(f) = \mathrm{div}(q) - \mathrm{div}(r)$. Now the number of zeroes of q is simply the degree, and q has no poles in $\mathbb{C}$, but a pole of order $\deg(q)$ at ∞. Therefore the sum $\sum v_p(q)$ of coefficients of $\mathrm{div}(q)$ is 0, and the same is true for r, and therefore f.) Our proof of dimension formulas for $M_k(1)$ will rely on an analogue of this formula for modular forms.

We remark that (5.1.1) is analogous to the formula

$$\prod_v |x|_v = 1$$

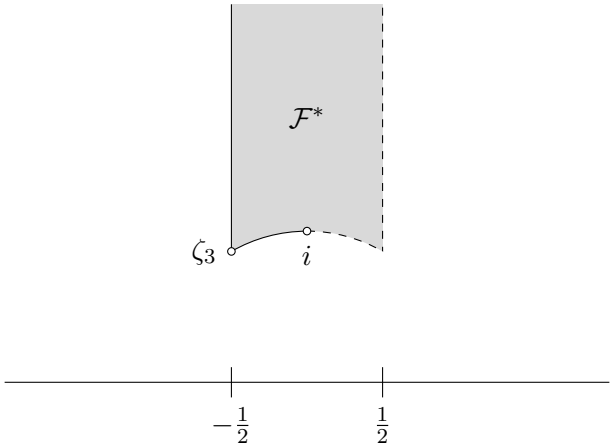
from algebraic number theory, where $x \in \mathbb{Q}$ and v runs over the places of $\mathbb{Q}$. (And both of these formulas are trivial consequences of the definition of the order of a function at a point and the p-adic valuation of a rational number. The analogy between algebraic number theory and algebraic geometry runs much deeper than this of course.)

Theorem 5.1.4. *Let f be a nonzero meromorphic modular form of weight k for $\mathrm{PSL}_2(\mathbb{Z})$. For $p \in \overline{\mathfrak{H}}$, let C_p be the elliptic subgroup of $\mathrm{PSL}_2(\mathbb{Z})$ stabilizing p (cf. Section 3.5), and let*

$$\mathcal{F}^* = \left\{ z \in \mathfrak{H} : -\frac{1}{2} \leq \mathrm{Re}(z) < \frac{1}{2},\ |z| \geq 1,\ \textit{and } |z| > 1 \textit{ if } \mathrm{Re}(z) \geq 0 \right\} - \{\zeta_3\},$$

so $\mathcal{F}^* \cup \{i, \zeta_3\}$ is in bijection with $\mathrm{PSL}_2(\mathbb{Z})\backslash\mathfrak{H}$. Then

$$\sum_{p \in \mathrm{PSL}_2(\mathbb{Z})\backslash\overline{\mathfrak{H}}} \frac{1}{|C_p|} v_p(f) = v_{i\infty}(f) + \frac{1}{2} v_i(f) + \frac{1}{3} v_{\zeta_3}(f) + \sum_{p \in \mathcal{F}^*} v_p(f) = \frac{k}{12}. \tag{5.1.2}$$



We wrote the expression $\sum \frac{1}{|C_p|} v_p(f)$ on the left of (5.1.2) both to give a more uniform presentation of the sum and make more clear the analogy with (5.1.1). However it is clear this expression equals $v_{i\infty}(f) + \frac{1}{2} v_i(f) + \frac{1}{3} v_{\zeta_3}(f) + \sum_{p \in \mathcal{F}^*} v_p(f)$ from Section 3.5, and this latter expression is what we will use in the proof.

Proof. Note that we can relate the orders of zeroes and poles of f to residues of the logarithmic derivative $\frac{f'}{f}$. Namely if $f(z) = \sum_{n=m} a_n (z-p)^n$ with $a_m \neq 0$, then

$$\frac{f'(z)}{f(z)} = \frac{m a_m (z-p)^{m-1} + (z-p)^m \left((m+1)a_m + \cdots\right)}{a_m (z-p)^m (1 + \cdots)} = \frac{m}{z-p} + c_0 + z(c_1 + c_2 z^2 + \cdots),$$

so

$$\mathrm{Res}_{z=p} \frac{f'}{f} = v_p(f).$$

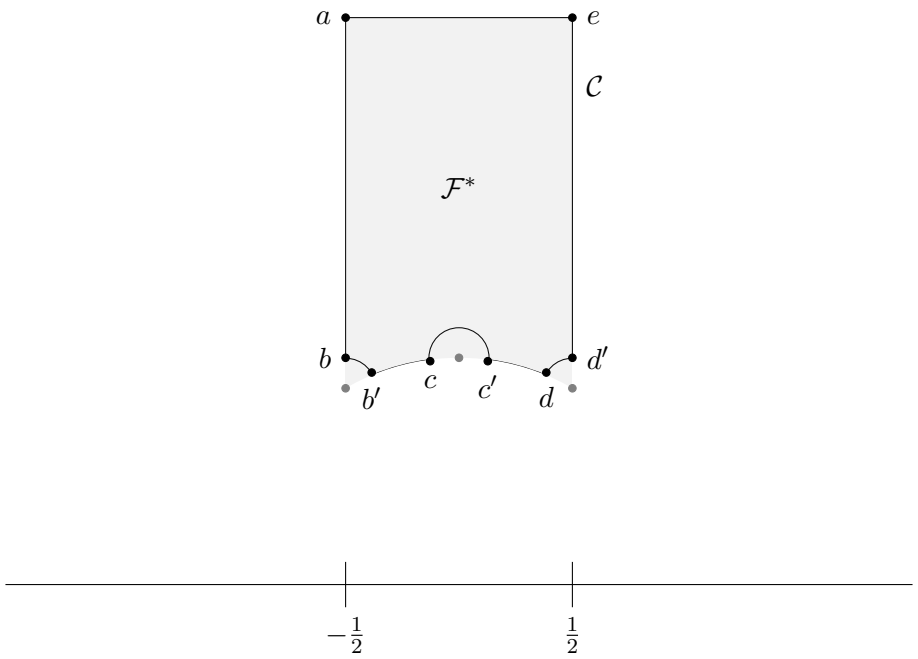
Consequently, if f is holomorphic on U and $\mathcal{C} \subset U$ is a simple closed curve which misses all zeroes and poles of f (i.e., misses all poles of $\frac{f'}{f}$), then the residue theorem says

$$\int_{\mathcal{C}} \frac{f'}{f} dz = 2\pi i \sum v_p(f), \tag{5.1.3}$$

where the sum is over all points in the region enclosed by $\mathcal{C}$.

First we assume that f has no zeroes or poles on the boundary $\partial\mathcal{F}^*$ of $\mathcal{F}^*$. Note $\partial\mathcal{F}^* = \partial\mathcal{F}$ where $\mathcal{F}$ is the standard fundamental domain for $\mathrm{PSL}_2(\mathbb{Z})$); here, we mean the boundary inside $\mathfrak{H}$, so a zero or pole at $i\infty$ is not ruled out. In this case we will consider our simple closed curve $\mathcal{C}$ to be of the following type.

The curve $\mathcal{C}$ is mostly easily described in the diagram below, but here is how I would describe it in words also. The curve $\mathcal{C}$, oriented counterclockwise, will begin at some point a high up on the left boundary $\mathrm{Re}(z) = -\frac{1}{2}$ of $\mathbb{F}^*$, travel down the vertical line $\mathrm{Re}(z) = -\frac{1}{2}$, stopping at a point b just short of $\zeta(3)$, then arcing inside $\mathcal{F}^*$ to travel along the bottom boundary $|z| = 1$, making a small arc inside $\mathcal{F}^*$ to avoid i, then continuing along the bottom boundary, again making a small arc to avoid ζ_6, then climbing up the right boundary $\mathrm{Re}(z) = \frac{1}{2}$ to a point e such that $\mathrm{Im}(e) = \mathrm{Im}(a)$, then traveling along the horizontal line $\mathrm{Im}(z) = \mathrm{Im}(a)$ to end back at a.



Note that this curve $\mathcal{C}$ is made in such a way that no zeroes of poles of f inside $\mathcal{F} - \{i, \zeta_3, \zeta_6\}$ lie outside of $\mathcal{C}$. This is possible since there are only finitely many zeroes and poles inside $\mathcal{F}$. To see this, recall the open balls $B_r(i\infty) = \{i\infty\} \cup \{z \in \mathfrak{H} : \mathrm{Im}(z) > r\}$ form a

basis of neighborhoods of $i\infty$. So f being meromorphic at $i\infty$ means we can choose a ball $B_r(i\infty)$ such that f has no zeroes or poles in $B_r(i\infty)$ except possibly at $i\infty$. Consequently all zeroes and poles of f inside $\mathcal{F}$ must lie inside a truncated fundamental domain $\mathcal{F}_r = \{z \in \mathcal{F} : \mathrm{Im}(z) \leq r\}$. But the number of zeroes and poles inside $\mathcal{F}_r$ must be finite because $\mathcal{F}_r$ is compact.

We also assume $\mathcal{C}$ is symmetric under reflection about $i\mathbb{R}^+$.

By (5.1.3), we know

$$\int_{\mathcal{C}} \frac{f'}{f} dz = 2\pi i \sum_{p \in \mathcal{F}^*} v_p(f), \tag{5.1.4}$$

On the other hand, we can compute the integral along $\mathcal{C}$ piecewise. Since $f(z+1) = f(z)$ we have

$$\int_a^b \frac{f'}{f} dz + \int_{d'}^e \frac{f'}{f} dz = 0.$$

Thus

$$\int_{\mathcal{C}} \frac{f'}{f} dz = \int_b^{b'} \frac{f'}{f} dz + \int_{b'}^c \frac{f'}{f} dz + \int_c^{c'} \frac{f'}{f} dz + \int_{c'}^d \frac{f'}{f} dz + \int_d^{d'} \frac{f'}{f} dz + \int_e^a \frac{f'}{f} dz.$$

First note the change of variables $z \mapsto q$ transforms the integral from e to a to a circle around $q = 0$ with negative orientation. Then Cauchy's Residue Theorem gives

$$\int_e^a \frac{f'}{f} dz = -2\pi i v_{i\infty}(f).$$

Next, if we integrate $\frac{f'}{f}$ around the circle α_r of radius r containing the arc from b to b', we get $2\pi i v_{\zeta_3}(f)$. (By making the arc of small enough radius, we may assume f has no zeros or poles inside this circle except possibly at ζ_3.) If we take the radius r going to 0, its angle tends to $\frac{\pi}{3}$, so you might guess

$$\int_b^{b'} \frac{f'}{f} dz \to -\frac{\pi i}{3} v_{\zeta_3}(f).$$

This is true, and it follows from applying the following lemma to $g(z) = \frac{f'(z-\zeta_3)}{f(z-\zeta_3)}$, which has at most a simple pole at 0 whose residue is $v_{\zeta_3}(f)$.

Lemma 5.1.5. *Suppose g is meromorphic with at most a simple pole at the origin. Let S_r be the circle of radius r around the origin, $0 \leq \theta_1 \leq \theta_2 \leq 2\pi$, and consider the points $a_r = re^{i\theta_1}$ and $b_r = re^{i\theta_2}$ on S_r. Then*

$$\lim_{r \to 0} \int_{a_r}^{b_r} g(z) dz = (\theta_2 - \theta_1) i \, \mathrm{Res}_{z=0} \, g(z),$$

where the integral is taken along the arc in S_r from a_r to b_r. In particular, in the limit as $r \to 0$, the integral only depends upon the length of the arc.

Proof. First write $g(z) = a_{-1}z^{-1} + h(z)$ where $h(z)$ is holomorphic at 0 and $a_{-1} = \operatorname{Res}_{z=0} g(z)$. It is clear that $\lim_{r\to 0} \int_{a_r}^{b_r} h(z)dz = 0$, so it suffices to consider

$$\lim_{r\to 0} \int_{a_r}^{b_r} \frac{a_{-1}}{z} dz.$$

Write $z = re^{i\theta}$ so $dz = ire^{i\theta}d\theta$. Then

$$\lim_{r\to 0} \int_{a_r}^{b_r} \frac{a_{-1}}{z} dz = \lim_{r\to 0} \int_{\theta_1}^{\theta_2} \frac{a_{-1}}{re^{i\theta}} ire^{i\theta} d\theta = (\theta_2 - \theta_1) i a_{-1}.$$

□

Similarly, we have

$$\int_c^{c'} \frac{f'}{f} dz \to -\pi i v_i(f)$$

and

$$\int_d^{d'} \frac{f'}{f} dz \to -\frac{\pi i}{3} v_{\zeta_6}(f) = -\frac{\pi i}{3} v_{\zeta_3}(f).$$

Putting all this together gives

$$2\pi i \left(v_{i\infty}(f) + \frac{1}{2} v_i(f) + \frac{1}{3} v_{\zeta_3}(f) + \sum_{p \in \mathcal{F}^*} v_p(f) \right) = \lim_{\mathcal{C} \to \partial\mathcal{F}} \int_{b'}^{c} \frac{f'}{f} dz + \int_{c'}^{d} \frac{f'}{f} dz.$$

By our symmetry assumption on $\mathcal{C}$, $S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ takes the arc from b' to c to the arc from d to c' (reverses orientation). Recall $Sz = -\frac{1}{z}$, so differentiating $f(Sz) = z^k f(z)$ gives

$$f'(Sz)\frac{d}{dz}Sz = f'(Sz)\frac{d}{dz}\frac{-1}{z} = \frac{1}{z^2} f'(Sz) = kz^{k-1} f(z) + z^k f'(z).$$

Therefore,

$$\frac{f'(Sz)}{f(Sz)} = \frac{kz^{k+1} f(z) + z^{k+2} f'(z)}{z^k f(z)} = kz + z^2 \frac{f'(z)}{f(z)}.$$

so if $z = Sw$, then $dz = \frac{1}{w^2} dw$ and

$$\int_{b'}^{c} \frac{f'(z)}{f(z)} dz = \int_d^{c'} \frac{f'(Sw)}{f(Sw)} \frac{dw}{w^2} = \int_d^{c'} \left(\frac{k}{w} + \frac{f'(w)}{f(w)} \right) dw$$

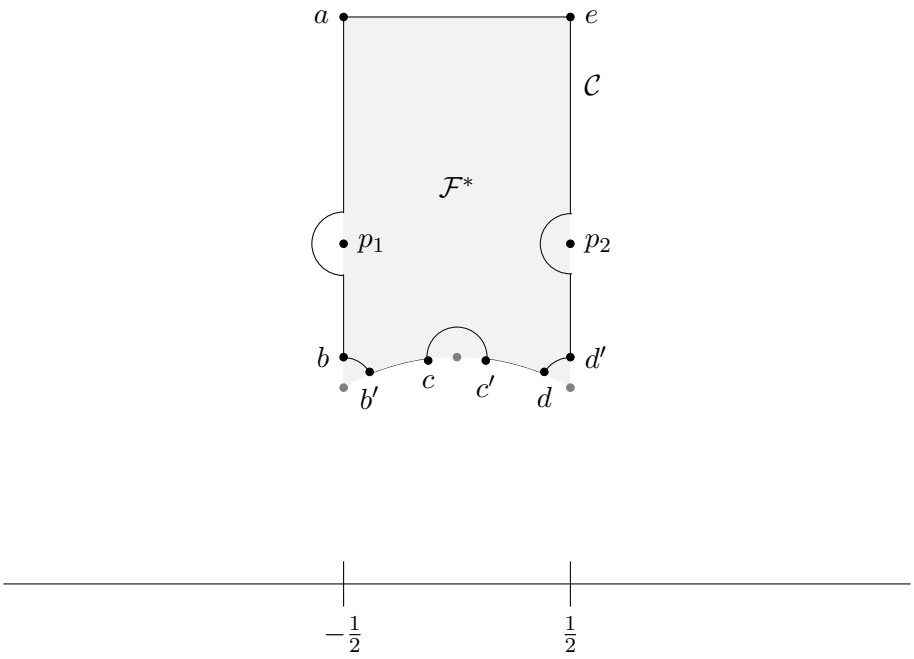
Since the path from d to c' is taken along $|w| = 1$,

$$\int_{b'}^{c} \frac{f'}{f} dz + \int_{c'}^{d} \frac{f'}{f} dz = k \int_d^{c'} \frac{dw}{w} \to k \int_{\pi/3}^{\pi/2} d\theta = k\frac{\pi i}{6},$$

as $\mathcal{C} \to \partial\mathcal{F}$.

This proves the theorem with our assumptions about the zeroes and poles along the boundary of ∂F. Now suppose there are m points $p_1, \ldots, p_m$, along $\partial F - \{i, \zeta_3, \zeta_6\}$ at which

f is zero or infinite. (There are at most finitely many by meromorphy.) Then we can use the above argument, by suitably modifying the type of our curve $\mathcal{C}$ so as to just avoid $p_1, \ldots, p_m$. Note that if p_i is such a point along the left boundary $\mathrm{Re} z = -\frac{1}{2}$ (resp. bottom left boundary $\mathrm{Re} z < 0$, $|z| = 1$), then Tp_i (resp. Sp_i) is also such a point along the right (resp. bottom right) boundary. For example if there are only two such points, say p_1 on the left boundary, and $p_2 = Tp_1$ along the right boundary, we can take our curve of the following type.



Here our previous argument goes through unchanged, since the integrals from a to b and d' to e still cancel each other out. This obviously works for any (finite) number of zeroes and poles along the side boundary. It is also easy to see one can treat zeroes and poles on the lower boundary similarly (cf. exercise below). □

Exercise 5.1.6. *With notation as in the previous theorem, suppose f has a zero or pole at some point p_1 lying on the lower left boundary $\left\{z \in \mathfrak{H} : |z| = 1, -\frac{1}{2} < \mathrm{Re}(z) < 0\right\}$ of $\partial\mathcal{F}$. Assume f has no zeroes or poles on $\partial F - \{i, \zeta_3, \zeta_6\}$ except at p_1 and $p_2 = Sp_1$. Check (5.1.2) still holds.*

Now we can use this to start computing dimensions of spaces of modular forms.

Example 5.1.7. *While we have been assuming $k \geq 0$ even up to now, one can consider the same definitions for $k < 0$ even, and (5.1.2) still holds. If $k < 0$, then (5.1.2) says f must have a pole. Consequently, there are no (nonzero) holomorphic modular forms of negative weight, justifying our "omission" of negative weights. (Meromorphic modular forms*

of negative weights do of course exist—just take $\frac{1}{f}$ for any nonzero holomorphic modular form f of positive weight.)

Proposition 5.1.8. *We have*

$$M_k(1) = \begin{cases} \mathbb{C} & k = 0 \\ \{0\} & k = 2 \\ \mathbb{C}E_k & k = 4, 6, 8, 10, 14. \end{cases}$$

This will complete the proof of the relation between σ_7 and σ_3 in 4.3.13.

Proof. We already know $M_0(1) = \mathbb{C}$ from Corollary 4.1.12.

Next note any nonzero term on the left of (5.1.2) is at least $\frac{1}{3}$, so (5.1.2) has no solutions of $k = 2$.

If $k = 4$, there is precisely one solution to (5.1.2), namely $v_{\zeta_3}(f) = 1$ and all other $v_p(f) = 0$. Similarly, one easily sees (5.1.2) has only one solution for $k = 6, 8, 10$ or 14. However, we already know E_k is one nonzero modular form in $M_k(1)$. Consequently if $f \in M_k(1)$ for $k = 4, 6, 8, 10$ or 14, then (5.1.2) having only one solution means the zeroes of f must agree with the zeroes of E_k. Consequently, f/E_k is a meromorphic modular form of weight 0 with no zeroes or poles, i.e., a constant since $M_0(1) = \mathbb{C}$. □

Note if $k = 12$, then there are infinitely many solutions to (5.1.2), and we need to use another argument to determine $M_{12}(1)$ (and similarly for $M_k(1)$ for $k > 14$). We already know two forms in this space, the Eisenstein series E_{12}, and the cusp form Δ.

Lemma 5.1.9. *Let $k \geq 4$ even. Then $M_k(1) = S_k(1) \oplus \mathbb{C}E_k$.*

Proof. Let $f \in M_k(1)$. We know $E_k \in M_k(1)$ is 1 at $i\infty$. Thus $g(z) = f(z) - f(i\infty)E_k(z) \in M_k(1)$ and vanishes at $i\infty$, i.e., $g \in S_k(1)$. □

In other words, every modular form in $M_k(1)$ is the sum of a cusp form and a multiple of the Eisenstein series E_k. In particular, to determine $\dim M_k(1)$ it suffices to determine $\dim S_k(1) = \dim M_k(1) - 1$. Note that by the previous proposition, there no nonzero cusp forms of level 1 for weights 0 through 10 or 14, so Δ is the first (smallest weight) instance of a cusp form.

Proposition 5.1.10. *The space $S_{12}(1) = \mathbb{C}\Delta$, i.e., $M_{12}(1) = \mathbb{C}\Delta \oplus \mathbb{C}E_{12}$.*

This will complete the proof that $\tau(n) \equiv \sigma_{11}(n) \bmod 691$ in Example 4.5.11.

Proof. Note that with $k = 12$, (5.1.2) only has one solution with $v_{i\infty}(f) > 0$, namely $v_{i\infty}(f) = 1$ and $v_p(f) = 0$ for any $p \neq i\infty$. This applies to any nonzero $f \in S_{12}(1)$, since such an f must vanish at $i\infty$. I.e., any nonzero $f \in S_{12}(1)$ has a simple (order 1) zero at $i\infty$ and no zeroes in $\mathfrak{H}$. In particular this is true for Δ. Then, as above, for any $f \in S_{12}(1)$, we have $f/\Delta \in M_0(1)$, i.e., f is a scalar multiple of Δ. □

For $k = 16, 18, 20, 22, 26$, again one can observe there is only one solution to (5.1.2) with $v_{i\infty}(f) > 0$ to see $\dim S_k(1) = 1$ and $\dim M_k(1) = 2$. However, we would prefer to be able to explicitly construct the cusp forms.

How can we construct cusp forms? One way would be to construct a form in $M_k(1)$ which isn't a multiple of E_k and subtract off an appropriate multiple of E_k, as in the proof of Lemma 5.1.9. More suitable for our present purposes is just to multiply a cusp form (namely Δ) with another form.

Exercise 5.1.11. *For any congruence subgroup Γ, let $f \in M_k(\Gamma)$ and $g \in S_\ell(\Gamma)$. Then $fg \in S_{k+\ell}(\Gamma)$.*

Now we show that any cusp form can be constructed as a multiple of Δ and a modular form of smaller weight.

Lemma 5.1.12. *Let $k \geq 16$. Then $S_k(1) = \Delta \cdot M_{k-12}(1)$. In particular, $\dim M_k(1) = 1 + \dim M_{k-12}(1)$.*

Proof. Let $f \in S_k(1)$. Then $v_{i\infty}(f) \geq 1$. On the other hand, Δ has a simple zero at $i\infty$ and is nonzero on $\mathfrak{H}$, therefore f/Δ is a holomorphic modular form of weight $k - 12$. □

We can put all this information together in the following theorem.

Theorem 5.1.13 (Dimension formula in level 1)**.** *Let $k \geq 0$ even. Then*

$$\dim M_k(1) = \begin{cases} \lfloor \frac{k}{12} \rfloor + 1 & k \not\equiv 2 \bmod 12 \\ \lfloor \frac{k}{12} \rfloor & k \equiv 2 \bmod 12. \end{cases}$$

By Lemma 5.1.9, this also tells us the dimensions of $S_k(1)$. Note the reason for the difference in the case $k \equiv 2 \bmod 12$ is because $E_2 \notin M_2(1)$.

In fact, we know more than the dimensions. What we did above, allows us to a basis for any M_k in terms of Eisenstein series and Δ.

Example 5.1.14. *Consider $M_{40}(1)$. We can write*

$$S_{40}(1) = \Delta \cdot M_{28}(1) = \Delta \cdot (\mathbb{C}E_{28} \oplus S_{28}(1)) = \Delta \cdot (\mathbb{C}E_{28} \oplus \Delta (\mathbb{C}E_{16} \oplus \Delta E_4)) .$$

Consequently, a basis for $M_{40}(1)$ is $\{E_{40}, \Delta E_{28}, \Delta^2 E_{16}, \Delta^3 E_4\}$.

Exercise 5.1.15. *Write a basis for $M_{36}(1)$ in terms of Eisenstein series and Δ.*

Since we can write any modular form of level 1 in terms of Eisenstein series and Δ, we can write any modular form of level 1 simply in terms of Eisenstein series by the relation (now justified)

$$\Delta = \frac{691}{432000} \left(E_4^3 - E_{12}\right)$$

that we obtained in Example 4.5.11. In fact we can write Δ in terms of strictly lower weight Eisenstein series, which is in many places how Δ is defined.

Exercise 5.1.16. *Show that*

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

Note the above relation shows the j-invariant defined in Exercise 4.2.13 is simply

$$j(z) = \frac{E_4(z)^3}{\Delta(z)}.$$

In other words, the j-invariant of an elliptic curve (which is actually an invariant of analytic isomorphism classes) is closely related to the discriminant Δ of the elliptic curve (which is not an invariant for isomorphism classes).

In fact all modular forms for $\mathrm{PSL}_2(\mathbb{Z})$ can be written in terms of E_4 and E_6.

Proposition 5.1.17. *Let $f \in M_k(1)$. Then f is polynomial in E_4 and E_6.*

Proof. Since the above procedure shows any f is a polynomial in $\Delta = \frac{E_4^3 - E_6^2}{1728}$ and $E_4, E_6, \ldots, E_k$, it suffices to show each E_j is a polynomial in E_4 and E_6. We use induction. Suppose it is true for E_j with $j \leq k-2$, and consider E_k.

We may assume $k \geq 8$ even. Then there exist $r, s \in \mathbb{Z}_{\geq 0}$ such that $k = 4r + 6s$. Hence $E = E_4^r E_6^s \in M_k(1)$. Since neither E_4 nor E_6 vanish at infinity, E this is not a cusp form. Further, because $M_k(1) = \mathbb{C}E_k \oplus S_k(1)$, there exists $g \in S_k(1)$ such that $E - g = cE_k$ for some $c \neq 0$. (Note the Fourier expansions of $E - g$ and E_k both start with 1, so in fact $c = 1$). On the other hand, since $g = \Delta g'$ for some $g' \in M_{k-12}(1)$, by induction g is a polynomial in E_4 and E_6. Therefore E_k also is. □

Exercise 5.1.18. *(i) Write E_{10} and E_{14} as polynomials in E_4 and E_6.*

(ii) Using a relation between E_{14}, E_{10} and E_4, show

$$\sigma_{13}(n) \equiv 11\sigma_9(n) - 10\sigma_3(n) \bmod 2640.$$

5.2 Finite dimensionality for congruence subgroups

As you can see from the proof of Theorem 5.1.13, trying to carry out the same approach for an arbitrary congruence subgroup does not seem so appealing. Of course, for a given congruence subgroup, one can fix a fundamental domain and mimic the proof of Theorem 5.1.4 and use this to compute dimension formulas.

However, there is an general approach using some Riemann surface theory and algebraic geometry, namely the Riemann–Roch theorem. We will not cover this here (see, e.g., [DS05]), but simply remark that one can prove an analogue of Theorem 5.1.4, which states if f is a meromorphic modular form of weight k on Γ, then

$$\sum_{p \in \Gamma\backslash\overline{H}} \frac{1}{|C_p|} v_p(f) = \frac{k}{2}\left(\frac{1}{2}\epsilon_2 + \frac{2}{3}\epsilon_3 + \epsilon_\infty + 2g - 2\right), \tag{5.2.1}$$

where C_p is the stabilizer of p in Γ, ϵ_2 (resp. ϵ_3) is the number of elliptic points of order 2 for Γ, ϵ_∞ is the number of cusps for Γ, and g is the genus of $\Gamma\backslash\overline{H}$. Note for $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$, one has $\epsilon_2 = 1$, $\epsilon_3 = 1$, $\epsilon_\infty = 1$ and $g = 0$ so this agrees with Theorem 5.1.4.

Consequently, one can show if $k \geq 0$ even, then

$$\dim M_k(\Gamma) = \begin{cases} (k-1)(g-1) + \lfloor \frac{k}{4} \rfloor \epsilon_2 + \lfloor \frac{k}{3} \rfloor \epsilon_3 + \frac{k}{2}\epsilon_\infty & k \geq 2 \\ 1 & k = 0, \end{cases} \tag{5.2.2}$$

and

$$\dim S_k(\Gamma) = \begin{cases} (k-1)(g-1) + \lfloor \frac{k}{4} \rfloor \epsilon_2 + \lfloor \frac{k}{3} \rfloor \epsilon_3 + \left(\frac{k}{2} - 1\right) \epsilon_\infty & k \geq 4 \\ g & k = 2 \\ 0 & k = 0. \end{cases} \tag{5.2.3}$$

For $\Gamma = \Gamma_0(N)$, explicit formulas for g, ϵ_2, ϵ_3 and ϵ_∞ in terms of N are given in [DS05] (cf. Exercise 3.5.6 and Exercise 3.5.14), making the formulas for $\dim M_k(N)$ and $\dim S_k(N)$ quite explicit. For instance, we remark

$$\dim S_2(p) = \frac{1}{12}(p+1) - \frac{1}{4}\left(1 + \left(\frac{-1}{p}\right)\right) - \frac{1}{3}\left(1 + \left(\frac{-3}{p}\right)\right) \tag{5.2.4}$$

and

$$\dim S_2(p^2) = \begin{cases} \frac{1}{12}(p+1)(p-6) + 1 - \frac{1}{4}\left(1 + \left(\frac{-1}{p}\right)\right) - \frac{1}{3}\left(1 + \left(\frac{-3}{p}\right)\right) & p \geq 5 \\ 0 & p = 2, 3 \text{ (or } 5). \end{cases} \tag{5.2.5}$$

We tabulate some explicit dimensions at the end of this chapter.

Instead, we will take a simpler, and in some sense more practical, approach (as in [Kil08]) using Sturm's bound, which will allow us to conclude finite dimensionality with a bound for dimensions, which in some cases will be sharp.

Theorem 5.2.1. (Sturm's bound) *Let Γ be a congruence subgroup and $f \in M_k(\Gamma)$. Let $\rho_1, \ldots, \rho_t$ be the cusps of Γ. If*

$$\sum_{\rho_j} v_{\rho_j}(f) > \frac{k[\mathrm{PSL}_2(\mathbb{Z}) : \Gamma]}{12},$$

then $f = 0$.

Proof. First suppose $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$ and $f \in M_k(1)$ with $r := v_{i\infty}(f) > \frac{k}{12}$. Since Δ has only a simple zero at $i\infty$ and is nonzero on $\mathfrak{H}$, $\frac{f}{\Delta^r}$ is a holomorphic modular form of weight $k - 12r < 0$. Since there are no nonzero holomorphic modular forms of negative weight (Example 5.1.7), this means $f = 0$.

Now suppose $\Gamma \neq \mathrm{PSL}_2(\mathbb{Z})$ and put $M = [\mathrm{PSL}_2(\mathbb{Z}) : \Gamma]$ and $f \in M_k(\Gamma)$ with $v_{i\infty}(f) > \frac{kM}{12}$. We will use f to construct a modular form for $\mathrm{PSL}_2(\mathbb{Z})$.

Let $\{\tau_i\}_{1 \leq i \leq M}$ be a set of coset representatives for $\Gamma \backslash \mathrm{PSL}_2(\mathbb{Z})$ with $\tau_1 \in \Gamma$. Recall

$$f|_{\tau,k}(z) = j(\tau, z)^{-k} f(\tau z) = (cz + d)^{-k} f(\tau z),$$

where $\tau = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, so $f|_{\tau_1,k} = f$. Let

$$F(z) = \prod_{i=1}^{M} f|_{\tau_i,k}(z) = f(z) \prod_{i=2}^{M} f|_{\tau_i,k}(z).$$

Note

$$f|_{\gamma\tau_i,k} = (f|_{\gamma,k})\,|_{\tau_i,k} = f|_{\tau_i,k}$$

for any $\gamma \in \Gamma$, i.e., $f|_{\tau_i,k}$ does not depend upon the choice of coset representative.

Consider any $\tau \in \mathrm{PSL}_2(\mathbb{Z})$. Note $\{\tau_i' := \tau\tau_i\}$ is just another set of coset representatives for $\Gamma\backslash\mathrm{PSL}_2(\mathbb{Z})$, hence

$$F|_{\tau,kM} = \left(\prod_{i=1}^{M} f|_{\tau_i,k}\right)|_{\tau,k} = \prod_{i=1}^{M} (f|_{\tau_i,k})\,|_{\tau,k} = \prod_{i=1}^{M} f|_{\tau\tau_i,k} = F,$$

i.e., $F \in M_{kM}(1)$. Since $v_{i\infty}(F) = \sum_{i=1}^{M} v_{\tau_i^{-1}\cdot i\infty}(f) > \frac{kM}{12}$, by the $\mathrm{PSL}_2(\mathbb{Z})$ case we know $F = 0$, which implies $f = 0$. □

Remark 5.2.2. *In the course of the proof we have shown how to construct a modular form of full level at the expense of increasing the weight. The same argument evidently shows the following.*

Let $\Gamma' \subseteq \Gamma \subseteq \mathrm{PSL}_2(\mathbb{Z})$ be congruence subgroups, $f \in M_k(\Gamma')$ and $M = [\Gamma : \Gamma']$. Let $\{\tau_i\}_{1\le i\le M}$ be a set of coset representatives for $\Gamma'\backslash\Gamma'$. Then

$$F(z) = \prod_{i=1}^{M} f|_{\tau_i,k} \in M_{kM}(\Gamma).$$

What Sturm's bound really means is the following.

For a congruence subgroup Γ, we put $q = e^{2\pi iz}$ as usual if $T \in \Gamma$. If not, then $T^N \in \Gamma$ for some N, and we set $q = e^{2\pi iz/N}$ for the minimal such $N \in \mathbb{N}$. Then $f \in M_k(\Gamma)$ has period N so we can write the Fourier expansion as $f(z) = \sum a_n q^n$.

Corollary 5.2.3. *Let $f(z) = \sum a_n q^n$, $g(z) = \sum b_n q^n \in M_k(\Gamma)$. If $a_n = b_n$ for $n \le \frac{k[\mathrm{PSL}_2(\mathbb{Z}):\Gamma]}{12}$, then $f = g$.*

Proof. Apply Sturm's bound to $f - g$. □

Corollary 5.2.4. *For a congruence subgroup Γ,*

$$\dim M_k(\Gamma) \le \frac{k}{12}[\mathrm{PSL}_2(\mathbb{Z}) : \Gamma] + 1.$$

In particular,

$$\dim M_k(N) \le \frac{k}{12} N \prod_{p|N}\left(1 + \frac{1}{p}\right) + 1.$$

Proof. Sturm's bound says the linear map given by

$$f(z) = \sum_{n\ge 0} a_n q^n \mapsto (a_0, a_1, \ldots, a_r)$$

where $r = \frac{k}{12}[\mathrm{PSL}_2(\mathbb{Z}) : \Gamma]$ is an injective map from $M_k(\Gamma)$ to $\mathbb{C}^{r+1}$.

The second statement follows from Corollary 3.4.3. □

We remark that the bound for $\dim M_k(N)$ is asymptotically of the right order (for fixed k and N large)—the "main term" in the general dimension formulas for both $\dim M_k(N)$ and $\dim S_k(N)$ is $\frac{k-1}{12} N \prod_{p|N} \left(1 + \frac{1}{p}\right)$.

Example 5.2.5. *Let $k = 0$, then we see for any Γ, $\dim M_0(\Gamma) \leq 1$. On the other hand any constant function lies in $M_0(\Gamma)$, so $M_0(\Gamma) = \mathbb{C}$. This provides another proof that there are no nonconstant holomorphic modular functions (cf. Corollary 4.1.12).*

Example 5.2.6. *Let $N = 4$. Then we see*

$$\dim M_k(4) \leq \frac{k}{12} 4 \cdot \frac{3}{2} + 1 = \frac{k}{2} + 1.$$

In particular

$$\dim M_4(4) \leq \frac{4}{2} + 1 = 3.$$

On the other hand, we have constructed 3 linearly independent elements of $M_4(4)$ (none of which are cusp forms), namely $E_{4,4}$, $E_{4,2}$ and E_4 (cf. Example 4.4.7). Hence

$$\dim M_4(4) = 3.$$

This justifies the formulas for the number of representations of n as a sum of 8 squares (resp. triangular numbers) in Example 4.4.7 (resp. Exercise 4.4.9).

Exercise 5.2.7. *Show $\dim M_k(\Gamma_0(2)) = 2$ for $k = 4, 6$.*

Observe that even in small weights, $\dim M_k(N) > 1$ for higher levels, in contrast to the level 1 case, where the dimension did not jump to 2 until the occurrence of the cusp form Δ in weight 12. The reason for this difference is that there is an Eisenstein series $E_{k,d} \in M_k(N)$ for each $d|N$, and these are all different. In general Γ will have multiple cusps, and one can construct an "Eisenstein series for each cusp" (which turn out to be the same as the $E_{k,d}$ for $d|N$ when $\Gamma = \Gamma_0(N)$ and N is a product of distinct primes or $N = 4$).

While we haven't shown this yet, $E_4 - E_{4,4}$ (or anything in $M_4(4)$) is not a cusp form, even though it vanishes at $i\infty$. We point this out to emphasize that in higher level, one really needs to check a form vanishes at *all* cusps in order to verify it is a cusp form.

However, one sometimes has 1-dimensional spaces in higher level (besides in the trivial weight $k = 0$).

Exercise 5.2.8. *(a) Consider the Eisenstein series $E_{2,2}(z) = E_2(z) - 2E_2(2z)$ from Exercise 4.2.18. Show $E_{2,2}(z) \in M_2(2)$ and $E_{2,2}(2z) \in M_2(4)$.*

(b) Deduce $\dim M_2(2) = 1$ and $\dim M_2(4) = 2$.

(c) Show $r_4(n) = \begin{cases} 8\sigma_1(n) & n \text{ odd} \\ 24\sigma_1(n_0) & n = 2^r n_0,\ n_0 \text{ odd.} \end{cases}$

One can show the first cusp form for $\Gamma_0(4)$ occurs at weight 6.

Exercise 5.2.9. *(a) Show $F_\eta(z) = \eta^{12}(2z) \in S_6(4)$.*

(b) Using Sturm's bound, show $\dim S_6(4) = 1$.

(c) For $k \geq 6$ even, show $S_k(4) = F_\eta \cdot M_{k-6}(4)$.

(d) Show for $k \geq 0$,

$$\dim M_k(4) = \frac{k}{2} + 1.$$

Example 5.2.10. *Let us now consider the number $r_{12}(n)$ of representations of n as a sum of 12 squares. By Proposition 4.4.4, we know*

$$\vartheta^{12}(z) = \sum_{n=0}^{\infty} r_{12}(n)q^n = 1 + 2 \cdot 12q + 4 \cdot 66q^2 + 8 \cdot 220q^3 + \cdots \in M_6(4).$$

From the previous exercise, we know $M_6(4) = \langle E_6, E_{6,2}, E_{6,4}, F_\eta \rangle$. Note

$$E_6(z) = 1 - 504 \sum_{n=1}^{\infty} \sigma_5(n)q^n = 1 - 504 \left(q + 33q^2 + 244q^3 + \cdots\right)$$

$$E_{6,2}(z) = 1 - \frac{504}{31} \sum_{n=1}^{\infty} \left(32\sigma_5(n/2) - \sigma_5(n)\right) q^n = 1 + \frac{504}{31} \left(q + q^2 + 244q^3 + \cdots\right)$$

$$E_{6,4}(z) = E_{6,2}(2z) = 1 - \frac{504}{31} \sum_{n=1}^{\infty} \left(32\sigma_5(n/4) - \sigma_5(n/2)\right) q^n = 1 + \frac{504}{31} q^2 + \cdots$$

$$F_\eta(z) = \eta^{12}(2z) = q \prod_{n=1}^{\infty} (1 - q^{2n})^{12} = \sum_{n>0} a_n q^n = q - 12q^3 + \cdots .$$

One can verify that

$$\vartheta^{12} = -\frac{5}{336} E_6 + \frac{31}{1008} E_{6,2} + \frac{62}{63} E_{6,4} + 16F_\eta.$$

Consequently, one gets a formula for $r_{12}(n)$ in terms of σ_5 and the Fourier coefficients a_n of the cusp form F_η:

$$r_{12}(n) = 8\sigma_5(n) - 512\sigma_5(n/4) + 16a_n.$$

Because of the appearance of cusp forms of level 4 starting at weight 6, for $2k \geq 12$ with k even, one no longer gets as elementary an expression for $r_{2k}(n)$ as the Fourier coefficients of cusp forms are in some sense more mysterious. (While we've avoided modular forms of odd weight, there is also a cusp form of weight 5 for $\Gamma_1(4)$, so the formula for $r_{10}(n)$ also involves a cusp form.) However, as mentioned before, Hecke's bound will show the Fourier coefficients for cusp forms grow at a much slower rate, so one is at least able to get a nice asymptotic for $r_{2k}(n)$ just in terms of the Fourier coefficients of Eisenstein series. We will say a little more about this once we get to Hecke's bound.

We remark that in specific cases, it is still possible to get elementary formulas for $r_{2k}(n)$ with more work. For instance, Glaisher in fact derived an explicit elementary formula for $r_{12}(n)$ as

$$r_{12}(n) = (-1)^{n-1} 8 \sum_{d|n} (-1)^{d+n/d} d^5 + 4 \sum_{N(\alpha)=n} \alpha^4,$$

where α runs over all Hurwitz integers with norm n.

Appendix: Dimension Tables

Explicit dimensions for small k, N are given in Table 5.1 and Table 5.2. Note that, for a fixed N and $k \geq 4$, $\dim M_k(N) - \dim S_k(N)$ is constant. One may see this from comparing (5.2.2) and (5.2.3), and one sees $\dim M_k(N) - \dim S_k(N) = \epsilon_\infty$ for $k \geq 4$. This corresponds to the fact that the various cusps give rise to linearly independent Eisenstein series on $M_k(N)$, which are all holomorphic if $k \geq 4$, and these Eisenstein series together with the cusp forms linearly generate $M_k(N)$.

Table 5.1: Dimensions of spaces of modular forms for $2 \leq k \leq 8$

N	$M_2(N)$	$S_2(N)$	$M_4(N)$	$S_4(N)$	$M_6(N)$	$S_6(N)$	$M_8(N)$	$S_8(N)$
1	0	0	1	0	1	0	1	0
2	1	0	2	0	2	0	3	1
3	1	0	2	0	3	1	3	1
4	2	0	3	0	4	1	5	2
5	1	0	3	1	3	1	5	3
6	3	0	5	1	7	3	9	5
7	1	0	3	1	5	3	5	3
8	3	0	5	1	7	3	9	5
9	3	0	5	1	7	3	9	5
10	3	0	7	3	9	5	13	9
11	2	1	4	2	6	4	8	6
12	5	0	9	3	13	7	17	11
13	1	0	5	3	7	5	9	7
14	4	1	8	4	12	8	16	12
15	4	1	8	4	12	8	16	12
16	5	0	9	3	13	7	17	11
17	2	1	6	4	8	6	12	10
18	7	0	13	5	19	11	25	17
19	2	1	6	4	10	8	12	10
20	6	1	12	6	18	12	24	18
21	4	1	10	6	16	12	20	16
22	5	2	11	7	17	13	23	19
23	3	2	7	5	11	9	15	13
24	8	1	16	8	24	16	32	24
25	5	0	11	5	15	9	21	15
26	5	2	13	9	19	15	27	23
27	6	1	12	6	18	12	24	18
28	7	2	15	9	23	17	31	25
29	3	2	9	7	13	11	19	17
30	10	3	22	14	34	26	46	38
31	3	2	9	7	15	13	19	17
32	8	1	16	8	24	16	32	24
33	6	3	14	10	22	18	30	26
34	6	3	16	12	24	20	34	30
35	6	3	14	10	22	18	30	26
36	12	1	24	12	36	24	48	36
37	3	2	11	9	17	15	23	21
38	7	4	17	13	27	23	37	33
39	6	3	16	12	26	22	34	30
40	10	3	22	14	34	26	46	38

Table 5.2: Dimensions of spaces of modular forms for $10 \leq k \leq 16$

N	$M_{10}(N)$	$S_{10}(N)$	$M_{12}(N)$	$S_{12}(N)$	$M_{14}(N)$	$S_{14}(N)$	$M_{16}(N)$	$S_{16}(N)$
1	1	0	2	1	1	0	2	1
2	3	1	4	2	4	2	5	3
3	4	2	5	3	5	3	6	4
4	6	3	7	4	8	5	9	6
5	5	3	7	5	7	5	9	7
6	11	7	13	9	15	11	17	13
7	7	5	9	7	9	7	11	9
8	11	7	13	9	15	11	17	13
9	11	7	13	9	15	11	17	13
10	15	11	19	15	21	17	25	21
11	10	8	12	10	14	12	16	14
12	21	15	25	19	29	23	33	27
13	11	9	15	13	15	13	19	17
14	20	16	24	20	28	24	32	28
15	20	16	24	20	28	24	32	28
16	21	15	25	19	29	23	33	27
17	14	12	18	16	20	18	24	22
18	31	23	37	29	43	35	49	41
19	16	14	20	18	22	20	26	24
20	30	24	36	30	42	36	48	42
21	26	22	32	28	36	32	42	38
22	29	25	35	31	41	37	47	43
23	19	17	23	21	27	25	31	29
24	40	32	48	40	56	48	64	56
25	25	19	31	25	35	29	41	35
26	33	29	41	37	47	43	55	51
27	30	24	36	30	42	36	48	42
28	39	33	47	41	55	49	63	57
29	23	21	29	27	33	31	39	37
30	58	50	70	62	82	74	94	86
31	25	23	31	29	35	33	41	39
32	40	32	48	40	56	48	64	56
33	38	34	46	42	54	50	62	58
34	42	38	52	48	60	56	70	66
35	38	34	46	42	54	50	62	58
36	60	48	72	60	84	72	96	84
37	29	27	37	35	41	39	49	47
38	47	43	57	53	67	63	77	73
39	44	40	54	50	62	58	72	68
40	58	50	70	62	82	74	94	86

Chapter 6

Hecke operators

Recall the conjecture of Ramanujan that $\tau(mn) = \tau(m)\tau(n)$ for m, n relatively prime. While this specific result was shown by Mordell, it was Hecke who developed a general theory to determine which modular forms have multiplicative Fourier coefficients (i.e., $a_{mn} = a_m a_n$ whenever $\gcd(m, n) = 1$—note this imposes no condition on a_0). Note that if the Fourier coefficients a_n of some modular form are multiplicative, then they are determined by a_0 and the prime power Fourier coefficients a_{p^r}.

The following elementary exercise shows the Eisenstein series series E_k essentially have multiplicative Fourier coefficients.

Exercise 6.0.1. *If* $\gcd(m, n) = 1$*, show* $\sigma_k(m)\sigma_k(n) = \sigma_k(mn)$*.*

Namely, if we write

$$E_k(z) = 1 - \frac{2k}{B_k} \sum_{n \geq 1} \sigma_{k-1}(n) q^n = \sum a_n q^n,$$

the above exercise shows

$$a_m a_n = \left(\frac{2k}{B_k}\right)^2 \sigma_{k-1}(m)\sigma_{k-1}(n) = \left(\frac{2k}{B_k}\right)^2 \sigma_{k-1}(mn) = -\frac{2k}{B_k} a_{mn} = a_1 a_{mn}.$$

Or, if we simply consider the renormalized Eisenstein series,

$$E_k^*(z) = -\frac{B_k}{2_k}(z) E_k(z) = -\frac{B_k}{2k} + \sum_{n \geq 1} \sigma_{k-1}(n) q^n, \tag{6.0.1}$$

we see E_k^* has multiplicative Fourier coefficients. Now we can say that by "essentially multiplicative" Fourier coefficients, we mean some nonzero multiple of our modular form has multiplicative Fourier coefficients.

Hecke's idea for determining if a modular form has (essentially) multiplicative Fourier coefficients, very roughly, is the following. Consider an operator

$$U_p \left(\sum a_n q^n \right) = \sum a_{pn} q^n. \tag{6.0.2}$$

If this operator preserves $S_{12}(1)$, then

$$U_p\Delta = \lambda_p\Delta$$

for some $\lambda_p \in \mathbb{C}$, since $S_{12}(1) = \langle\Delta\rangle$. On the other hand,

$$U_p\Delta = U_p(q - 24q^2 + 252q^3 - \cdots) = \tau(p)q - 24\tau(p)q^2 + \cdots$$

so $\lambda_p = \tau(p)$ and therefore we would have $\tau(pn) = \tau(p)\tau(n)$ for all n.

However, τ is not *totally multiplicative*, i.e., $\tau(p^2) \neq \tau(p)^2$, so U_p does not preserve Δ. Instead, since Ramanujan predicted $\tau(p^2) = \tau(p)^2 - p^{11}$, one would expect $\tau(pn) = \tau(p)\tau(n) - p^{11}\tau(n/p)$ for $p|n$. Thus one can guess the correct operator (for weight 12) should be

$$T_p\left(\sum a_n q^n\right) = \sum_{p\nmid n} a_{pn}q^n + \sum_{p|n}\left(a_{pn} + p^{11}a_{n/p}\right)q^n. \tag{6.0.3}$$

Then one just needs to show T_p preserves $S_{12}(1)$.

6.1 Hecke operators for $\Gamma_0(N)$

While we could define the Hecke operators directly by their action on Fourier expansions along the lines of (6.0.3), it will be helpful (and more motivated) to think of them as acting on lattices.

First let's consider the case of $\mathrm{PSL}_2(\mathbb{Z})$. Recall $\mathrm{PSL}_2(\mathbb{Z})\backslash\mathfrak{H}$ parameterizes the space of lattices up to homothety (equivalence by $\mathbb{C}^\times$). Hence a weak modular form of weight 0 is simply a meromorphic function on equivalence classes of lattices.

What about weight k? Given $f \in M_k(1)$, consider the lattice $\Lambda = \langle 1, \tau\rangle$ with $\tau \in \mathfrak{H}$. Put $F(\Lambda) = f(\tau)$. Since $\langle\omega_1, \omega_2\rangle = \langle\omega_1', \omega_2'\rangle$ if and only if

$$\begin{pmatrix} a & b \\ c & d\end{pmatrix}\begin{pmatrix}\omega_1\\ \omega_2\end{pmatrix} = \begin{pmatrix}\omega_1'\\ \omega_2'\end{pmatrix}$$

one easily sees that $\langle 1, \tau\rangle = \langle 1, \tau'\rangle$ (with $\tau' \in \mathfrak{H}$ also) if and only if $\tau' = \tau + d$ for some $d \in \mathbb{Z}$. Thus f having period 1 implies $F(\Lambda)$ is well-defined on lattices of the form $\langle 1, \tau\rangle$.

Now consider an arbitrary lattice $\Lambda = \langle\omega_1, \omega_2\rangle$. We can write $\Lambda = \omega_1\langle 1, \omega_2/\omega_1\rangle = \lambda\langle 1, \tau\rangle$ where $\lambda = \omega_1$ and $\tau = \omega_2/\omega_1$. Thus, to see how to extend F to a function on all lattices it suffices to determine how F should behave under multiplying a lattice $\Lambda = \langle 1, \tau\rangle$ by a scalar λ. It's reasonable to ask that a scalar λ should just transform F by some factor, i.e.,

$$F(\lambda\Lambda) = c(\lambda)F(\Lambda).$$

The only condition we have is that this should be compatible with our definition of $F(\langle 1, \tau\rangle) = f(\tau)$. In other words, if $\lambda\langle 1, \tau'\rangle = \langle 1, \tau\rangle$, we need to ensure $F(\lambda\langle 1, \tau'\rangle) = F(\langle 1, \tau\rangle)$.

This boils down to the case where $\lambda = \tau$ and $\tau' = -\frac{1}{\tau}$. Here we require

$$c(\tau)f\left(-\frac{1}{\tau}\right) = F(\tau\langle 1, \frac{1}{\tau}\rangle) = F(\langle 1, \tau\rangle) = f(\tau).$$

But now the modularity of f implies $c(\tau) = \frac{1}{\tau^k}$. In other words, the weight k modular form f can be viewed as a homogeneous function F of degree $-k$ on the space of lattices of $\mathbb{C}$, i.e., a function F such that

$$F(\lambda\Lambda) = \frac{1}{\lambda^k}F(\Lambda)$$

for $\lambda \in \mathbb{C}^\times$. Call the space of such F by $\mathcal{L}(k)$.

Conversely, if F is a function of lattices such that $F(\lambda\Lambda) = \lambda^{-k}F(\Lambda)$ with $k \geq 0$ even, then one can check $f(z) = F(\langle 1, z\rangle) \in M_k(1)$. Hence there is a bijection between $M_k(1)$ and $\mathcal{L}(k)$.

Then for $\mathrm{PSL}_2(\mathbb{Z})$ we can define the n-th Hecke operator in terms of $F \in \mathcal{L}(k)$:

$$T_n(F(\Lambda)) = F\left(\sum_{\Lambda' \overset{n}{\subset} \Lambda} \Lambda'\right),$$

where the sum is over all $\Lambda' \subseteq \Lambda$ of index n. In other words T_n averages F over all sublattices of index n. It is clear that T_nF is still a function of lattices, and it is homogeneous of degree $-k$. Clearly $T_1F = F$.

Thus the correspondence between modular forms $f \in M_k(1)$ and $F \in \mathcal{L}(k)$ induces an action of the Hecke operators

$$T_n : M_k(1) \to M_k(1).$$

Let's see how T_n translates to an operator on $M_k(1)$.

First observe that the sublattices Λ' of $\Lambda = \langle 1, \tau\rangle$ of index n are precisely $\Lambda' = \langle 1, \tau'\rangle$, where

$$\begin{pmatrix}\tau' \\ 1\end{pmatrix} = \begin{pmatrix}a & b \\ c & d\end{pmatrix}\begin{pmatrix}\tau \\ 1\end{pmatrix} \quad \text{for } \begin{pmatrix}a & b \\ c & d\end{pmatrix} \in \mathcal{M}_n(\mathbb{Z})$$

where $\mathcal{M}_n(\mathbb{Z})$ denotes the 2×2 integer matrices of determinant n. For such a

$$\Lambda' = \langle a\tau + b, c\tau + d\rangle = (c\tau + d)\langle\frac{a\tau + b}{c\tau + d}, 1\rangle,$$

we have

$$F(\Lambda') = F\left((c\tau + d)\langle\frac{a\tau + b}{c\tau + d}, 1\rangle\right) = (c\tau + d)^{-k} f\left(\frac{a\tau + b}{c\tau + d}\right) = f|_{\mu,k}(\tau)$$

where we extend the slash operator

$$f|_{\mu,k}(\tau) = (c\tau + d)^{-k} f\left(\frac{a\tau + b}{c\tau + d}\right) \quad \text{for} \quad \mu = \begin{pmatrix}a & b \\ c & d\end{pmatrix} \in \mathcal{M}_n(\mathbb{Z}).$$

(Note in general for $\mu = \begin{pmatrix}a & b \\ c & d\end{pmatrix} \in \mathrm{GL}_2^+(\mathbb{Q})$ (the superscript + means positive determinant) one typically defines the slash operator with a factor of $\det(\mu)^{k/2}$, so our notation departs from the standard here, but seems the most reasonable for our present purpose.)

Given $\mu_1, \mu_2 \in \mathcal{M}_n(\mathbb{Z})$, one can check $\mu_1\langle 1, \tau\rangle = \mu_2\langle 1, \tau\rangle$ if and only if $\mu_2 = \gamma\mu_1$ for some $\gamma \in \mathrm{SL}_2(\mathbb{Z})$. Hence on the level of modular forms, we have

$$T_n f = \sum_{\mu \in \mathrm{SL}_2(\mathbb{Z})\backslash \mathcal{M}_n(\mathbb{Z})} f|_{\mu,k}.$$

For our arithmetic purposes, it is better to introduce a normalization factor of n^{k-1} in the Hecke operator, which we will do below. The above discussion was just motivation for how to define Hecke operators for $\mathrm{PSL}_2(\mathbb{Z})$.

While we will not go through the details (cf. [Kob93]), a similar argument can be made for the modular groups $\Gamma_0(N)$ (as well as for $\Gamma_1(N)$ and $\Gamma(N)$). The idea is that $\Gamma_0(N)\backslash\mathfrak{H}$ parameterizes pairs (Λ, C) where Λ is a lattice in $\mathbb{C}$ and C is a cyclic subgroup of Λ of order N. Then one can identify modular forms $f \in M_k(N)$ with homogeneous functions of degree $-k$ on pairs (Λ, C) and define Hecke operators T_n similarly, though some care must be taken when $\gcd(n, N) > 1$.

Let

$$\mathcal{M}_{n,N}(\mathbb{Z}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathcal{M}_n(\mathbb{Z}) | c \equiv 0 \bmod N \right\} / \{\pm I\}.$$

Definition 6.1.1. *Suppose* $\gcd(n, N) = 1$. *We define the* n*-th* Hecke operator T_n *on* $M_k(N)$ *by*

$$T_n f = n^{k-1} \sum_{\mu \in \Gamma_0(N)\backslash \mathcal{M}_{n,N}(\mathbb{Z})} f|_{\mu,k}.$$

We will see the normalization factor n^{k-1} will make the action on Fourier coefficients nice.

Observe that for $f \in M_k(N)$, $\mu \in \mathcal{M}_{n,N}(\mathbb{Z})$ and $\gamma \in \Gamma_0(N)$ we have

$$f|_{\gamma\mu,k} = (f|_{\gamma,k})|_{\mu,k} = f|_{\mu,k},$$

so the above sum over coset representatives is well defined. It is also easy to see that T_n is linear. Now we show it actually is an operator on $M_k(N)$, as well as $S_k(N)$.

Unless otherwise specified, *we assume* $\gcd(n, N) = 1$ *in what follows.*

Proposition 6.1.2. *We have* $T_n : M_k(N) \to M_k(N)$ *and* $T_n : S_k(N) \to S_k(N)$.

Proof. Let $\gamma \in \Gamma_0(N)$. Then

$$\begin{aligned}(T_n f)|_{\gamma,k} = n^{k-1} \left(\sum_{\mu \in \Gamma_0(N)\backslash \mathcal{M}_{n,N}(\mathbb{Z})} f|_{\mu,k} \right) |_{\gamma,k} &= n^{k-1} \sum_{\mu \in \Gamma_0(N)\backslash \mathcal{M}_{n,N}(\mathbb{Z})} f|_{\mu\gamma,k} \\ &= n^{k-1} \sum_{\mu \in \Gamma_0(N)\backslash \mathcal{M}_{n,N}(\mathbb{Z})} f|_{\mu,k} = T_n f\end{aligned}$$

since right multiplication by $\gamma \in \Gamma_0(N)$ preserves $\mathcal{M}_{n,N}(\mathbb{Z})$, and therefore simply permutes the cosets $\Gamma_0(N)\backslash\mathcal{M}_{n,N}(\mathbb{Z})$.

Note that each $f|_{\mu,k}$ is holomorphic on $\mathfrak{H}$, therefore $T_n f$ is. To see $T_n f$ is holomorphic at each cusp, let $\tau \in \mathrm{PSL}_2(\mathbb{Z})$ and consider $f|_{\mu,k}|_{\tau,k} = f|_{\mu\tau,k} = f|_{\mu',k}$ where $\mu' =$

$\mu\tau = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathcal{M}_n(\mathbb{Z})$ (but not necessarily $\mathcal{M}_{n,N}(\mathbb{Z})$). As $\mathrm{Im}(z) \to \infty$, $f|_{\mu',k}(z) \to \lim_{\mathrm{Im}(z)\to\infty} \frac{1}{(cz)^k} f\left(\frac{a}{c}\right)$ which has a finite limit because f is holomorphic at the cusp $\frac{a}{c}$. Thus each $f|_{\mu'}$ is holomorphic at $i\infty$, and $T_n f$ is holomorphic at the cusps.

The same argument we used for holomorphy at the cusps shows that, if f vanishes at the cusps, so does $T_n f$. □

Lemma 6.1.3. *A set of coset representatives for* $\Gamma_0(N)\backslash\mathcal{M}_{n,N}(\mathbb{Z})$ *is*

$$\left\{ \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} : a, d > 0,\ ad = n,\ 0 \le b < d \right\}.$$

Proof. Take $\begin{pmatrix} a & b \\ cN & d \end{pmatrix} \in \mathcal{M}_{n,N}(\mathbb{Z})$ representing some coset in $\Gamma_0(N)\backslash\mathcal{M}_{n,N}(\mathbb{Z})$. Since $\gcd(n, N) = 1$, we know $\gcd(a, cN) = 1$. Then there exist $x, y \in \mathbb{Z}$ such that $\begin{pmatrix} x & y \\ cN & -a \end{pmatrix} \in \Gamma_0(N)$. Now observe

$$\begin{pmatrix} x & y \\ cN & -a \end{pmatrix} \begin{pmatrix} a & b \\ cN & d \end{pmatrix} = \begin{pmatrix} ax + cyN & bx + dy \\ 0 & bcN - ad \end{pmatrix},$$

which means we may assume $c = 0$ for our coset representative.

Further, given $\begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \in \mathcal{M}_{n,N}(\mathbb{Z})$, we may replace it by the coset representative

$$\begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} = \begin{pmatrix} a & b + dx \\ 0 & d \end{pmatrix}$$

where we choose x so that $0 \le b + dx < d$, i.e., we may just assume $0 \le b < d$. Since $n = ad > 0$, a and d have the same sign, we may multiply by $-I$ if necessary to assume $a, d > 0$. This shows any coset has a representative of the desired form.

Now consider $\begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$ and $\begin{pmatrix} a' & b' \\ 0 & d' \end{pmatrix}$ such that $ad = a'd' = n$, $0 \le b < d$ and $0 \le b' < d'$. Assume they represent the same coset, i.e.,

$$\begin{pmatrix} a' & b' \\ 0 & d' \end{pmatrix} = \begin{pmatrix} r & s \\ tN & u \end{pmatrix} \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} = \begin{pmatrix} ra & rb + sd \\ tNa & tNb + ud \end{pmatrix}$$

for some $\begin{pmatrix} r & s \\ tN & u \end{pmatrix} \in \Gamma_0(N)$. First we see this means $t = 0$, which means $r = u = 1$ (up to ± 1). This means $a' = a$ $d' = d$, and $b' = b + sd$. However $0 \le b' < d' = d$ implies we need $s = 1$, i.e., $\begin{pmatrix} r & s \\ tN & u \end{pmatrix} = I$, and therefore any two distinct matrices of the prescribed form represent distinct cosets. □

Theorem 6.1.4. *Let* $f(z) = \sum a_n q^n \in M_k(N)$ *and suppose* $\gcd(m, N) = 1$. *Then*

$$(T_m f)(z) = \sum b_n q^n$$

where

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

In particular, if $m = p$ is prime, then

$$b_n = \begin{cases} a_{pn} & p \nmid n \\ a_{pn} + p^{k-1} a_{n/p} & p|n \end{cases}$$

so

$$(T_p f)(z) = \sum_{n \not\equiv 0 \bmod p} a_{pn} q^n + \sum_{n \equiv 0 \bmod p} \left(a_{pn} + p^{k-1} a_{n/p} \right) q^n.$$

Proof. By the previous lemma, we have

$$(T_m f)(z) = m^{k-1} \sum f|\begin{pmatrix} a & b \\ 0 & d \end{pmatrix},k$$

where the sum runs over $a, b, d \in \mathbb{Z}_{\geq 0}$ such that $ad = m$ and $0 \leq b < d$. Note

$$f|\begin{pmatrix} a & b \\ 0 & d \end{pmatrix},k(z) = \frac{1}{d^k} f\left(\frac{az+b}{d}\right) = \frac{1}{d^k} \sum_{n=0}^{\infty} a_n e^{2\pi i \frac{a}{d} zn} e^{2\pi i \frac{b}{d} n} = \frac{1}{d^k} \sum_{n=0}^{\infty} \zeta_d^{bn} a_n q^{\frac{a}{d}n}.$$

For fixed a, d we will be considering the sum

$$\sum_{0 \leq b < d} \frac{1}{d^k} f\left(\frac{az+b}{d}\right) = \frac{1}{d^k} \sum_{n=0}^{\infty} \left(\sum_{0 \leq b < d} \zeta_d^{bn} \right) a_n q^{\frac{a}{d}n}.$$

Note that the inner sum $\sum_{0 \leq b < d} \zeta_d^{bn}$ will just be a sum over all d-th roots of unity, and therefore vanish, unless $d|n$, in which case it is just d. Thus

$$\sum_{0 \leq b < d} \frac{1}{d^k} f\left(\frac{az+b}{d}\right) = \frac{1}{d^{k-1}} \sum_{n=0}^{\infty} a_{dn} q^{an}.$$

Hence

$$\begin{aligned}(T_m f)(z) &= m^{k-1} \sum_{ad=m} \sum_{0 \leq b < d} \frac{1}{d^k} f\left(\frac{az+b}{d}\right) = m^{k-1} \sum_{ad=m} \left(\frac{1}{d^{k-1}} \sum_{n=0}^{\infty} a_{dn} q^{an} \right) \\ &= \sum_{n=0}^{\infty} \sum_{a|m} \left(a^{k-1} a_{mn/a} q^{an} \right) = \sum_{n=0}^{\infty} \sum_{d|\gcd(m,n)} d^{k-1} a_{mn/d^2} q^n.\end{aligned}$$

□

Corollary 6.1.5. *The Ramanujan tau function satisfies $\tau(mn) = \tau(m)\tau(n)$ whenever $\gcd(m, n) = 1$ and $\tau(p^r) = \tau(p)\tau(p^{r-1}) - p^{11}\tau(p^{r-2})$.*

Proof. Consider

$$(T_p\Delta)(z) = \sum_{p\nmid n} \tau(pn)q^n + \sum_{p|n} \left(\tau(pn) + p^{11}\tau(n/p)\right) q^n.$$

On the other hand, T_p acts on $S_{12}(1) = \mathbb{C}\Delta$ so

$$(T_p\Delta)(z) = \lambda\Delta(z)$$

for some $\lambda \in \mathbb{C}^\times$. The 1st Fourier coefficients of Δ and $T_p\Delta$ are just 1 and $\tau(p)$, so $\lambda = \tau(p)$. Comparing the n-th Fourier coefficients, we see

$$\begin{cases} \tau(pn) = \lambda\tau(n) = \tau(p)\tau(n) & p \nmid n \\ \tau(pn) + p^{11}\tau(n/p) = \lambda\tau(n) = \tau(p)\tau(n) & p|n. \end{cases}$$

The former equation proves the multiplicativity when $m = p$ and the latter equation proves the recursion relation for $\tau(p^r)$.

To obtain the general multiplicativity law $\tau(mn) = \tau(m,n)$ for $\gcd(m,n) = 1$, we simply use the same argument as above with T_m instead of T_p. □

The above argument applies in a more general setting.

Exercise 6.1.6. *Suppose* $f(z) = \sum a_n q^n \in S_k(N)$ *(resp.* $M_k(N)$*) and* $\dim S_k(N) = 1$ *(resp.* $\dim M_k(N) = 1$*). Show*

(i) If $\gcd(m,N) = \gcd(n,N) = \gcd(m,n) = 1$*, then* $a_1a_{mn} = a_ma_n$*. In particular, if* $f(z) \neq 0$ *we must have* $a_1 \neq 0$*, and if we normalize* f *such that* $a_1 = 1$*, the Fourier coefficients are multiplicative.*

(ii) If $a_1 = 1$ *and* $p \nmid N$*, then* $a_{p^n} = a_p a_{p^{n-1}} - p^{k-1}a_{p^{n-2}}$ *for* $n \geq 2$*.*

What this means is we can use Hecke operators to *compute* values of Fourier coefficients from just knowing what the T_p's are.

Exercise 6.1.7. *Using the fact that* $\dim S_{16}(1) = 1$*, use the previous exercise to help compute the first 10 Fourier coefficients of* $E_4\Delta$*.*

Now you might ask, for what modular forms f are the Fourier coefficients multiplicative in the sense $a_1a_{mn} = a_ma_n$ for $\gcd(m,n) = 1$? By Exercise 6.0.1, we know this is true for the Eisenstein series E_k, and now we have seen it is true for Δ, and by Exercise 6.1.6, any $E_k\Delta$ where $k = 4, 6, 8, 10, 14$.

In general, if you have two power series (or Fourier expansions) f and g with multiplicative coefficients, the formal product fg will not have multiplicative Fourier coefficients, so there is no reason to expect all modular forms—or even products of modular forms with multiplicative Fourier coefficients—to have multiplicative coefficients. For instance, starting in weight 24 for the full modular group, we have a space of cusp forms of dimension greater than 1, namely $S_{24}(1) = \langle \Delta^2, E_{12}\Delta \rangle$. Since

$$\Delta(z) = q - 24q^2 + 252q^3 - 1472q^4 + 4830q^4 - 6048q^6 + \cdots$$

we compute

$$\Delta^2(z) = q^2 - 48q^3 + 1080q^4 - 2944q^5 + 143820q^6 + \cdots .$$

Right away, we see two things: 1) the Fourier coefficients of q^2 and q^3 do not multiply to give the Fourier coefficient of q^6, and 2) $a_1 = 0$ so the multiplicative property $a_1 a_n = a_m a_n$ for $\gcd(m, n) = 1$ would mean $a_n = 0$ whenever n is divisible by two different primes, which is obviously not the case. (In fact something stronger is true, see Corollary 6.1.11 below.)

One can also check that

$$E_{12} = 1 + \frac{65520}{691}\left(q + 2049q^2 + 177148q^3 + 4196353q^4 + 48828126q^5 + \cdots\right)$$

so

$$E_{12}\Delta = q + \frac{65520}{691}\left(\frac{2039}{2730}q^2 + \frac{527191}{260}q^3 + \frac{525013709}{4095}q^4 + \cdots - \frac{666536766}{65}q^6 + \cdots\right)$$

While the first Fourier coefficient is 1 here, again one sees the products of the q^2 and q^3 coefficients does not give the q^6 coefficient (it's not even the right sign!). So one cannot hope that the Fourier coefficients of $E_k\Delta$ are multiplicative in general.

You might now ask, are there *any* elements of $S_{24}(1)$ with multiplicative Fourier coefficients or does this property of having multiplicative Fourier coefficients only occur in small weight? We will see there always are cusp forms with multiplicative Fourier coefficients. First, let's make an observation in the weight 12 case: while $\frac{691}{65520}E_{12}$ and Δ have multiplicative Fourier coefficients, one can't expect this for any modular form in $M_{12}(1)$. Namely if f and g have multiplicative Fourier coefficients, $f + g$ will generally not. So while most forms in $M_{12}(1)$ do not have multiplicative Fourier coefficients, $M_{12}(1)$ is generated by two "nice" forms, $\frac{691}{65520}E_{12}$ and Δ, which do.

Using the theory of Hecke operators, we will show that there exists such a "nice" basis for $M_k(N)$ and $S_k(N)$. The basic idea is to show that the Hecke operators T_m and T_n commute for $\gcd(m, n) = 1$ (and m, n prime to N). Then one defines an inner product $\langle \cdot, \cdot \rangle$, called the Petersson inner product, on $M_k(N)$ and shows each T_n is Hermitian with respect to $\langle \cdot, \cdot \rangle$, i.e., $\langle T_n f, g\rangle = \langle f, T_n g\rangle$. Then a well known theorem in linear algebra says that a commuting family of operators which are Hermitian with respect to $\langle \cdot, \cdot \rangle$ can be simultaneously diagonalized by some orthonormal basis with respect $f_1, \ldots, f_r$ to $\langle \cdot, \cdot \rangle$ In other words, each f_i is an eigenform for T_n, i.e., $T_n f_i = \lambda f_i$ for some $\lambda \in \mathbb{C}$.

Definition 6.1.8. *Let $f \in M_k(N)$ be nonzero. We say f is a* **(Hecke) eigenform** *if, for each n relatively prime to N, there exists a* **(Hecke) eigenvalue** *$\lambda_n \in \mathbb{C}$ such that $T_n f = \lambda_n f$.*

We say f is a **normalized eigenform** *if the first nonzero Fourier coefficient is* 1.

The arguments from Corollary 6.1.5 and Exercise 6.1.6 apply to show any Hecke eigenform has multiplicative Fourier coefficients. Precisely, work out

***Exercise 6.1.9.** *Let $f = \sum a_n q^n \in M_k(N)$ be a Hecke eigenform. Suppose* $\gcd(m, n) = \gcd(m, N) = \gcd(n, N) = 1$. *Then $a_1 a_{mn} = a_m a_n$.*

In other words, the basis $f_1, \dots, f_r$ which diagonalizes the T_n's asserted above is a basis of Hecke eigenforms, and therefore a basis of $M_k(N)$ with multiplicative Fourier coefficients in the sense of the previous exercise.

We remark that there is a technicality here we have ignored, namely that the Petersson inner product $\langle f, g \rangle$ is only well defined when at least f or g is a cusp form. So the above argument will only technically show that $S_k(N)$ has a basis of eigenforms but, at least when $N = 1$, we will see how this implies one can extend the basis of eigenforms of $S_k(N)$ to a basis of eigenforms for $M_k(N)$.

We remark that for a eigencusp form (a Hecke eigenform which is a cusp form) the first nonzero Fourier coefficient should be a_1 in order for the Hecke operators to not force every Fourier coefficient to be zero.

Lemma 6.1.10. *Let $f = \sum a_n q^n \in M_k(1)$ be a Hecke eigenform and $k > 0$. Then $a_1 \neq 0$.*

Proof. Write $T_m f = \lambda_m f = \sum b_n q^n$. Then $b_1 = a_m$. Consequently

$$a_m = \lambda_m a_1.$$

In other words, if $a_1 = 0$, then $a_m = 0$ for all $n \geq 0$. Consequently $f(z) = a_0 \in M_0(1)$. □

Corollary 6.1.11. *Let $f \in M_k(1)$. Then $\Delta^2 f$ is not a Hecke eigenform.*

In particular, this shows Δ^2 is not an eigenform.

Now let's get to the first step in showing the existence of Hecke eigenforms, which is showing the Hecke operators are commutative.

Lemma 6.1.12. *Suppose $\gcd(m, N) = \gcd(n, N) = 1$ and $\gcd(m, n) = 1$. Then the Hecke operators on $M_k(N)$ satisfy $T_m T_n = T_{mn}$.*

Proof. Take $f(z) = \sum a_r q^r \in M_k(N)$. Write $T_n f(z) = \sum b_r q^r$ and $T_m(T_n f)(z) = \sum c_r q^r$. Then

$$b_r = \sum_{d \mid \gcd(n,r)} d^{k-1} a_{nr/d^2}$$

so

$$c_r = \sum_{e \mid \gcd(m,r)} e^{k-1} b_{mr/e^2} = \sum_{e \mid \gcd(m,r)} e^{k-1} \sum_{d \mid \gcd(n, mr/e^2)} d^{k-1} a_{mnr/d^2e^2}.$$

Since $\gcd(m, n) = 1$, d runs over the divisors of $\gcd(n, r/e)$, and therefore $d' = de$ runs over the divisors of mn and

$$c_r = \sum_{d' \mid \gcd(mn,r)} (d')^{k-1} a_{mnr/(d')^2}$$

so $T_m T_n = T_{mn}$. □

Lemma 6.1.13. *Suppose $p \nmid N$. The Hecke operators on $M_k(N)$ satisfy*

$$T_{p^r} T_{p^s} = \sum_{d \mid \gcd(p^r, p^s)} d^{k-1} T_{p^{r+s}/d^2}. \tag{6.1.1}$$

Proof. The proof here is modeled on the one in [Apo90, Theorem 6.13], but there appear to me to be errors in the $r = 1$ case of *loc. cit.* I believe I have corrected them.

We may as well assume $r \leq s$.

First let's show the $r = 1$ case, which just says

$$T_p T_{p^s} = T_{p^{s+1}} + p^{k-1} T_{p^{s-1}}.$$

For $f \in M_k(n)$,

$$\begin{aligned} T_{p^s} f(z) &= p^{s(k-1)} \sum_{0 \leq i \leq s} \sum_{0 \leq b < p^i} f \Big| \begin{pmatrix} p^{s-i} & b \\ 0 & p^i \end{pmatrix}, k \\ &= p^{s(k-1)} \sum_{0 \leq i \leq s} p^{-ik} \sum_{0 \leq b < p^i} f\left(\frac{p^{s-i} z + b}{p^i} \right). \end{aligned} \tag{6.1.2}$$

In particular

$$T_p g(z) = p^{(k-1)} g(pz) + p^{-1} \sum_{0 \leq b' < p} g\left(\frac{z + b'}{p} \right)$$

so

$$T_p T_{p^s} f(z) = p^{(s+1)(k-1)} \sum_{0 \leq i \leq s} p^{-ik} \sum_{0 \leq b < p^i} f\left(\frac{p^{s+1-i} z + b}{p^i} \right) \tag{6.1.3}$$

$$+ p^{-1} p^{s(k-1)} \sum_{0 \leq b' < p} \sum_{0 \leq i \leq s} p^{-ik} \sum_{0 \leq b < p^i} f\left(\frac{p^{s-i}(z + b') + pb}{p^{i+1}} \right). \tag{6.1.4}$$

Note the $i = s$ term from (6.1.4) is

$$p^{-1-s} \sum_{0 \leq b' < p} \sum_{0 \leq b < p^s} f\left(\frac{z + b' + pb}{p^{s+1}} \right) = p^{-1-s} \sum_{0 \leq b < p^{s+1}} f\left(\frac{z + b}{p^{s+1}} \right).$$

Thus adding the $i = s$ term from (6.1.4) to (6.1.3) gives (6.1.2) with s replaced by $s + 1$, i.e., they sum to $T_{p^{s+1}} f(z)$.

Now the remaining terms, i.e., the $i < s$ terms from (6.1.4), sum to

$$p^{-1} p^{s(k-1)} \sum_{0 \leq b' < p} \sum_{0 \leq i \leq s-1} p^{-ik} \sum_{0 \leq b < p^i} f\left(\frac{p^{s-1-i} z + b + p^{s-1-i} b'}{p^i} \right)$$

If $i \leq \frac{s-1}{2}$, then $\frac{p^{s-1-i} b'}{p^i} \in \mathbb{Z}$, so by periodicity of f there is no dependence on b' and the contribution for such a fixed i is

$$p^{s(k-1)} p^{-ik} \sum_{0 \leq b < p^i} f\left(\frac{p^{s-1-i} z + b}{p^i} \right). \tag{6.1.5}$$

In fact, for any i, $b + p^{s-1-i} b'$ mod p^i runs over the set of residue classes mod p^i exactly p times, so the contribution for any i is given again by (6.1.5). This proves the $r = 1$ case.

Now we suppose (6.1.1) is true up to some fixed r. By the $r = 1$ case we have

$$T_{p^{r+1}}T_{p^s} = (T_pT_{p^r})\,T_{p^s} - p^{k-1}T_{p^{r-1}}T_{p^s}.$$

By the inductive assumption that (6.1.1) is true for r, we also have

$$T_p\,(T_{p^r}T_{p^s}) = \sum_{0\le i\le r} p^{i(k-1)}T_pT_{p^{r+s-2i}}.$$

Comparing these, and making another use of the $r = 1$ case, gives

$$\begin{aligned} T_{p^{r+1}}T_{p^s} &= \sum_{0\le i\le r} p^{i(k-1)}T_pT_{p^{r+s-2i}} - p^{k-1}T_{p^{r-1}}T_{p^s} \\ &= \sum_{0\le i\le r} \left(p^{i(k-1)}T_{p^{r+s+1-2i}} + p^{(i+1)(k-1)}T_{p^{r+s-1-2i}}\right) - p^{k-1}T_{p^{r-1}}T_{p^s}. \end{aligned}$$

Expanding out the last term using the $r-1$ case of (6.1.1) gives

$$T_{p^{r+1}}T_{p^s} = \sum_{0\le i\le r+1} p^{i(k-1)}T_{p^{r+s+1-2i}}.$$

□

The previous two lemmas combine to give the following.

Corollary 6.1.14. *The Hecke operators $\{T_n\}_{\gcd(n,N)=1}$ on $M_k(N)$ commute.*

The former lemma also shows knowing what the prime power Hecke operators do tells you what all Hecke operators do, and the latter lemma shows that the prime power Hecke operators are determined by the prime Hecke operators. In particular, we see

Corollary 6.1.15. *If $f \in M_k(N)$ is an eigenform for each T_p with $p \nmid N$ prime, i.e., $T_pf = \lambda_p f$ for some $\lambda_p \in \mathbb{C}$, then f is a Hecke eigenform.*

Proof. The $r = 1$ case of (6.1.1) says, for $j = s - 1 \ge 2$,

$$T_{p^j} = T_pT_{p^{j-1}} - p^{k-1}T_{p^{j-2}}.$$

If $T_{p^i}f = \lambda_{p^i}f$ for some λ_{p^i} whenever $i < j$, then we see

$$T_{p^j}f = T_p(\lambda_{p^{j-1}}f) - p^{k-1}\lambda_{p^{j-2}}f = \lambda_{p^j}f$$

where

$$\lambda_{p^j} = \left(\lambda_p\lambda_{p^{j-1}} - p^{k-1}\lambda_{p^{j-2}}\right). \tag{6.1.6}$$

Hence by induction, we see if $T_pf = \lambda_p f$ then $T_{p^j}f = \lambda_{p^j}f$ for some λ_{p^j}—precisely, with λ_p^j satisfying the recursion in (6.1.6).

Now let n be relatively prime to N, and write $n = p_1^{e_1}\dots p_r^{e_r}$. By Lemma 6.1.12, we have

$$T_nf = \left(\prod_{i=1}^{r} T_{p_i^{e_i}}\right) f = \prod_{i=1}^{r} \lambda_{p_i^{e_i}} f.$$

□

6.2 Petersson inner product

We know that $M_k(\Gamma)$ is a finite dimensional complex vector space, and therefore can be made into a Hilbert space, i.e., an inner product space. The standard way to make a function space into a Hilbert space is with the L^2 inner product. Namely, if $f, g \in L^2(X)$ for some space X, then the inner product is given by

$$\langle f, g \rangle = \int_X f(x)\overline{g(x)}\, dx.$$

While modular forms are not L^2 on $\mathfrak{H}$, they are essentially L^2 on relevant Riemann surface $X = \Gamma\backslash\mathfrak{H}$. We say essentially here, because they are of course not actually functions on $\Gamma\backslash\mathfrak{H}$ due to the automorphy transformation factor except in the uninteresting (constant) case of $k = 0$.

Let's see what happens when we naively try to make $f \in M_k(\Gamma)$ into a function on $X = \Gamma\backslash\mathfrak{H}$. For simplicity, consider $\Gamma = \mathrm{PSL}_2(\mathbb{Z}) = \langle S, T\rangle$. We already have $f(Tz) = f(z+1) = f(z)$ so f descends to a function on the infinite cylinder $\langle T\rangle\backslash\mathfrak{H}$. We would like to modify f to a function $F = F_f$ which still satisfies $F(z+1) = F(z)$ but also satisfies $F(Sz) = F(-1/z) = F(z)$.

Since the invariance under T is a transformation rule in $x = \mathrm{Re}(z)$ which we don't want to mess up, we might just try to impose invariance under S by modifying $y = \mathrm{Im}(z)$. Since $f(S \cdot iy) = f(-1/iy) = f(i/y) = (iy)^k f(iy)$, it makes sense to consider the function (which will be called a **Maass form**)

$$F(z) := y^{k/2} f(z) = \mathrm{Im}(z)^{k/2} f(z). \tag{6.2.1}$$

Then we still have $F(z+1) = F(z)$ for all z and now

$$F(-1/iy) = F(i/y) = y^{-k/2} f(i/y) = \pm i^k y^{k/2} f(iy) = (-1)^{k/2} F(iy),$$

so at least $F(-1/iy) = F(iy)$ when $k \equiv 0 \bmod 4$. (One could let $F(z) = (iy)^{k/2} f(z)$ so that we actually have $F(-1/iy) = F(iy)$, but one typically considers F as defined in (6.2.1), since the sign $(-1)^{k/2}$ will not matter in the end.) It doesn't quite satisfy $F(-1/z) = F(z)$ for all z but it will be close, and we can say how close.

In general, for any $f \in M_k(\Gamma)$, we can associate to f the Maass form F defined by (6.2.1). Noting that

$$|j(\gamma, z)|^2 = \frac{\mathrm{Im}(z)}{\mathrm{Im}(\gamma z)}, \quad \gamma \in \mathrm{SL}_2(\mathbb{R}) \tag{6.2.2}$$

(cf. (3.2.1)), we see

$$F(\gamma z) = \mathrm{Im}(\gamma z)^{k/2} f(\gamma z) = \mathrm{Im}(\gamma z)^{k/2} j(\gamma, z)^k f(z) = j_0(\gamma, z)^k \mathrm{Im}(z)^{k/2} f(z) = j_0(\gamma, z)^k F(z)$$

for $\gamma \in \Gamma$, where

$$j_0(\gamma, z) = \frac{j(\gamma, z)}{|j(\gamma, z)|}.$$

In other words, the Maass form F (not holomorphic) is not quite invariant under Γ, but it is up to a factor $j_0(\gamma, z)$ of absolute value 1. This is good enough for our purposes.[1] Namely, if also $g \in M_k(\Gamma)$ and $G(z) = \mathrm{Im}(z)^{k/2} g(z)$ is the associated Maass form, it means the product

$$F(\gamma z)\overline{G(\gamma z)} = j_0(\gamma, z)^k \overline{j_0(\gamma, z)}^k F(z)\overline{G(z)} = F(z)\overline{G(z)}$$

is invariant under Γ.

Consequently, we can define the inner product

$$\langle f, g \rangle = \int_{\Gamma\backslash\mathfrak{H}} F(z)\overline{G(z)}\, d\omega,$$

where $d\omega$ is an area measure on the Riemann surface $X = \Gamma\backslash\mathfrak{H}$. Now in practice, you want to express this as an integral on a fundamental domain for Γ inside $\mathfrak{H}$, so we need to know what the hyperbolic measure should be.

Lemma 6.2.1. *The measure $\frac{dx\,dy}{y^2} = \frac{dz\,d\bar{z}}{\mathrm{Im}(z)^2}$ is an invariant measure on the hyperbolic plane $\mathfrak{H}$, i.e., for $\gamma \in \mathrm{PSL}_2(\mathbb{R})$ we have*

$$\int_{\mathfrak{H}} f(\gamma z)\, \frac{dx\,dy}{y^2} = \int_{\mathfrak{H}} f(z)\, \frac{dx\,dy}{y^2},$$

for any integrable (w.r.t. to this measure) function f on $\mathfrak{H}$.

Proof. Write $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, and let

$$w = \gamma z = \frac{(ad+bc)x + bd + acy^2 + iy}{c^2x^2 + c^2y^2 + d^2} = u + iv.$$

Note

$$\frac{dw}{dz} = \frac{1}{|cz+d|^2} = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x} = \frac{\partial v}{\partial y} - i\frac{\partial u}{\partial y}.$$

Then the Jacobian determiant

$$\left|\begin{matrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{matrix}\right| = \left|\begin{matrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{matrix}\right|^{-1} = \left|\begin{matrix} |cz+d|^{-2} & 0 \\ 0 & |cz+d|^{-2} \end{matrix}\right|^{-1} = |j(\gamma, z)|^4.$$

Since $|j(\gamma, z)|^2 = \frac{\mathrm{Im}(z)}{\mathrm{Im}(\gamma z)} = \frac{y}{v}$, we have

$$\int_{\mathfrak{H}} f(\gamma z)\, \frac{dx\,dy}{y^2} = \int_{\mathfrak{H}} f(w)|j(\gamma, z)|^4\, \frac{du\,dv}{y^2} = \int_{\mathfrak{H}} f(w)\, \frac{du\,dv}{v^2}.$$

□

[1] We remark that one can make f invariant under Γ at the expense of working on a larger space than $\mathfrak{H}$. Namely, one has the surjective map $\mathrm{PSL}_2(\mathbb{R}) \to \mathfrak{H}$ given by $\gamma \mapsto \gamma \cdot i$. Since $\mathrm{SO}(2)$ stabilizes i, one can view $\mathrm{PSL}_2(\mathbb{R})/\mathrm{SO}(2) \simeq \mathfrak{H}$. Thus one can lift f to a function on $\mathrm{PSL}_2(\mathbb{R})$ by $f(\gamma) = f(\gamma \cdot i)$ and define the **automorphic form** $\phi_f(\gamma) = j(\gamma, i)^{k/2} f(\gamma)$. This makes ϕ_f invariant under Γ so it is a function on $\Gamma\backslash\mathrm{PSL}_2(\mathbb{R})$. Hence, viewed as functions on $\mathrm{PSL}_2(\mathbb{R})$, the passage from modular forms to automorphic forms trades right $\mathrm{SO}(2)$ invariance for left Γ invariance.

Definition 6.2.2. *Let $f, g \in M_k(\Gamma)$. The* **Petersson inner product** *of f with g is defined to be*

$$\langle f, g \rangle := \frac{1}{[\mathrm{PSL}_2(\mathbb{Z}) : \Gamma]} \int_{\Gamma\backslash\mathfrak{H}} y^k f(z)\overline{g(z)} \frac{dx\,dy}{y^2} \tag{6.2.3}$$

whenever the integral converges.

The above discussion shows that the integrand is invariant under Γ, so the defining integral makes sense, provided it converges. Consequently, we can also write this integral as

$$\langle f, g \rangle := \frac{1}{[\mathrm{PSL}_2(\mathbb{Z}) : \Gamma]} \int_{\mathcal{F}} y^k f(z)\overline{g(z)} \frac{dx\,dy}{y^2} \tag{6.2.4}$$

where $\mathcal{F}$ is any fundamental domain for Γ.

The normalization factor $[\mathrm{PSL}_2(\mathbb{Z}) : \Gamma]^{-1}$ roughly cancels out the volume of $\Gamma\backslash\mathfrak{H}$, defined to be

$$\mathrm{vol}(\Gamma\backslash\mathfrak{H}) = \int_{\Gamma\backslash\mathfrak{H}} \frac{dx\,dy}{y^2}.$$

Lemma 6.2.3. *Let Γ be a congruence subgroup in $\mathrm{PSL}_2(\mathbb{Z})$. Then $\mathrm{vol}(\Gamma\backslash\mathfrak{H})$ is finite.*

Proof. Let $\mathcal{F}'$ be a fundamental domain for Γ. By Lemma 3.4.9,

$$\mathcal{F}' = \bigcup_{i=1}^{[\mathrm{PSL}_2(\mathbb{Z}):\Gamma]} \alpha_i \mathcal{F},$$

where $\mathcal{F}$ is the standard fundamental domain for $\mathrm{PSL}_2(\mathbb{Z})$ and $\alpha_i \in \mathrm{PSL}_2(\mathbb{Z})$. Since the area measure is invariant under the action of $\mathrm{PSL}_2(\mathbb{R})$,

$$\mathrm{vol}(\mathcal{F}') = \sum \mathrm{vol}(\alpha_i \mathcal{F}) = [\mathrm{PSL}_2(\mathbb{Z}) : \Gamma]\mathrm{vol}(\mathcal{F}),$$

so it suffices to show $\mathcal{F}$ has finite volume. Note

$$\mathrm{vol}(\mathcal{F}) = \int_{-1/2}^{1/2} \int_{\sqrt{1-x^2}}^{\infty} \frac{dx\,dy}{y^2} \leq \int_{-1/2}^{1/2} \int_{1/2}^{\infty} \frac{dx\,dy}{y^2} = 2\int_{-1/2}^{1/2} dx = 2.$$

□

The proof is of course valid not just for congruence subgroups, but any finite index subgroup of $\mathrm{PSL}_2(\mathbb{Z})$, and consequently any subgroup of $\mathrm{PSL}_2(\mathbb{R})$ commensurable with $\mathrm{PSL}_2(\mathbb{Z})$.

Exercise 6.2.4. *Compute* $\mathrm{vol}(X_0(1)) = \mathrm{vol}(\mathrm{PSL}_2(\mathbb{Z})\backslash\mathfrak{H})$.

Note the above lemma implies the Petersson inner product converges for (the admittedly not very interesting case of) $f, g \in M_0(\Gamma) = \mathbb{C}$. However for general weight k we do not always have convergence. For instance, in the standard fundamental domain of $\mathrm{PSL}_2(\mathbb{Z})$, the y^k can grow too fast for the Petersson inner product to converge unless $f(z)\overline{g}(z) \to 0$ as $y \to \infty$. Similarly, if we tend to a cusp in $\mathbb{Q}$, $f(z)\overline{g}(z)$ may grow too fast (cf. Section 4.3) for the inner product to converge. This suggests that when $k > 0$ we need f or g to be a cusp form for $\langle f, g \rangle$ to converge.

Proposition 6.2.5. *Let $f, g \in M_k(\Gamma)$. The Petersson inner product $\langle f, g\rangle$ converges if either $f \in S_k(\Gamma)$ or $g \in S_k(\Gamma)$.*

Proof. We assume $f \in S_k(\Gamma)$. The case $g \in S_k(\Gamma)$ is similar.

Let $\mathcal{F}$ be a fundamental domain for $\Gamma\backslash\mathfrak{H}$ and $\overline{\mathcal{F}}$ be its closure in $\overline{\mathfrak{H}}$. Write $\overline{\mathcal{F}} = K \cup \bigcup U_i$, where this is a finite union of subsets with K compact in $\mathfrak{H}$ and each $U_i \subset \overline{\mathfrak{H}}$ containing exactly one cusp, say z_i. Since $y^{k-2} f(z)\overline{g(z)}$ is bounded on K it suffices to show this is bounded on each U_i.

As in the proof of Lemma 4.5.8, we know $f(z) \to 0$ exponentially fast as $z \to z_i$, whereas $y^{k-2}\overline{g(z)}$ has at most a finite order pole at z_i. Thus $y^{k-2} f(z)\overline{g(z)}$ is bounded on each U_i, and because $\mathrm{vol}(\overline{\mathcal{F}}) = \mathrm{vol}(\mathcal{F}) < \infty$, the Petersson inner product converges. □

What this means is that the Petersson inner product defines an inner product (it is clearly sesquilinear, i.e., linear in the first vector and anti, or conjugate, linear in the second) on the space of cusp forms $S_k(N)$. This inner product almost extends to an inner product on $M_k(N)$, but $\langle f, g\rangle$ need not converge when both f and g are not cusp forms.

While we have failed to make $M_k(N)$ a Hilbert space, we have at least made $S_k(N)$ one. Most importantly, the inner product behaves symmetrically with respect to the action of the Hecke operators.

Lemma 6.2.6. *There exists a complete set of representatives $\{\mu_i\}$ for $\Gamma_0(N)\backslash\mathcal{M}_{n,N}(\mathbb{Z})$ such that $\{n\mu_i^{-1}\}$ is also a complete set of representatives for $\Gamma_0(N)\backslash\mathcal{M}_{n,N}(\mathbb{Z})$.*

Proof. First we note that both $\Gamma_0(N)\backslash\mathcal{M}_{n,N}(\mathbb{Z})$ and $\mathcal{M}_{n,N}(\mathbb{Z})/\Gamma_0(N)$ (which equals $\sigma_1(n) = \sum_{d|n} d$ by Lemma 6.1.3). This follows because the map $\mu \mapsto n\mu^{-1}$, which maps $\mathcal{M}_{n,N}(\mathbb{Z})$ to itself, induces a bijection of the right cosets $\Gamma_0(N)\backslash\mathcal{M}_{n,N}(\mathbb{Z})$ with the left cosets $\mathcal{M}_{n,N}(\mathbb{Z})/\Gamma_0(N)$.

Then we can write

$$\mathcal{M}_{n,N}(\mathbb{Z}) = \bigsqcup_{i=1}^{r} \Gamma_0(N)\alpha_i = \bigsqcup_{i=1}^{r} \beta_i\Gamma_0(N),$$

for some collections, α_i and β_j in $\mathcal{M}_{n,N}(\mathbb{Z})$. It is easy to see there must be a permutation $\pi : \{1, \dots, r\} \to \{1, \dots, r\}$ such that $\Gamma_0(N)\alpha_i \cap \beta_{\pi(i)}\Gamma_0(N) \neq \emptyset$ for all i (otherwise there is a proper subset of indices i such that $\Gamma_0(N)\alpha_i = \mathcal{M}_{n,N}(\mathbb{Z})$).

Take $\mu_i \in \Gamma_0(N)\alpha_i \cap \beta_{\pi(i)}\Gamma_0(N)$ for $i = 1, \dots, r$. Then $\{\mu_i\}$ forms a complete set of representatives both for $\Gamma_0(N)\backslash\mathcal{M}_{n,N}(\mathbb{Z})$ and for $\mathcal{M}_{n,N}(\mathbb{Z})/\Gamma_0(N)$. The latter fact composed with the map $\mu \mapsto n\mu^{-1}$ shows that $\{n\mu_i^{-1}\}$ is also a complete set of representatives for $\Gamma_0(N)\backslash\mathcal{M}_{n,N}(\mathbb{Z})$. □

Note: even for $n = p$, the representatives given in Lemma 6.1.3, do not satisfy the conditions given in the above lemma. At least some places in the literature are sloppy about this, and as a consequence have a gap or error in the proof of the following theorem below (this includes an earlier version of these notes, as I originally copied this mistake from elsewhere, and I thank Roberto Miatello for pointing this out to me—the approach of using the above lemma can be found in [Miy06]).

Theorem 6.2.7. *The Hecke operators $\{T_n\}$ on $S_k(N)$ are hermitian with respect to the Petersson inner product, i.e.,*

$$\langle T_n f, g\rangle = \langle f, T_n g\rangle$$

for $f, g \in S_k(N)$.

Proof. Fix a fundamental domain $\mathcal{F}$ for $\Gamma_0(N)$. For simplicity of notation, we extend $\langle f, g\rangle$ by (6.2.4) to any functions on $\mathfrak{H}$ for which the integral converges.

First we claim that for any $\tau = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PSL}_2(\mathbb{R})$, we have

$$\langle f|_{\tau,k}\,,\, g\rangle = \det(\tau)^{-k}\langle f,\, g|_{\tau^{-1},k}\rangle,$$

where we extend $|_{\tau,k}$ to $\tau \in \mathrm{GL}_2(\mathbb{R})^+ = \{\alpha \in \mathrm{GL}_2(\mathbb{R}) : \det(\alpha) > 0\}$ by (4.3.1). (The condition of positive determinant ensures τ maps $\mathfrak{H}$ to $\mathfrak{H}$. We remark that if we had followed standard notation and included a factor of $\det(\tau)^{k/2}$ in our definition of the slash operator, the determinant factor would not appear in the above identity.)

This follows since

$$y^k f|_{\tau,k}(z)\,\overline{g(z)} = \frac{\mathrm{Im}(z)^k}{(cz+d)^k} f(\tau z)\overline{g(z)} = \det(\tau)^{-k}\mathrm{Im}(w)^k\overline{(cz+d)}^k f(w)\overline{g(\tau^{-1}w)}.$$

where we put $w = \tau z$ and used the identity

$$|j(\tau, z)|^2 = \det(\tau)\frac{\mathrm{Im}(z)}{\mathrm{Im}(\tau z)}, \quad \tau \in \mathrm{GL}_2(\mathbb{R})^+,$$

which is a simple generalization of (6.2.2). Then observing $(cz+d)^k = j(\tau, z)^k = j(\tau^{-1}, w)^{-k}$ shows this equals

$$\det(\tau)^{-k}\mathrm{Im}(w)^k f(w)\overline{g|_{\tau^{-1},k}(w)},$$

which implies our claim.

Recall

$$T_n f = n^{k-1}\sum_{\mu \in \Gamma_0(N)\backslash\mathcal{M}_{n,N}(\mathbb{Z})} f|_{\mu,k} = n^{k-1}\sum_{i=1}^{r} f|_{\mu_i,k},$$

where $\mu_1, \ldots, \mu_r$ is a set of representatives for $\Gamma_0(N)\backslash\mathcal{M}_{n,N}(\mathbb{Z})$, which we take to be as in Lemma 6.2.6. Now

$$\langle T_n f, g\rangle = n^{k-1}\sum\langle f|_{\mu_i,k}, g\rangle = n^{-1}\sum\langle f, g|_{\mu_i^{-1},k}\rangle$$

Since $g|_{n\tau,k} = n^{-k}g|_{\tau,k}$, this implies

$$\langle T_n f, g\rangle = n^{k-1}\sum\langle f, g|_{n\mu_i^{-1},k}\rangle = \langle f, T_n g\rangle,$$

by our choice of μ_i. □

Corollary 6.2.8. *The space $S_k(N)$ has a basis consisting of Hecke eigenforms.*

Proof. Since $\{T_n\}$ is a family of commuting operators which are hermitian with respect to $\langle \cdot, \cdot \rangle$, a well known theorem in linear algebra tells us that an orthonormal basis with respect to $\langle \cdot, \cdot \rangle$ is a basis of eigenvectors for all $\{T_n\}$. □

By Exercise 6.1.9, we know the normalized ($a_1 = 1$) Hecke eigenforms in $S_k(N)$ have Fourier coefficients a_n which are multiplicative for n relatively prime to the level N. Hence the cusp forms are generated by forms whose Fourier coefficients are some kind of arithmetic sequences.

There are some more questions we can ask at this point to push our theory further.

- Does the whole space of modular forms $M_k(N)$ actually have a basis of Hecke eigenforms?
- Can we define the Hecke operators T_n on $M_k(N)$ when $\gcd(n, N) > 1$?
- Can we actually find a basis for $S_k(N)$ (or $M_k(N)$) whose Fourier coefficients a_n are multiplicative for all n?

Here we will just briefly discuss these questions.

First, observe that Theorem 6.2.7 is actually valid for $f, g \in M_k(N)$ when $\langle T_n f, g \rangle$ converges. One can use this to deduce that $M_k(N)$ does have a basis of eigenforms, and in fact we already know it at least one case.

Example 6.2.9. *The space $M_k(1)$ has a basis of Hecke eigenforms. To see this, recall $M_k(1) = \mathbb{C}E_k \oplus S_k(1)$. By Exercise 6.0.1, we know E_k is an eigenform, hence the basis of eigenforms for $S_k(1)$ can be extended to a basis of eigenforms for $M_k(1)$.*

In fact, the Eisenstein series (for $M_k(N)$, or more generally for $M_k(\Gamma)$) are orthogonal to the space of cusp forms in the sense that $\langle E, f \rangle = 0$ whenever E is an Eisenstein series and f is a cusp form. Often one uses this relation to *define* Eisenstein series: the space of Eisenstein series for $M_k(\Gamma)$ will be the elements $E \in M_k(\Gamma)$ such that $\langle E, f \rangle = 0$ for all $f \in S_k(\Gamma)$.

Next, one can define Hecke operators T_n on $M_k(N)$ when $\gcd(n, N) > 1$, however a definition in terms of $\mathcal{M}_{n,N}(\mathbb{Z})$ is somewhat trickier as the analogue Lemma 6.1.3 is not as nice. For simplicity, let's just talk about T_p where $p|N$, since all the T_n's can be constructed out of just the T_p's. One can define

$$T_p f = \sum_{0 \le j < p} f| \begin{pmatrix} 1 & j \\ 0 & p \end{pmatrix}_{,k},$$

and if $f(z) = \sum a_n q^n$, one obtains

$$T_p f(z) = \sum a_{pn} q^n.$$

The definition of Hecke operators can be made more uniform (including for arbitrary congruence subgroups Γ) by working in terms of double cosets (see, e.g., [DS05] or [Kob93]).

While the Hecke operators T_p for $p|N$ can be defined without much trouble, one does not in general have a basis of eigenforms for $S_k(N)$ for all T_n. Roughly the issue is the following.

Say $N = pq$ with p, q distinct primes and $f \in S_k(q)$ is a Hecke eigenform. In particular, f is an eigenform for T_p (on $S_k(q)$) Then automatically $f \in S_k(N)$. However T_p on $S_k(N)$ acts differently, so f is no longer an eigenform for T_p.

More generally, if $d|N$ and $f(z) \in S_k(N/d)$, then $f(z), f(dz) \in S_k(N)$ and we the subspace generated by all such elements the space $S_k^{\text{old}}(N)$ of *oldforms*. Define the space of *newforms* $S_k^{\text{new}}(N)$ to be its orthogonal complement in $S_k(N)$ (w.r.t. the Petersson inner product). Then it is true that $S_k^{\text{new}}(N)$ has a basis of eigenforms for *all* T_p, as opposed to just for $p \nmid N$. However, it can happen that oldforms are also eigenform for all T_p, or that oldforms have multiplicative Fourier coefficients but are not eigenforms for all T_p. (Conversely, if f is an eigenform for all T_p, the Fourier coefficients may not be multiplicative, but they essentially are—they are if $a_1 = 1$.)

The notion of newforms versus oldforms leads to another interesting question:

- Given some form $f \in S_k(N)$, how can we determine if it comes from a form of smaller level or if it is a newform?

We will study all of these issues in Chapter 8.

Finally we remark that the theory of Hecke operators for an arbitrary congruence subgroup Γ is similar. See [Ste07] for how to computationally find a basis of Hecke eigenforms for a given space $M_k(\Gamma)$.

Chapter 7

L-functions

One of the main themes in modern number theory, is to associate to various objects (number fields, Dirichlet characters, Galois representations, elliptic curves, modular forms) an analytic gadget called an L-function. The idea comes from the theory of the Riemann zeta function, so before we introduce L-functions for modular forms, we recall some basic facts about the zeta function and Dirichlet L-functions.

NOTE: This chapter is readable, but not as complete as I would like. Specifically, I didn't have enough time in my mind to say why you should care about L-functions of modular forms, e.g., the relation with elliptic curves and Fermat's last theorem. I hope to someday flesh out this chapter more.

7.1 Degree 1 L-functions

Let's begin by reviewing the zeta function. Then we will review some facts about Dirichlet L-functions. Both of these are considered "degree 1" L-functions, which will be explained later.

7.1.1 The Riemann zeta function

First recall the **Riemann zeta function** is defined by

$$\zeta(s) = \sum \frac{1}{n^s} \tag{7.1.1}$$

for $s \in \mathbb{C}$ with $\mathrm{Re}(s) > 1$. Then $\zeta(s)$ can be (uniquely) meromorphically continued to a function of the whole complex plane with only a simple pole at $s = 1$. Euler (who considered $\zeta(s)$ for s real) observed

$$\zeta(s) = \sum \frac{1}{n^s} = \prod_p \frac{1}{1 - p^{-s}}. \tag{7.1.2}$$

Since each factor on the right converges at $s = 1$, the fact that $\zeta(s)$ has a pole at $s = 1$ implies there are infinitely many prime numbers. This is the first hint that the analytic behaviour of $\zeta(s)$ can encodes deep arithmetic information.

One important feature of $\zeta(s)$ is its *functional equation* relating the values at s to the values at $1-s$. Namely, recall the **gamma function** is defined by

$$\Gamma(s) = \int_0^\infty t^{s-1}e^{-s}dt \tag{7.1.3}$$

for $\text{Re}(s) > 0$. Then $\Gamma(s)$ has meromorphic continuation to $\mathbb{C}$ with simple poles at $s = 0, -1, -2, \ldots$. Further Γ satisfies the functional equation

$$\Gamma(s+1) = s\Gamma(s)$$

and one easily computes from the definition that $\Gamma(1) = 1$. Hence, by induction, the functional equation implies $\Gamma(n+1) = n!$ for $n \in \mathbb{N}$.

Then the **functional equation** for $\zeta(s)$ can be written

$$\zeta(s) = 2^s\pi^{s-1}\sin\left(\frac{\pi s}{2}\right)\Gamma(1-s)\zeta(1-s).$$

One can rewrite this functional equation more simply in terms of the **completed zeta function**

$$Z(s) = \frac{s(s-1)}{2\pi^{s/2}}\Gamma\left(\frac{s}{2}\right)\zeta(s)$$

as

$$Z(s) = Z(1-s).$$

Either way, the functional equation means one can obtain the values for $\zeta(s)$ from the values of $\zeta(1-s)$. In particular, if $\text{Re}(s) < 0$, then $\text{Re}(1-s) > 1$ so one can also use (7.1.1) to evaluate $\zeta(s)$ on the left half-plane $\text{Re}(s) < 0$. The in-between region, $0 \le \text{Re}(s) \le 1$ is called the **critical strip** and line of symmetry $\text{Re}(s) = \frac{1}{2}$ is called the **critical line**.

As is now famous, the zeros of the Riemann zeta function are intimately connected with prime numbers. Specifically, let

$$\psi(x) = \sum_{p^k \le x} \log p$$

be the Chebyshev function, which counts the number of prime powers up to x. Then one can show the astounding **explicit formula**

$$\psi(x) = x - \sum_\rho \frac{x^\rho}{\rho} - \log 2\pi, \tag{7.1.4}$$

where ρ runs over the zeroes of $\zeta(s)$ (including the "trivial zeroes" to the left of the the critical strip). Consequently, if we knew where all the zeroes of $\zeta(s)$ lay, we would know exactly where all the prime powers are. Regardless, one can already use (7.1.4) to prove the prime number theorem

$$\pi(x) := \#\{p : p \le x\} \sim \frac{\psi(x)}{\log x} \sim \frac{x}{\log x}.$$

The very deep *Riemann hypothesis* asserts that all nontrivial zeroes (all zeroes inside the critical strip) actually lie on the critical line, and this would give an optimal bound on the error term in the prime number theorem.

In another direction, *special values* of $\zeta(s)$ tell us interesting arithmetic information. For instance, Euler showed that for $n \in \mathbb{N}$,

$$\zeta(2n) = (-1)^{n+1} \frac{B_{2n} \cdot (2\pi)^{2n}}{2(2n)!}. \tag{7.1.5}$$

One can interpret $1/\zeta(2n)$ as the probability that $2n$ integers chosen at random are relatively prime (this is a simple consequence of (7.1.2)). Based on this expression, one might ask, when n is odd, if $\zeta(n) = c\pi^n$ for some $c \in \mathbb{Q}$? While this is not known (even for $\zeta(3)$), it is expected the answer is no.

One heuristic reason why values at the even and odd integers should be different comes by looking the functional equation. Namely, the expression (7.1.5) simplifies to

$$\zeta(1-2n) = -\frac{B_{2n}}{2n}.$$

Thus it appears the arithmetic of the special values of $\zeta(2n)$ is made more clear by looking at $\zeta(1-2n)$. On the other hand by the function equation and that $\Gamma(s)$ has poles at negative integers, in order for $\zeta(s)$ to be holomorphic we need $\zeta(-2n) = 0$. (These are the trivial zeroes—note the functional equation does not produce zeroes at negative odd integers because then the vanishing of $\sin\left(\frac{\pi s}{2}\right)$ cancels the pole from $\Gamma(1-s)$.) Since $\zeta(2n+1)$ corresponds to $\zeta(-2n) = 0$, there is no indication that $\zeta(s)$ at positive odd integers should follow the same sort of arithmetic behaviour as at positive even integers. (In light of the probabilistic interpretation of $1/\zeta(n)$, this is perhaps analogous to the difference between $r_k(n)$, i.e., $\vartheta^k(z)$, for k odd and even.)

7.1.2 Dirichlet L-functions

Definition 7.1.1. *We say $\chi : \mathbb{Z} \to \mathbb{C}$ is a* **Dirichlet character mod** N *if*
(i) $\chi(n) = \chi(n+N)$ for all $n \in \mathbb{Z}$;
(ii) $\chi(n) = 0 \iff \gcd(n, N) > 1$;
(iii) $\chi(1) = 1$; and
(iv) $\chi(mn) = \chi(m)\chi(n)$ for all $m, n \in \mathbb{Z}$.

The standard way to construct a Dirichlet character is to take a character χ of $(\mathbb{Z}/N\mathbb{Z})^\times$, extend it to $\mathbb{Z}/N\mathbb{Z}$ by setting it to 0 on all the noninvertible residue classes, then lifting it to a function of $\mathbb{Z}$. It is easy to see this satisfies the above definition, and conversely any Dirichlet character arises in such a way.

Then one defines the **Dirichlet L-function** for a Dirichlet character χ to be

$$L(s, \chi) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s}, \tag{7.1.6}$$

for $\mathrm{Re}(s) > 1$. (Condition (i) in the definition implies χ is bounded, so the series converges absolutely for $\mathrm{Re}(s) > 1$.) Since χ is multiplicative, again one an **Euler product** expansion

$$L(s, \chi) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s} = \prod_p \frac{1}{1 - \chi(p)p^{-s}}. \tag{7.1.7}$$

As with $\zeta(s)$, one can show $L(s,\chi)$ has meromorphic continuation to $\mathbb{C}$, but this time it will in fact be entire whenver χ is a nontrivial character (i.e., does not come from the trivial character of $(\mathbb{Z}/N\mathbb{Z})^\times$).

Because the Dirichlet characters mod N allow one to distinguish among the different residue classes mod N, these Dirichlet L-functions provide a way to study primes lying in arithmetic progressions. Specifically, Dirichlet showed that $L(1,\chi) \neq 0$ for each nontrivial χ, and used this to prove that for any b relatively prime to N, there are infinitely many primes of the form $aN+b$, $a,b \in \mathbb{N}$.

Again in analogy with the Riemann zeta function, Dirichlet L-functions have functional equations. If χ is a "primitive" Dirichlet character mod N, the completed Dirichlet L-function is

$$\Lambda(s,\chi) = \left(\frac{N}{\pi}\right)^{\frac{s+\epsilon}{2}} \Gamma\left(\frac{s+\epsilon}{2}\right) L(s,\chi) \tag{7.1.8}$$

where $\epsilon \in \{0,1\}$ is the order of $\chi(-1)$, i.e., $\epsilon = 0$ if $\chi(-1) = 1$ and $\epsilon = 1$ if $\chi(-1) = -1$. Then the functional equation reads

$$\Lambda(s,\chi) = (-i)^\epsilon \sqrt{N} \left(\sum_{n=1}^{N} \chi(n) e^{2\pi i n/N}\right) \Lambda(1-s,\overline{\chi}). \tag{7.1.9}$$

Note here the functional equation does not relate $L(s,\chi)$ with $L(1-s,\chi)$ in general, but rather with $L(1-s,\overline{\chi})$, where $\overline{\chi}$ is the complex conjugate of χ (also a Dirichlet character mod N). However if χ is real valued, which is equivalent to χ^2 is trivial, then $\overline{\chi} = \chi$ and this functional equation relates $L(s,\chi)$ with $L(1-s,\chi)$.

In any case, one can still recover the values of $L(s,\chi)$ for $\mathrm{Re}(s) < 0$ using the function equation (7.1.9) together with the series for $L(1-s,\overline{\chi})$. The *generalized Riemann hypothesis (GRH)* asserts that any zeroes of $L(s,\chi)$ inside the critical strip $0 \leq \mathrm{Re}(s) \leq 1$ in fact lie on the critical line $\mathrm{Re}(s) = \frac{1}{2}$. As one might expect from analogy with the Riemann hypothesis, GRH would tell us, up to a very precise error, how many primes $\leq x$ lie in a given arithmetic progression. (An asymptotic is already known by the Chebotarev density theorem.) What is perhaps more surprising, is that GRH also implies the *weak Goldbach conjecture* that every odd number > 7 is a sum of 3 odd primes.[1]

Again, just like with $\zeta(s)$, there are formulas for special values of $L(s,\chi)$. Let d be squarefree, and consider the quadratic extension $K = \mathbb{Q}(\sqrt{d})$ of $\mathbb{Q}$ of discriminant Δ (here $\Delta = 4d$ if $d \equiv 2,3 \bmod 4$ and $\Delta = d$ if $d \equiv 1 \bmod 4$). and the Dirichlet character mod Δ given by

$$\chi_\Delta(n) = \left(\frac{\Delta}{n}\right),$$

where $\left(\frac{\Delta}{n}\right)$ is the Kronecker symbol. In particular, if p is an odd prime not dividing Δ, $\chi_\Delta(p) = 1$ if Δ is a square mod p and $\chi_\Delta(p) = -1$ otherwise.

If $d < 0$, i.e., K is imaginary quadratic, the *Dirichlet class number formula* says

$$L(1,\chi_\Delta) = \frac{2\pi h_K}{w\sqrt{|\Delta|}},$$

[1]Since writing these notes, the weak Goldbach conjecture was proved by my buddy Harald Helfgott! Thanks, Harald!

where h_K is the class number of K and w is the number of roots of unity in K, i.e., $w = 6$ if $d = -3$, $w = 4$ if $d = -1$ and $w = 2$ otherwise. If $d > 0$, i.e., K is a real quadratic field, then Dirichlet's class number formula says

$$L(1, \chi_\Delta) = \frac{2 \log \eta h_K}{\sqrt{\Delta}},$$

where $\eta > 1$ is the fundamental unit in the ring of integers of K. Since the class number is a fundamental, yet mysterious, invariant of a number field, we see that special values (in this case at $s = 1$) of L-functions encode interesting arithmetic information.

7.2 The philosophy of L-functions

In general, what is an L-function? There is no widely accepted answer yet as to what precisely constitutes an L-function, or what objects give rise to L-functions. However based on examples, there are certain properties L-functions should satisfy, similar to $\zeta(s)$ and the Dirichlet L-functions.

Suppose one has an object X described by some data $a_1, a_2, a_3, \ldots$, one can study the sequence (a_n) forming the (formal) *L-series* (or *Dirichlet series*) for X

$$L(s, X) = \sum_{n=1}^{\infty} \frac{a_n}{n^s}.$$

Exercise 7.2.1. *Suppose $|a_n| = O(n^m)$ for some m, i.e. there exists a constant C such that $|a_n| \leq Cn^m$ for all n. Show $L(s, X)$ converges absolutely for $\mathrm{Re}(s) > 1 + m$.*

No matter what the original object X was, the L-series only depends on the sequence (a_n), so for the present purposes we could just assume the object X is the sequence (a_n), and so we will sometimes write $X = (a_n)$ below.

An L-series is just a series of the form $\sum \frac{a_n}{n^s}$, but just requiring convergence on a right half plane does not give all the properties one would like to have an L-function. Based on the theory of Dirichlet L-functions and $\zeta(s)$, in order for $L(s, X)$ to be called an *L-function*, one might also ask that

- $L(s, X)$ has meromorphic continuation to $\mathbb{C}$;
- $L(s, X)$ has a functional equation $L(s, X) = \sigma(s, X)L(k - s, \check{X})$ for some "simple" function $\sigma(s, X)$, some $k \in \mathbb{R}$ and some related object $\check{X}$;
- $L(s, X)$ should have an *Euler product* $L(s, X) = \prod_p L_p(s, X)$, for some "simple" local factors $L_p(s, X)$.[2]

Example 7.2.2. *Let $X = K$ be a number field, i.e., a finite extension of $\mathbb{Q}$. Let $\mathcal{O}_K$ be its ring of integers, and a_n be the number of integral ideals of norm n. In particular, if $K = \mathbb{Q}$, then $a_n = 1$ for all n so $L(s, \mathbb{Q}) = \zeta(s)$. More generally, $L(s, K) = \zeta_K(s)$ is the* **Dedekind zeta function** *for K. The Dedekind zeta function also satisfies all properties listed above.*

[2]One may not be able to make $L(s, X)$ have an Euler product for all X in the space of interest—e.g., modular forms—but only certain "nice" X—e.g., Hecke eigenforms.

Example 7.2.3. *Let χ be a Dirichlet character and $a_n = \chi(n)$. Then $L(s, \chi)$ is the Dirichlet L-function we defined earlier.*

Let's elaborate on the Euler product condition a little more. How did we get the Euler product for Dirichlet L-functions? It came from the fact that $\chi(n)$ is multiplicative—in fact totally multiplicative. Recall a sequence (a_n) is **multiplicative** (resp. **totally multiplicative**) if $a_{mn} = a_m a_n$ for $\gcd(m, n) = 1$ (resp. for all m, n).

Exercise 7.2.4. *Let $X = (a_n)$ be a multiplicative sequence such that $|a_n| = O(n^m)$, and define the* **local L-factor**

$$L_p(s, X) = 1 + \frac{a_p}{p^s} + \frac{a_{p^2}}{p^{2s}} + \frac{a_{p^3}}{p^{3s}} + \cdots$$

Show $L_p(s, X)$ converges absolutely for $\mathrm{Re}(s) > 1 + m$, and that on this half-plane, we have the product formula

$$L(s, X) = \sum_{n=1}^{\infty} \frac{a_n}{n^s} = \prod_p L_p(s, X).$$

A priori, there is no reason for $L_p(s, X)$ to have a simple expression. In the case of Dirichlet L-functions, the local L-factor $L_p(s, \chi) = \frac{1}{1-\chi(p)p^{-s}}$ simply because $\chi(p^j) = \chi(p)^j$. More generally, we need, for each p, a relation among the a_{p^j}'s in order for $L_p(s, X)$ to simplify. Suppose, for a fixed p, the terms a_{p^j} satisfy a degree d linear recurrence relation:

$$a_{p^{j+d}} = c_0 a_{p^j} + c_1 a_{p^{j+1}} + \cdots + c_{d-1} a_{p^{j+d-1}}.$$

Then one can show the local L-factor is of the form

$$L_p(s, X) = \frac{1}{F(p^{-s})}$$

where F is a polynomial (depending on $c_0, \ldots, c_{d-1}$, which depend upon p) of degree $\leq d$. For instance, for a Dirichlet character χ and a prime p, then the polynomial $F(t) = 1 - \chi(p)t$. For $L_p(s, X)$ of this form, call the degree of F the degree of $L_p(s, X)$.

If $X = f$ is a Hecke eigenform and the a_n's are its Fourier coefficients, then the theory of Hecke operators tells us the Fourier coefficients a_{p^j} satisfy a degree 2 linear recursion, so the local L-factors should be reciprocals of (typically) quadratic polynomials in p^{-s}. (For a finite number "bad" primes, the L-factor may be trivial or (a reciprocal of a) linear (polynomial).)

Typically, the "degree" of almost all (all but finitely many) local L-factors, $L_p(s, X)$, will be the same, and we will call this the number the **degree** of the L-function $L(s, X)$. This explains why the Riemann zeta function and Dirichlet L-functions are called degree 1 L-functions, and the L-functions of modular forms are called degree 2 L-functions.

The properties above are some properties we want L-functions to satisfy, but are probably not strong enough to say any function with these properties should be the L-function of some object. In the next section, we will describe some conditions that are *sufficient* to ensure an L-series is the L-function of a modular form.

Just as with Dirichlet L-functions, when one can construct an L-function $L(s, X)$ for some object X, one should expect the following.

- The analytic properties—namely the location of the zeroes and poles—of $L(s, X)$ should reveal deep information about X.
- For certain special points $s = s_0$, there should be a meaningful formula for *special value* $L(s_0, X)$.

7.3 L-functions for modular forms

As was suggested in the previous section, if $f(z) = \sum_{n=0}^{\infty} a_n q^n$ is a modular form, we will define its L-series by $L(s, f) = \sum_{n=1}^{\infty} \frac{a_n}{n^s}$. Obviously replacing f with a multiple cf replaces $L(s, f)$ with $cL(s, f)$, so one often assumes f is normalized so its first nonzero Fourier coefficient is 1.

For convergence of $L(s, f)$, we need the following simple estimate on Fourier coefficients of cusp forms, due to Hardy and Hecke.

Proposition 7.3.1. *(The Hecke, or trivial, bound) Let $f(z) = \sum a_n q^n \in S_k(N)$. Then, $|a_n| = O(n^{k/2})$, i.e., for some constant C, we have*

$$|a_n| \leq C n^{k/2}. \tag{7.3.1}$$

Proof. From our discussion at the beginning of Section 6.2, we know $f(z)\overline{f(z)}y^k = |f(z)y^{k/2}|^2$ is $\Gamma_0(N)$ invariant. If $\mathcal{F}$ is a fundamental domain for $\Gamma_0(N)$, f being a cusp form means $f(z) \to 0$ as $z \in \mathcal{F}$ tends to a boundary point in $\mathbb{Q} \cup \{i\infty\}$. Consequently, as in the proof of Proposition 6.2.5, we see $|f(z)y^{k/2}|$ is bounded on $\mathcal{F}$, and by invariance, bounded on $\mathfrak{H}$. Say $|f(z)y^{k/2}| \leq C_1$ for all $z \in \mathfrak{H}$.

Then for fixed $y > 0$,

$$a_n e^{-2\pi n y} = \left| \int_0^1 f(x + iy) e^{-2\pi i n z} e^{-2\pi n y} dx \right| = \left| \int_0^1 f(x + iy) e^{-2\pi i n x} dx \right| \leq C_1 y^{-k/2}.$$

Setting $y = \frac{1}{n}$ gives the desired bound. □

The Ramanujan conjecture (proved by Deligne) is that one actually has $|a_n| \leq C n^{(k-1)/2}$, for some C (depending on f but not n). This is very deep and it is the best general bound possible. (In analytic number theory, knowing the exponents in such bounds is crucial for applications, and even a small improvement in the exponent is often considered major progress.) We note there is also a more explicit version of Deligne's bound which says $|a_p| \leq 2p^{(k-1)/2}$ and therefore $|a_n| \leq d(n) n^{(k-1)/2}$, where $d(n)$ is the number of divisors of n, for suitable cusp forms f (namely f should be a normalized eigennewform, as defined in the next chapter—for now just think an eigenform with $a_1 = 1$, which is all that is actually meant in the case of full level $N = 1$.).

Compare these bounds with those for Eisenstein series, which are simple to obtain. We explained this earlier, but it may be good to go over it on your own now.

Exercise 7.3.2. *Write $E_k(z) = \sum a_n q^n$. Show $|a_n| = O(n^{k-1})$ and that this exponent is sharp, i.e., $|a_n| \neq O(n^{k-1-\epsilon})$ for any $\epsilon > 0$.*

Example 7.3.3. *Recall from Example 5.2.10, we have*

$$r_{12}(n) = 8\sigma_5(n) - 512\sigma_5(n/4) + 16a_n,$$

where the a_n's are the Fourier coefficients of the cusp form F_η. Hence Hecke's bound tells us

$$r_{12}(n) = 8\sigma_5(n) - 512\sigma_5(n/4) + O(n^3),$$

or more simply for $n \not\equiv 0 \bmod 4$,

$$r_{12}(n) = 8\sigma_5(n) + O(n^3).$$

As in the above exercise, we see that $\sigma_5(n)$ roughly grows at the rate of n^5.

The Ramanujan conjecture says in fact the bound on the error is $O(n^{5/2})$.

Definition 7.3.4. *Let $f(z) = \sum a_n q^n \in M_k(N)$. The* **(Hecke) L-function** *associated to f is given by*

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

where this sum converges. The **completed L-function** *associated to f is given by*

$$\Lambda(s, f) = \left(\frac{\sqrt{N}}{2\pi}\right)^s \Gamma(s) L(s, f).$$

By Exercises 7.2.1 and 7.3.2, we see if $f = E_k \in M_k(1)$ then $L(s, f)$ converges for $\mathrm{Re}(s) > k$. In fact this is true for any $f \in M_k(N)$. If $f \in S_k(N)$, then by Hecke's bound we see $L(s, f)$ converges for $\mathrm{Re}(s) > 1 + k/2$. In fact, for cusp forms the Ramanujan conjecture tells us $L(s, f)$ converges in the slightly larger right half-plane $\mathrm{Re}(s) > 1 + (k-1)/2$.

We remark that unless f is an eigenform, $L(s, f)$ will not have an Euler product in general, but we will still call these L-functions as they at least have meromorphic continuation and functional equation.

Lemma 7.3.5. *Let $f \in M_k(N)$ and consider the* **Fricke involution**

$$\check{f}(z) := \frac{1}{N^{k/2} z^k} f\left(\frac{-1}{Nz}\right).$$

Then $\check{f} \in M_k(N)$. Moreover, if $f \in S_k(N)$, then so is $\check{f}$.

Proof. We can write

$$\check{f}(z) = \frac{(\det \omega)^{k/2}}{j(\omega, z)^k} f(\omega z) = (\det \omega)^{k/2} f|_\omega(z),$$

where

$$\omega = \begin{pmatrix} 0 & -1 \\ N & 0 \end{pmatrix}.$$

One sees

$$\omega \begin{pmatrix} a & b \\ c & d \end{pmatrix} \omega^{-1} = \begin{pmatrix} d & -c/N \\ -bN & a \end{pmatrix}$$

so ω normalizes $\Gamma_0(N)$. Then if $\gamma \in \Gamma_0(N)$, we compute

$$\check{f}|_\gamma = (\det \omega)^{k/2} f|_{\omega\gamma} = (\det \omega)^{k/2} f|_{\gamma'\omega} = (\det \omega)^{k/2} f|_\omega = \check{f},$$

where $\gamma' = \omega\gamma\omega^{-1} \in \Gamma_0(N)$. Hence $\check{f}$ satisfies the correct modular transformation law.

Holomorphy and holomorphy at the cusps follow automatically. Similarly if f is zero at the cusps it is easy to see $\check{f}$ is also. □

Note if $N = 1$, then $\check{f} = f$.

Theorem 7.3.6. *Let $f(z) = \sum a_n q^n \in M_k(N)$. The completed L-function $\Lambda(s, f)$ can be continued to a meromorphic function in s satisfying*

$$\Lambda(s, f) = (-1)^{k/2} \Lambda(k - s, \check{f}),$$

with at most simple poles at $s = 0$ and $s = k$. Furthermore, if $f \in S_k(N)$, then $\Lambda(s, f)$ is entire and bounded in vertical strips, i.e., $\Lambda(s, f)$ is bounded when $\mathrm{Re}(s)$ is.

Proof. (Sketch) The key to the proof comes from using an *integral representation* for $\Lambda(s, f)$. Namely, we look at the **Mellin transform** of f for $\mathrm{Re}(s) > k$,

$$\int_0^\infty f(iy) y^{s-1} dy = \int_0^\infty \sum a_n e^{-2\pi n y} y^{s-1} dy = \frac{\Gamma(s)}{(2\pi)^s} L(s, f) = N^{-s/2} \Lambda(s, f).$$

Write $\check{f}(z) = \sum b_n q^n$. Then one can show

$$\Lambda(s, f) + \frac{a_0}{s} + (-1)^{k/2} \frac{b_0}{k - s} = \int_1^\infty f(iy/\sqrt{N}) - a_0) y^{s-1} dy + \int_1^\infty (\check{f}(iy/\sqrt{N}) - b_0) y^{k-s-1} dy.$$

Since $f(iy/\sqrt{N}) - a_0$ and $(\check{f}(iy/\sqrt{N}) - b_0$ are of rapid decay when $y \to \infty$, the integrals on the right converge absolutely, analytically continue to entire functions of s, and are bounded on vertical strips. A similar expression for $\Lambda(s, \check{f})$ gives the functional equation. □

Theorem 7.3.7. *Suppose $f(z) = \sum a_n q^n \in S_k(1)$ be a Hecke eigenform. Then*

$$L(s, f) = \prod_p \frac{1}{1 - a_p p^{-s} + p^{k-1} p^{-2s}}.$$

Theorem 7.3.8. (Hecke's converse theorem) *Let (a_n) be a sequence such that $|a_n| = O(n^m)$ for some m, and $k \in 2\mathbb{N}$. Suppose*

$$\Lambda(s) = \frac{\Gamma(s)}{(2\pi)^s} \sum_{n=1}^\infty \frac{a_n}{n^s}$$

has meromorphic continuation to $\mathbb{C}$,

$$\Lambda(s) = (-1)^{k/2} \Lambda(k - s)$$

and

$$\Lambda(s) + \frac{a_0}{s} + (-1)^{k/2}\frac{a_0}{k-s}$$

is entire and bounded in vertical strips. Then

$$f(z) = \sum_{n=0}^{\infty} a_n q^n \in M_k(1).$$

This provides an analytic way to show something is a modular form! Roughly it says, given a sequence (a_n) of polynomial growth, the Dirichlet series $L(s) = \sum \frac{a_n}{n^s}$ is the L-function of a modular form $L(s) = L(s, f)$ if it satisfies certain nice properties (stated above in terms of $\Lambda(s)$). In other words, if the Dirichlet series is nice, then the sequence (a_n) is actually the Fourier coefficients for a modular form!

There is also a converse theorem for $M_k(N)$, due to Weil, but it requires more than just the niceness of a single function $L(s)$. It roughly says, given a sequence (a_n) of polynomial growth, the Dirichlet series $L(s) = \sum \frac{a_n}{n^s}$ is the L-function of a modular form in $M_k(N)$ if the set of twisted L-series

$$L_\chi(s) = \sum \frac{a_n \chi(n)}{n^s}$$

satisfy certain nice properties for all *Hecke characters* χ. (Hecke characters are a generalization of Dirichlet chacters.)

To put this in context, we will summarize the primary methods of constructing modular forms:

- Use an averaging procedure (e.g., as with our Eisenstein series and Hecke operators).
- Construct a suitable η-quotient.
- Use theta series.
- Use the converse theorem.
- Use the *trace formula* (which we will not describe).

Both the method of the converse theorem and the trace formula are not truly constructive, but they provide a way of showing something you suspect should be a modular form actually is a modular form.

The simplest instance of this is the following. Since the L-function $L(s, f)$ of a modular (eigen)form is a degree 2 Euler product, one might ask if you can construct some such f by asking that $L(s, f)$ is a product of 2 known degree 1 L-functions.

Example 7.3.9. *Put*

$$L(s) = \zeta(s)\zeta(s-k+1)$$

for $k \geq 4$ even. One can see by Hecke's converse theorem, that $L(s)$ is the L-function of a modular form of weight k. In fact

$$L(s) = L(s, E_k).$$

More generally, if χ_1 and χ_2 are primitive Dirichlet characters mod N_1 and N_2, then

$$L(s, \chi_1)L(s - k + 1, \chi_2) = L(s, f)$$

for some

$$f \in M_k(\Gamma_1(N_1N_2)).$$

If $\chi_1(-1) = \chi_2(-1)$, then in fact $f \in M_k(N_1N_2) = M_k(\Gamma_0(N_1N_2))$. Explicitly,

$$f = \sum_{n \geq 1} a_n q^n, \quad a_n = \sum_{ad=n} \chi_1(a)\chi_2(d)d^{k-1}.$$

From the Euler product for $L(s, f)$, one can deduce that f is a Hecke eigenform. However, this construction will not yield a cusp form (even though $a_0 = 0$, f does not vanish at all cusps). In fact, one can construct all Eisenstein series on $\Gamma_1(N)$ in this way by including degenerate cases (when χ_1 and χ_2 are trivial, so non-primitive if $N_1, N_2 > 1$, then $f = E_k$). Thus the space of Eisenstein series, and therefore $M_k(\Gamma_1(N))$, and by restriction $M_k(N)$, also has a basis consisting of eigenforms.

Part II

Selected topics

This part is still in progress, and currently only has a mostly finished chapter on new-forms, and the beginning of a chapter on Hilbert modular forms, and is also likely to contain errors. After finishing the chapter on Hilbert modular forms, I will probably work on a chapter on half-integral weight forms.

In any event, as remarked in the preface, it is not intended to contain complete proofs.

Chapter 8

Newforms and oldforms

Recall the main points of Chapter 6 were: (i) $M_k(N)$ has a basis of eigenforms for the Hecke operators T_p, $p \nmid N$ (or equivalently, T_n, $\gcd(n, N) = 1$); and (ii) if f is an eigenform, then its Fourier coefficients a_n are multiplicative for n relatively prime to N.

In this chapter we will try to resolve some of the issues brought up at the end of Chapter 6, principally:

- Can we find modular forms in $M_k(N)$ whose Fourier coefficients are multiplicative for all n?
- To what extent do such forms make up the space $M_k(N)$?
- Given a modular form $f \in M_k(N)$, how can we tell if it comes from a form of smaller level?

One application of the first question is that if $f \in M_k(N)$ has multiplicative Fourier coefficients, then the L-series $L(s, f)$ will have an Euler product.

To put the third question in context, recall that $M_k(d) \subset M_k(N)$ if $d|N$, i.e., any modular form of level N is also of level Nm for any $m \in \mathbb{N}$. Thus given a modular form f, it's natural to ask what its true level? (This is often called the *exact* level of f.) Forms that can be constructed (as linear combinations) of forms of smaller level are called *oldforms*, and the forms orthogonal to all of these are called *newforms*. Then the third question can be essentially be restated as: how can we distinguish newforms from oldforms?[1]

Because of the connection of multiplicativity of Fourier coefficients and being eigenforms of Hecke operators, it makes sense to try to answer at least the first two questions by extending our theory of Hecke operators T_n to all n. It will turn out that the space of newforms is spanned by forms which are simultaneous eigenforms for all T_n. This is not true for the oldforms, though some oldforms may still be eigenforms for all of the T_n. However, following Atkin and Lehner [AL70], we can pick out the new eigenforms by introducing some additional operators W_p (for $p|N$) which also commute with all of the T_n's (and therefore

[1] More precisely, these questions are equivalent when restricted to eigenforms, but not quite equivalent in general. For instance, say f has exact level 3 and g has exact level 5, so $f + g$ has level 15. It is not a newform because it is just a linear combination of forms of smaller level, though $f + g$ itself does not come from level 3 or level 5.

the space spanned by simultaneous eigenforms for all of these operators will be precisely the space of newforms).

Another important application of these Atkin–Lehner operators is that they will allow us to distinguish eigenforms. It's possible for two eigenforms (which are not scalar multiples of each other) to have the same Hecke eigenvalues for all T_n. Often when proving theorems about modular forms, you want to use Hecke operators to isolate individual eigenforms. This is possible sometimes, e.g., for $S_2(p)$ with p prime (all forms will be newforms in this case), but in general one has to use something like these Atkin–Lehner operators in conjunction with the Hecke operators.

Our first goal will be to define Hecke operators T_n on $M_k(N)$ for all n. As we remarked in Chapter 6, our earlier approach of using lattices to define Hecke operators leads to complications when $\gcd(n, N) > 1$, so we begin by explaining the double coset approach to Hecke operators.

8.1 Hecke operators via double cosets

Going back to our definition of Eisenstein series, we saw one of the basic ideas for making modular forms on a congruence subgroup Γ is to take the weighted average of a function over Γ. For instance, (4.2.2) told us that

$$E_{k,N}(z) = \sum_{\gamma \in P\backslash\Gamma_0(N)} j(\gamma, z)^{-k} = \sum_{\gamma \in P\backslash\Gamma_0(N)} 1|_{\gamma,k}(z),$$

where $P = \langle T \rangle$ is the unipotent subgroup of $\mathrm{PSL}_2(\mathbb{Z})$ and 1 denotes the constant function.

Now let's suppose we have two congruence subgroups Γ_1, Γ_2, and $f \in M_k(\Gamma_1)$. If we want to make a modular form for Γ_2 out of f, we can try to do it by averaging, i.e., a sum of the form $\sum f|_{\gamma,k}(z)$. Of course if we let γ run over all of Γ_2, this will not converge—for instance f is already invariant by $|_{\gamma,k}$ for $\gamma \in \Gamma_1 \cap \Gamma_2$, which is infinite. This is why we already had to mod out by P in the definition of $E_{k,N}$. In the present case, it makes sense to sum over $\gamma \in (\Gamma_1 \cap \Gamma_2)\backslash\Gamma_2$. Alternatively, we can sum over $\gamma \in \Gamma_1\backslash\Gamma_1\Gamma_2$. This latter form generalizes to the double coset operators.

For $\alpha \in \mathrm{GL}_2(\mathbb{Q})^+$, we define the associated **weight k double coset operator**

$$[\Gamma_1 \alpha \Gamma_2]_k f(z) = \sum_{\gamma \in \Gamma_1\backslash\Gamma_1\alpha\Gamma_2} \det \gamma^{k-1} f|_{\gamma,k}(z)$$

Lemma 8.1.1. *For $f \in M_k(\Gamma_1)$, $[\Gamma_1 \alpha \Gamma_2]_k f(z) \in M_k(\Gamma_2)$.*

Proof. See [DS05, Sec 5.1]. □

Depending on your terminology, this may or may not technically be an operator when $\Gamma_1 \neq \Gamma_2$ (it's not a map from a space to itself), but our interest is in the case where $\Gamma_1 = \Gamma_2 = \Gamma_0(N)$.

For any prime p, define the **p-th Hecke operator** on $M_k(N)$ to be

$$T_p f = [\Gamma_0(N) \begin{pmatrix} 1 & \\ & p \end{pmatrix} \Gamma_0(N)]_k.$$

When we compute our Hecke operators below, we will see that this agrees with the previous definition of T_p when $p \nmid N$.

One can similarly define more general Hecke operators T_n in terms of double cosets, however the T_p's are the crucial part, and for simplicity, we will avoid this and just define the T_n's using the T_p's and the Hecke operator relations we already found in Chapter 6.[2] We inductively define

$$T_{p^{r+1}} = \begin{cases} T_p T_{p^r} - p^{k-1} T_{p^{r-1}} & p \nmid N \\ T_p T_{p^r} & p|N, \end{cases}$$

and then extend by multiplicativity: set

$$T_{mn} = T_m T_n$$

when $\gcd(m,n) = 1$. The only part that's unmotivated is the case of $T_{p^{r+1}}$ when $p|N$ (so we simply have $T_{p^r} = (T_p)^r$ for $p|N$). Basically the difference arises from the difference in the double coset decompositions:

Lemma 8.1.2. *For $p \nmid N$,*

$$\Gamma_0(N) \begin{pmatrix} 1 & \\ & p \end{pmatrix} \Gamma_0(N) = \bigsqcup_{b=0}^{p-1} \Gamma_0(N) \begin{pmatrix} 1 & b \\ 0 & p \end{pmatrix} \bigsqcup \Gamma_0(N) \begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix}.$$

For $p|N$,

$$\Gamma_0(N) \begin{pmatrix} 1 & \\ & p \end{pmatrix} \Gamma_0(N) = \bigsqcup_{b=0}^{p-1} \Gamma_0(N) \begin{pmatrix} 1 & b \\ 0 & p \end{pmatrix}.$$

Proof. It is simple to derive this from Lemma 6.1.3. We remark [DS05, Sec 3.8] also proves an analogue for $\Gamma_1(N)$ in (3.16). □

We have the following generalization of Theorem 6.1.4.

Theorem 8.1.3. *Let $f(z) = \sum a_n q^n \in M_k(N)$. If $p \nmid N$, then*

$$(T_p f)(z) = \sum_{n \not\equiv 0 \bmod p} a_{pn} q^n + \sum_{n \equiv 0 \bmod p} \left(a_{pn} + p^{k-1} a_{n/p} \right) q^n.$$

If $p|N$, then

$$(T_p f)(z) = \sum_{n=0}^{\infty} a_{pn} q^n.$$

Proof. The proof is just a calculation similar to the proof of Theorem 6.1.4 using the double coset decomposition in the last lemma. See also [DS05, Prop 5.2.2(a)]. □

[2] For instance when $p|N$, T_{p^r} is given by the double coset operator for $\Gamma_0(N) \begin{pmatrix} 1 & \\ & p^r \end{pmatrix} \Gamma_0(N)$. See, e.g., [KL06, Sec 3.6] for the general definition of T_n in terms of double cosets.

In fact, one can generalize the operation of T_p when $p|N$ to get the U_p transformation (not assuming $p|N$)

$$U_p : M_k(N) \to M_k(pN)$$
$$(U_p f)(z) = \sum_{n=0}^{\infty} a_{pn} q^n.$$

(You may recall U_p from (6.0.2).) Namely, if $f \in M_k(N)$, then $f \in M_k(pN)$ by inclusion and $U_p f = T_p f$ (as an element of $M_k(pN)$). The fact that T_p operates on $M_k(pN)$ proves the image of U_p lands in $M_k(pN)$. We remark that some authors use U_p to denote the T_p operators in the case $p|N$, and reserve the notation T_p for the case $p \nmid N$. (E.g., Atkin and Lehner [AL70], who also use p only for primes prime to N, and q for primes dividing N, so you may hear people refer to T_p's and U_q's.)

Corollary 8.1.4. *If $f \in M_k(N)$ is an eigenform for every T_p, then f is an eigenform for every T_n.*

Proof. For $p|N$, since $T_{p^r} = (T_p)^r$ it is evident that if if f is an eigenform for T_p it is also an f is an eigenform for T_{p^r}. Now apply multiplicativity of the T_n's and Corollary 6.1.15. □

As for terminology, for $f \in M_k(N)$, if we just say f is an eigenform, then to be consistent with previous terminology this means f is an eigenform for all T_p with $p \nmid N$, i.e., for all T_n with $\gcd(n, N) = 1$. We will sometimes say f is a **complete eigenform** to mean f is an eigenform for all T_p, i.e., for all T_n. (This terminology is probably not standard.)

What it arithmetically means to be an eigenform for T_p with $p|N$ (or U_p) is easier to describe than for T_p with $p \nmid N$.

Corollary 8.1.5. *Suppose $f(z) = \sum a_n q^n \in M_k(N)$ and $p|N$.*

(i) Suppose f is an eigenform for T_p with eigenvalue λ_p. Then $a_{p^r} = \lambda_p^r a_1$. If $a_1 \neq 0$, then $\lambda_p = \frac{a_p}{a_1}$. If $a_0 \neq 0$, then $\lambda_p = 1$.

(ii) Suppose the a_n's are multiplicative and $f \neq 0$ so $a_1 = 1$. If $a_0 \neq 0$, then f is an eigenform for T_p if and only if $a_{p^r} = 1$ for all $r \geq 0$. If $a_0 = 0$, then f is an eigenform for T_p if and only if $a_{p^r} = (a_p)^r$ for all $r \geq 0$.

Proof. The theorem says that f is an eigenform for T_p if and only if there exists λ such that $a_{pn} = \lambda a_n$ for all n.

(i) The first two assertions follow from $a_{p^r} = \lambda_p a_{p^{r-1}}$ for $r \geq 1$. The latter from the fact that $a_{p \cdot 0} = a_0 = \lambda a_0$.

(ii) If f is an eigenform, the conditions on a_{p^r} follow from (i). Conversely, suppose $a_{p^r} = (a_p)^r$. Then for $n = p^r m$ with $p \nmid m$, we have $a_{pn} = a_{p^{r+1}m} = a_{p^{r+1}} a_m = a_p a_{p^r} a_m = a_p a_n$, i.e., f is an eigenform for T_p with eigenvalue a_p. □

Compare this with Exercise 6.1.6 and the following exercise.

Exercise 8.1.6. *Suppose $f(z) = \sum a_n q^n \in M_k(N)$ and $p \nmid N$.*

(i) Show that if $a_0 \neq 0$ and $T_p f = \lambda_p f$, then $\lambda_p = 1 + p^k$.

(ii) If the a_n's are multiplicative, determine a necessary and sufficient condition on the a_{p^r}'s for f to be an eigenform of T_p.

Recall that if a_n is a multiplicative sequence, it is determined by just the a_{p^r}'s. If the Fourier coefficients a_n of $f \in M_k(N)$ are multiplicative, being an eigenform for any T_p is equivalent to the a_{p^r}'s satisfying a certain recurrence relation, which depends on whether $p|N$ and on k for $p \nmid N$, and therefore that the a_n's are determined by just the a_p's.

The following result tells us conversely that being an eigenform for all T_n's (or equivalently by the first corollary, for all T_p's) implies the a_n's are multiplicative.

Proposition 8.1.7. *Let $f(z) = \sum a_n q^n \in M_k(N)$ be an eigenform for all T_n with eigenvalue λ_n. Then $a_n = \lambda_n a_1$. Hence, if f is a normalized (so $a_1 = 1$) eigenform for all T_n, then, the a_n's are multiplicative.*

This is why we defined the Hecke operators so that $T_{p^r} = (T_p)^r$ for $p|N$. From the above calculation of T_p, this is the right definition to make this proposition true.

Exercise 8.1.8. *Prove the above proposition.*

The following exercise explains the relation between "raising the level" of f at p and the Hecke operator T_p.

Exercise 8.1.9. *Suppose $f(z) = \sum a_n q^n \in M_k(N)$. The form $g(z) = f(pz)$ lies in $M_k(pN)$ and satisfies $T_p g = f$. Moreover, if f is an eigenform for T_n with $p \nmid n$, then so is g (on $M_k(pN)$) with the same eigenvalue.*

In particular, if $p|N$, then the map $f(z) \mapsto f(pz)$ gives a modular form $g(z) = \sum b_n q^n = \sum a_n q^{pn}$ obtained by "spreading out" the Fourier coefficients from $\mathbb{N} \cup \{0\}$ to $p\mathbb{N} \cup \{0\}$ and raising the level by p. The T_p operator reverses this process. Since T_p takes the linearly independent forms f and g to the 1-dimensional space generated by f (for $f \neq 0$), we see that the kernel of $T_p : M_k(pN) \to M_k(pN)$ is nontrivial, i.e., T_p is not an invertible linear transformation on $M_k(pN)$. Furthermore, if f is a complete eigenform on $M_k(N)$, $p|N$ and $T_p f = \lambda_p f$, then f and $f - \lambda_p g$ are complete eigenforms on $M_k(N)$ where $f - \lambda_p g$ will have T_p-eigenvalue 0.

8.2 Hecke operators on Eisenstein series

Here we investigate some details of Hecke operators on Eisenstein series. This will provide some concrete examples as well as motivation for the theory of newforms.

First we give an example where Fourier coefficients are multiplicative, but the form may not be an eigenform for all T_p.

Example 8.2.1. *For $k \geq 4$ be even, recall the renormalized Eisenstein series $E_k^*(z) = \sum a_n q^n \in M_k(1)$ from (6.0.1). It has multiplicative Fourier coefficients $a_n = \sigma_{k-1}(n)$ for $n \geq 1$, and is an eigenform of all T_p on $M_k(1)$. However, we can also regard $E_k^* \in M_k(N)$. On $M_k(N)$, E_k^* is not an eigenform for T_p with $p|N$ as $a_p = \sigma_{k-1}(p) \neq 1$ (Corollary 8.1.5(ii)).*

Now let's see what happens for our Hecke operators for Eisenstein series with level. For simplicity, we will focus on the case $k = 2$.

For $N = p$ prime, we can consider

$$E_{2,p}(z) = E_2(z) - pE_2(pz) \tag{8.2.1}$$

as in Exercise 4.2.18. This is holomorphic and $E_{2,p} \in M_2(p)$. From Exercise 4.2.18, you should have gotten the Fourier expansion

$$E_{2,p}(z) = 1 - p - 24\sum_{n=1}^{\infty} \sigma_{1,p}(n)q^n \tag{8.2.2}$$

where $\sigma_{1,p}(n) = \sigma_1(n) - p\sigma_1(n/p)$ and we interpret $\sigma_1(n/p)$ to be 0 if $p \nmid n$.

Exercise 8.2.2. *Check that $\sigma_{1,p}(p^r) = 1$, $\sigma_{1,p}(pn) = \sigma_{1,p}(n)$ if $p \nmid n$ and $\sigma_{1,p}$ is multiplicative. Conclude $\sigma_{1,p}(pn) = \sigma_{1,p}(n)$ for any n.*

This exercise tells us that, on $M_2(p)$, we have

$$T_p E_{2,p}(z) = 1 - p - 24\sum_{n=1}^{\infty} \sigma_{1,p}(pn)q^n = E_{2,p}(z)$$

and for $\ell \neq p$ prime

$$\begin{aligned} T_\ell E_{2,p}(z) &= (1+\ell)(1-p) - 24\sum_{n=1}^{\infty} \left(\sigma_{1,p}(\ell n) + \ell\sigma_{1,p}(n/\ell)\right) q^n \\ &= (1+\ell)E_{2,p}(z). \end{aligned}$$

For the last equality, you can write $n = \ell^r m$ with $\ell \nmid m$ and use the simple calculation $\sigma_{1,p}(\ell^{r+1}) + \ell\sigma_{1,p}(\ell^{r-1}) = (1+\ell)\sigma_{1,p}(\ell^r)$. Thus for all primes ℓ, $T_\ell E_{2,p} = \sigma_{1,p}(\ell)E_{2,p}$, and then by the above proposition, for all n,

$$T_n E_{2,p} = \sigma_{1,p}(n)E_{2,p}.$$

Hence $E_{2,p}$ is an eigenform for all Hecke operators on $M_2(p)$.

If we renormalize our Eisenstein series as

$$E^*_{2,p} = \frac{-1}{24}E_{2,p}(z) = \frac{p-1}{24} + \sum_{n=1}^{\infty} \sigma_{1,p}(n)q^n,$$

then we see $E^*_{2,p}$ has multiplicative Fourier coefficients.

Example 8.2.3. *Recall from Exercise 5.2.8 that $M_2(4)$ is 2-dimensional and generated by the Eisenstein series*

$$f(z) = E_{2,2}(z) = 1 + 24\sum_{n=1}^{\infty} \sigma_{1,2}(n)q^n = 1 + 24\left(q + q^2 + 4q^3 + q^4 + 6q^5 + \cdots\right)$$

and

$$g(z) = E_{2,2}(2z) = 1 + 24\sum_{n=1}^{\infty} \sigma_{1,2}(n)q^{2n} = 1 + 24\left(q^2 + q^4 + 4q^6 + q^8 + 6q^{10} + \cdots\right).$$

Since f is a complete eigenform on $M_2(2)$, Exercise 8.1.9 tells us that f and g are eigenforms on $M_2(4)$, but g is not a complete eigenform. (Moreover, since f and g have the same eigenvalues for T_n with $2 \nmid n$, any form in $M_2(4)$ is a (not necessarily complete) eigenform.) Moreover $T_2 g = f$ so $T_2(f-g) = T_2 f - f = f - f = 0$. Indeed, the Fourier expansion of $f - g$,

$$(f-g)(z) = 24\left(q + 4q^3 + 6q^5 + 8q^7 + \cdots\right)$$

contains no even powers of q. (We remark that even though the constant term of $f-g$ is 0, it is not a cusp form because it does not vanish at all other cusps.)

Hence f and $f-g$ give a basis of $M_2(4)$ of complete eigenforms. Moreover, we can normalize them to see $\frac{1}{24}f$ and $\frac{1}{24}(f-g)$ yield a basis of $M_2(4)$ consisting of forms with multiplicative Fourier coefficients.

To get a weight 2 Eisenstein series of a general level $N > 1$, you might consider $E_2(z) - NE_2(Nz)$. This is indeed in $M_2(N)$ and one can work out the Fourier expansion. However, it is not as nice as $E_{2,p}$ because it is not an eigenform for all T_p's (though it will be for $p \nmid N$), as the following exercise shows.

Exercise 8.2.4. *Suppose $N = p_1 p_2$, where p_1 and p_2 are two not necessarily distinct primes, and let $E(z) = E_2(z) - NE_2(Nz) = \sum a_n q^n$. Compute a_{p_1}, a_{p_2} and $a_{p_1 p_2}$. Show E is not an eigenform for T_{p_1} or T_{p_2} but is an eigenform for T_p with $p \nmid N$.*

Let us first define a weight 2 Eisenstein series for squarefree $N = p_1 \cdots p_r$. Set

$$E^*_{2,N}(z) = \frac{(-1)^{r+1}}{24} \sum_{d|N} \mu(\frac{N}{d}) d E_2(dz),$$

where μ is the Möbius function defined in Section 4.2 and r is the number of distinct prime divisors of N. One can check $E^*_{2,N}(z) \in M_0(N)$. Observe that if $N = p$ is prime, we have $E^*_{2,N}(z) = -\frac{1}{24}(-E_2(z) + pE_2(pz)) = E^*_{2,p}$, which coincides with our previous definition.

To compute the Fourier coefficients, it will be convenient to use some facts about Dirichlet convolution. Given two functions $f, g : \mathbb{N} \to \mathbb{C}$, we define their **Dirichlet convolution** (or **Dirichlet product**) by $f * g(n) = \sum_{d|n} f(n) g(\frac{n}{d})$.

Exercise 8.2.5. *(i) Show that if f and g are multiplicative, then $f * g$ is multiplicative.*

(ii) Show that $\mathrm{id} * \mu = \phi$, where $\mathrm{id} : \mathbb{N} \to \mathbb{N}$ is the identity map and ϕ is the Euler phi function.

We won't need this, but in case you're interested, you can check that Dirichlet convolution makes the space of functions $f : \mathbb{N} \to \mathbb{C}$ into a commutative ring.

Lemma 8.2.6. *For $N > 1$ squarefree, we have the Fourier expansion*

$$E^*_{2,N}(z) = (-1)^{r+1}\frac{\phi(N)}{24} + \sum_{n=1}^{\infty} \sigma_{1,N}(n) q^n,$$

where $\sigma_{1,N}$ is the multiplicative function given by $\sigma_{1,N}(n) = \sum_{d|\gcd(n,N)} \mu(d) d \sigma_1(\frac{n}{d})$.

Proof. The constant term of $E^*_{2,N}$ is $\frac{(-1)^r}{24}(\mathrm{id} * \mu)(N) = \frac{(-1)^r}{24}\phi(N)$. For $n \geq 1$, the q^n coefficient of $E^*_{2,N}$ is

$$\begin{aligned}(-1)^r \sum_{d|N} \mu(\frac{N}{d}) d\sigma_1(\frac{n}{d}) &= (-1)^r \sum_{d|\gcd(n,N)} \mu(\frac{N}{d}) d\sigma_1(\frac{n}{d}) \\ &= (-1)^r \mu(N) \sum_{d|\gcd(n,N)} \mu(d) d\sigma_1(\frac{n}{d}) \\ &= \sum_{d|\gcd(n,N)} \mu(d) d\sigma_1(\frac{n}{d}) = \sigma_{1,N}(n).\end{aligned}$$

It remains to show $\sigma_{1,N}(n)$ is multiplicative. Write $n = n_N m$ where $\gcd(n_N, N) = \gcd(n, N)$ and $\gcd(m, N) = 1$. Since σ_1 is multiplicative, we have $\sigma_{1,N}(n) = \sigma_1(m)\sigma_{1,N}(n_N)$. It remains to show $\sigma_{1,N}$ is multiplicative for powers of primes dividing N. Note

$$\sigma_{1,N}(n_N) = \sum_{d|n_N} \mu(d) d\sigma_1(\frac{n}{d}) = ((\mu \cdot \mathrm{id}) * \sigma_1)(n_N)$$

since μ is 0 on any p^e with $e > 1$. Since $\mu \cdot \mathrm{id}$ and σ_1 are multiplicative, so is their Dirichlet convolution by the exercise above, and we are done. □

Since $\sigma_{1,N}$ behaves like σ_1 for primes not dividing N, this means $E^*_{2,N}$ is an eigenfunction of each T_p with $p \nmid N$ on $M_2(N)$. (Alternatively, one can write $E_{2,N}$ as in terms of $E_{2,p}$'s for $p|N$ and apply Exercise 8.1.9.) And, for $p|N$, we see

$$\sigma_{1,N}(p^e) = \sum_{d|p^e} \mu(d) d\sigma_1(\frac{p^e}{d}) = \sigma_1(p^e) - p\sigma_1(p^{e-1}) = 1.$$

From the previous section, this means $E^*_{2,N}$ is a complete eigenfunction on $M_2(N)$.

What about when $N > 1$ is not squarefee? We can inductively define $E^*_{2,N}$ for cube-free $N > 1$ as follows. If $p|N$ but $p^2 \nmid N$, set

$$E^*_{2,pN}(z) = E^*_{2,N}(z) - E^*_{2,N}(pz)$$

Then by Exercise 8.1.9, we inductively see that $E^*_{2,pN}$ is an eigenfunction for all T_n with $p \nmid n$ and

$$T_p E^*_{2,pN}(z) = T_p E^*_{2,N}(z) - T_p E^*_{2,N}(pz) = E^*_{2,N}(z) - E^*_{2,N}(z) = 0.$$

Thus we have shown

Proposition 8.2.7. *For any cube-free $N > 1$, $E^*_{2,N}$ is a eigenform on $M_2(N)$, and a complete eigenform. Moreover, if $N = p^2M$ is cube-free, then $E^*_{2,pM}$ and $E^*_{2,N}$ are both complete eigenforms on $M_2(N)$, with $E^*_{2,pM}$ having T_p-eigenvalue 1 and $E^*_{2,N}$ having T_p-eigenvalue 0. These complete eigenforms are normalized so they have multiplicative Fourier coefficients.*

The issue with $p^3|N$ is as follows. For simplicity, assume $N = p^3$. Then we have

$$\begin{aligned} T_p E^*_{2,p}(z) &= E^*_{2,p}(z) \\ T_p E^*_{2,p}(pz) &= E^*_{2,p}(z) \\ T_p E^*_{2,p}(p^2 z) &= E^*_{2,p}(pz). \end{aligned}$$

Thus, taking these forms as a basis of this 3-dimensional subspace of $M_2(p^3)$, T_p acts as the matrix

$$T_p = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

This matrix has eigenvalue 1 with multiplicity 1 and eigenvalue 0 with multiplicity 2. But the eigenspace corresponding to eigenvalue 0 is 1-dimensional. Hence there is no basis of eigenforms for T_p for this 3-dimensional subspace of $M_2(p^3)$. This by itself does not automatically imply that $M_2(p^3)$ has no basis of complete eigenforms for general p, but we at least see it in the following example.

Example 8.2.8. *The space $M_2(8)$ is 3-dimensional, generated by $E^*_{2,2}(z)$, $E^*_{2,2}(2z)$ and $E^*_{2,2}(4z)$. By the above argument, it does not have a basis of complete eigenforms.*

One can similarly work out the action of the Hecke operators on Eisenstein series $E_{k,N}$ from Section 4.2 of higher weight. See, e.g., [DS05, Prop 5.2.3], which includes Eisenstein series coming from other cusps and gives criteria for these Eisenstein series to be complete eigenforms.

8.3 Atkin–Lehner operators

Let $p|N$, and say p^r is the highest power of p dividing N. The p-th **Atkin–Lehner operator** W_p on $M_k(N)$ is given by

$$W_p f = p^{kr/2} f| \begin{pmatrix} ap^r & b \\ cN & dp^r \end{pmatrix},$$

where $a, b, c, d \in \mathbb{Z}$ and the above matrix has determinant $adp^{2r} - bcN = p^r$.

Exercise 8.3.1. *Show W_p is independent of the choice of a, b, c, d. (This is [AL70, Lem 10].)*

Denote by W_N the Fricke involution

$$W_N f = \check{f} = N^{k/2} f| \begin{pmatrix} 0 & -1 \\ N & 0 \end{pmatrix}$$

introduced in Lemma 7.3.5. Note if $N = p$, then $W_N = W_p$.

Theorem 8.3.2. *The operators W_N and W_p on $M_k(N)$ for $p|N$ commute with T_m for $\gcd(m, N) = 1$.*

Proof. For W_p, see [AL70, Lem 11]. □

8.4 New and old forms

The problem with finding a basis of $M_k(N)$ which are eigenforms for all T_n is that some forms really come from smaller level, and the T_n's depend upon not just n but N.

To be precise, if $d|N$, then we get forms of level N from forms of level d just by inclusion $M_k(d) \subset M_k(N)$, simply because $\Gamma_0(d) \supset \Gamma_0(N)$. Let $f(z) = \sum a_n q^n$ be a form in $M_k(N)$ which is also in $M_k(d)$, and suppose p is a prime dividing N but not d. To distinguish the Hecke action in level d and level N, write T_p^d (resp. T_p^N) for the p-the Hecke operator on $M_k(d)$ (resp. $M_k(N)$). Then, by Theorem 8.1.3, we have

$$(T_p^d f)(z) - (T_p^N f)(z) = p^{k-1} \sum_{n \equiv 0 \bmod p} a_{n/p} q^n = p^{k-1} \sum_{n=0}^{\infty} a_n q^{pn} = p^{k-1} f(pz).$$

Supposing f is a normalized eigenform on $M_k(d)$, we see

$$(U_p f)(z) = (T_p^N f)(z) = (T_p^d f)(z) - p^{k-1} f(pz) = a_p f(z) - p^{k-1} f(pz).$$

(Note the similarity to (8.2.1).) Hence f is also an eigenform for T_p^N only if there exists λ such that $f(pz) = \lambda f(z)$ for all z, i.e., only if f is constant. Therefore, when we raise the level at p (via simple inclusion $M_k(d) \subset M_k(N)$), eigenforms for T_p do not generally remain eigenforms at p.

So, if we want to try to find an basis of eigenforms of level N for all T_n, it looks like we should try to get rid of all forms that come from smaller level.[3] Now you might ask if there are any other ways to construct modular forms of level N from those of smaller levels besides simple inclusion. We've already seen one related to the U_p transformations. Note that with d, N, and p as above, we have that for $f \in M_k(d)$,

$$(U_p f)(z) = (T_p^N f)(z) = (T_p^d f)(z) - p^{k-1} f(pz).$$

Since $T_p^d f \in M_k(d) \subset M_k(N)$, we see $U_p f = T_p^N f \in M_k(N)$ implies the function $z \mapsto f(pz)$ also lies in $M_k(N)$.

We can generalize this construction to replace p by an arbitrary integer:

Proposition 8.4.1. *Suppose $f(z) = \sum a_n q^n \in M_k(N)$. Then*

$$g(z) = f(mz) = \sum a_n q^{mn} \in M_k(mN).$$

Further, if $f \in S_k(N)$, show $g \in S_k(mN)$.

Exercise 8.4.2. *Prove the above proposition.*

The next exercise tells us more about the eigenspaces of $U_p = T_p^N$ in our earlier setup.

[3] We haven't shown this is strictly necessary, except in $M_2(8)$ with Eisenstein series earlier. In fact Exercise 8.4.3 and Remark 8.4.4, together with the Atkin–Lehner decomposition below, indicate that for $f \in S_k(d)$ a eigenform, f not being an eigenform for T_p^N in $S_k(N)$ should not be an honest obstruction to finding a basis of complete eigenforms for $S_k(N)$ at least when $p^2 \nmid N$.

Exercise 8.4.3. *Let $p, d|N$ with $p \nmid d$. Suppose $f(z) = \sum a_n q^n$ is a normalized eigenform, let $g(z) = f(pz)$, and V_f be the 2-dimensional subspace of $M_k(N)$ generated by $f = \sum a_n q^n$ and g.*

(i) Show $T_p^N g = f$.

(ii) Prove that V_f has a basis of eigenforms for T_p^N if and only if $|a_p| \neq 2\sqrt{p^{k-1}}$.

In this exercise you should observe that U_p restricted to the image of $M_k(d)$ in $M_k(N)$ under inclusion is simply the inverse of this inclusion map.

Remark 8.4.4. *You may recall we earlier mentioned Deligne's bound for cusp forms which says $|a_p| \leq 2\sqrt{p^{k-1}}$ in the above setting. The Sato–Tate conjecture (now a theorem) says that the a_p's are, in a suitable sense, uniformly distributed in this range (for p fixed and f varying). Thus we very rarely expect equality, and in fact it is conjectured that Deligne's inequality is really a strict inequality. This means that, in our earlier discussion, while an eigenform $f \in S_k(d)$ is no longer an eigenform for T_p in $S_k(pd)$, it can conjecturally be expressed as a linear combination of two eigenforms for T_p in $S_k(pd)$.*

Now we come to the definition of newforms and oldforms.

Definition 8.4.5. *We define the space of **oldforms** of $S_k(N)$ by*

$$S_k^{\text{old}}(N) = \text{Span}\left\{f(mz) : f \in S_k(d),\, dm|N,\, d < N\right\}.$$

Definition 8.4.6. *We define the space of **newforms** of $S_k(N)$ to be the orthogonal complement of $S_k^{\text{old}}(N)$ in $S_k(N)$, i.e.,*

$$S_k(N) = S_k^{\text{old}}(N) \oplus S_k^{\text{new}}(N).$$

Let us say a newform $f \in S_k(N)^{\text{new}}$ is an **eigennewform**[4] if it is an eigenform of all the T_p's (and thus all the T_n's), of all W_p's for $p|N$, and of W_N.

Theorem 8.4.7 (Atkin–Lehner)**.** *Let $f(z) = \sum a_n q^n \in S_k^{\text{new}}(N)$ be an eigennewform. Then*

1. *the eigenvalues λ_p's and λ_N for f under the action of the W_p's ($p|N$) and W_N are ± 1;*
2. *$\prod_{p|N} \lambda_p = \lambda_N$;*
3. *if $p|N$ but $p^2 \nmid N$, then $a_p = -\lambda_p p^{\frac{k}{2}-1}$;*
4. *if $p^2|N$, then $a_p = 0$.*

Theorem 8.4.8 (Atkin–Lehner)**.** *The space $S_k^{\text{new}}(N)$ has a basis consisting of eigennewforms.*

[4]In the literature, the term *newform* usually means a normalized eigennewform, so in this terminology the set of newforms are a basis for $S_k^{\text{new}}(N)$ rather than the whole space. For instance, in [AL70], the term newform means what we call eigennewform. You should probably stick to standard terminology and just pity my deep-seated need to call all the elements of $S_k(N)^{\text{new}}$ newforms coming from psychological inadequacies formed in childhood. I blame a clown poster in my bedroom.

Both of these theorems are contained in [AL70, Thm 3] (together with the observation that $\prod_{p|N} W_p = W_N$ for 2).

We remark the following consequence of the above theorems (along with multiplicativity of Fourier coefficients of eigennewforms), which gives an (only if) test for whether f lies in the space newforms of non-squarefree level, i.e., an easy way to detect oldforms in certain cases.

Corollary 8.4.9. *Let $f = \sum a_n q^n \in S_k(N)$ and suppose $p^2|N$ for some prime p. Then $f \in S_k^{\text{new}}(N)$ only if $a_n = 0$ for all $n \equiv 0 \bmod p$.*

Note there is no analogous test when N is squarefree. For instance, if $k = 2$ and $p|N$ but $p^2 \nmid N$, then there are typically eigennewforms $f, g \in S_2^{\text{new}}(N)$ where $a_p(f) = +1$ and $a_p(g) = -1$, so we can make the p-th Fourier coefficient of some linear combination of f and g arbitrary. Of course if one restricts to complete normalized eigenforms, we can use Theorem 8.4.7 directly.

Atkin and Lehner also describe the full space of cusp forms in terms of newforms of smaller levels. For $d|N$ and $m|\frac{N}{d}$, we have a map $\iota_m : S_k(d) \to S_k(N)$ via $f(z) \mapsto f(mz)$. When $m = p$, this is just the map given by the U_p operator.

Proposition 8.4.10. *We have*

$$S_k(N) = \bigoplus_{d|N} \bigoplus_{m|\frac{N}{d}} \iota_m(S_k^{\text{new}}(d)). \tag{8.4.1}$$

Example 8.4.11. *For prime and prime squared levels, the above decomposition explicitly reads*

$$S_k(p) = S_k(1) \oplus \iota_p(S_k(1)) \oplus S_k^{\text{new}}(p)$$

and

$$S_k(p^2) = S_k(1) \oplus \iota_p(S_k(1)) \oplus \iota_{p^2}(S_k(1)) \oplus S_k^{\text{new}}(p) \oplus \iota_p(S_k^{\text{new}}(p)) \oplus S_k^{\text{new}}(p^2).$$

Note that in the case of $S_k(p)$, this means the old space can be decomposed as a direct sum $S_k^{\text{old}}(p) = \bigoplus V_f$, where $V_f = \mathbb{C}f + \mathbb{C}\iota_p f$ and f runs over a basis of eigennewforms for $S_k(1)$. Then by Exercise 8.4.3 and Remark 8.4.4, this means the full cuspidal space $S_k(p)$ conjecturally has a basis of complete eigenforms.

Each $S_k^{\text{new}}(d)$ has a basis consisting of eigennewforms (of level d), and their Fourier coefficients are multiplicative for all n, so this gives a basis of $S_k(N)$ consisting of forms which have this multiplicativity property. Note there is a difference between the Fourier coefficients of f being multiplicative for all n and being an eigenform for all T_n.

We can use the dimension formulas given in Section 5.2 to compute $\dim S_k(N)^{\text{new}}$. The general case is treated in [Mar05]. We just illustrate a couple of special cases:

Corollary 8.4.12. *For $k \geq 2$ even, we have*

$$\dim S_k^{\text{new}}(p) = \dim S_k(p) - 2\dim S_k(1)$$

and

$$\dim S_k^{\text{new}}(p^2) = \dim S_k(p^2) - 2\dim S_k(p) + \dim S_k(1).$$

In particular, for $k = 2$, we have

$$\dim S_2^{\text{new}}(p) = \dim S_2(p) = \frac{p+1}{12} - \frac{1}{4}\left(1 + \left(\frac{-1}{p}\right)\right) - \frac{1}{3}\left(1 + \left(\frac{-3}{p}\right)\right),$$

and

$$\dim S_2^{\text{new}}(p^2) = \begin{cases} \frac{1}{12}(p+1)(p-8) + 1 + \frac{1}{4}\left(1 + \left(\frac{-1}{p}\right)\right) + \frac{1}{3}\left(1 + \left(\frac{-3}{p}\right)\right) & p \geq 7, \\ 0 & p \leq 5. \end{cases}$$

Proof. The first two dimension formulas follow immediately from the explicit decompositions in the previous example. When $k = 2$, note that $S_2(1) = \{0\}$, so $S_2^{\text{new}}(p) = S_2(p)$, and that dimension formula was given in (5.2.4). The explicit $k = 2$ for level p^2 then follows from (5.2.5). □

We remark that while we can't find a basis of $S_k(N)$ consisting of complete eigenforms in general, Atkin and Lehner showed one can find a basis of eigenforms which are also eigenfunctions of the W_q's.

Proposition 8.4.13 ([AL70], Lem 27)**.** *There exists a basis of $S_k(N)$ consisting of functions which are eigenforms for all T_p with $p \nmid N$ and W_q with $q|N$.*

Note that the decomposition (8.4.1) also induces a canonical basis of eigenforms of $S_k(N)$. Namely for each $d|N$, one takes the set of normalized eigennewforms f of $S_k(d)$. Then the union of the sets $\{f(mz) : m|\frac{N}{d}\}$ for all such d, f gives a uniquely defined basis for $S_k(N)$. However, the oldforms in this basis are not eigenfunctions of the W_q's. See [AL70, Sec 5] for a precise description of a basis as in Proposition 8.4.13 in terms of eigennewforms of $S_k(d)$, $d|N$, including a description of the eigenvalues of W_q on the oldforms.

Now we say describe a refinement of the decomposition in (8.4.1), related to such bases. For $d|N$ and f an eigennewform of $S_k(d)$, we consider the subspace of $S_k(N)$ generated by $f = \sum a_n q^n$:

$$\langle f \rangle_d^N := \bigoplus_{m|\frac{N}{d}} \iota_m(\mathbb{C}f) = \bigoplus_{m|\frac{N}{d}} \mathbb{C}f(mz).$$

Exercise 8.1.9 tells us that for $p \nmid m$, $f(mz)$ is an eigenform for T_p with eigenvalue a_p, i.e., ι_m preserves Hecke eigenvalues prime to m. Consequently T_p acts on $\langle f \rangle_d^N$ by scaling by a_p for $p \nmid N$, and thus T_n acts on this space by a_n when $\gcd(n, N) = 1$. Thus the decomposition (8.4.1) can be refined to

$$S_k(N) = \bigoplus_{d|N} \bigoplus_f \langle f \rangle_d^N, \tag{8.4.2}$$

where f runs over eigennewforms of $S_k(d)$. Moreover this latter decomposition decomposition is a decomposition into the collection of common eigenspaces for the T_n's with $\gcd(n, N) = 1$:

Theorem 8.4.14 (Atkin–Lehner)**.** *Suppose $g \in S_k(N)$ is an eigenfunction of all T_p's for $p \nmid N$. Then there exists a eigennewform $f \in S_k(d)$ for some $d|N$ such that $g \in \langle f \rangle_d^N$. Furthermore, if $g \neq 0$, there is a unique such f.*

In fact, since $S_k(N)$ is finite dimensional, there is some bound B (depending on k, N) such that the theorem is true if we only require g to be an eigenfunction of all T_p's with $p \nmid N$ and $p \leq B$. Namely, we can take B as in Sturm's bound (Theorem 5.2.1).

This first part of this theorem is [AL70, Thm 4], and the proof comes essentially by using the decomposition (8.4.2) (which then implies the uniqueness of f) and comparing L-functions of different eigennewforms.

We remark this theorem is also a special case of a theorem known as **strong multiplicity one** for automorphic representations. The representation theoretic connection is that each eigennewform f determines a unique cuspidal automorphic representation (the subject of another course) π_f, whose representation space restricted to $S_k(N)$ is precisely $\langle f \rangle_d^N$. Which is to say that these spaces $\langle f \rangle_d^N$ are quite natural to look at from the point of view of representation theory, and the Strong Multiplicity One Theorem says that one can distinguish different cuspidal automorphic representations π_f, $\pi_{f'}$ (here f' is some other eigennewform of possibly a different level, and even possibly a different weight) by looking at sufficiently many Hecke eigenvalues prime to the levels of f and f'.

What this theorem means in practice is that eigennewforms are determined by a finite collection of Hecke eigenvalues (once we specify the weight and bound the level, though in fact strong multiplicity one says it suffices to merely bound the weight and level), so for many purposes to understand $S_k(N)$ it suffices to determine (finite) systems of Hecke eigenvalues. Computing systems of Hecke eigenvalues λ_n's for $\gcd(n, N) = 1$ is then equivalent to obtaining the decomposition (8.4.2). There are known ways to compute the actions of Hecke operators T_n on $S_k(N)$ without knowing in advance the space $M_k(N)$ (say as given by Fourier expansions), e.g., using *modular symbols* or *Brandt matrices on quaternion algebras*.[5] These provide effective ways to compute $S_k(N)$.

We illustrate this by means of an example.

Example 8.4.15. *We can compute in Sage some matrices for Hecke operators on* $S_2(55)$*:*

$$T_2 = \begin{pmatrix} 0 & 2 & 0 & -1 & 1 \\ 1 & 0 & -2 & 3 & -1 \\ 0 & 0 & -2 & 3 & -3 \\ 0 & 1 & -2 & 3 & -3 \\ 0 & 0 & 0 & 0 & -2 \end{pmatrix}, \quad T_3 = \begin{pmatrix} 0 & 0 & 2 & -2 & 0 \\ 0 & -2 & 2 & -2 & 0 \\ 1 & -2 & 5 & -4 & -2 \\ 0 & -2 & 4 & -4 & -1 \\ 0 & 0 & 0 & 0 & -1 \end{pmatrix},$$

and

$$T_7 = \begin{pmatrix} -1 & 1 & 0 & -1 & 1 \\ 0 & -2 & 0 & 0 & 0 \\ -1 & -1 & -2 & 1 & -1 \\ -1 & -1 & 0 & -1 & -1 \\ 0 & 0 & 0 & 0 & -2 \end{pmatrix}$$

[5] Actually Brandt matrices compute Hecke operators on a slightly smaller space, which can be both computationally and theoretically advantageous. E.g., when $k \geq 4$ and $N = p$ is prime, the Brandt matrices naturally give the Hecke operators on $S_k^{\text{new}}(p)$.

Then T_2, T_3 and T_7 have as bases of eigenvectors

$$v_1 = \begin{pmatrix} 1 \\ 0 \\ -1 \\ -1 \\ 0 \end{pmatrix}, v_2 = \begin{pmatrix} 1 \\ 2+\sqrt{2} \\ 3 \\ 3+\sqrt{2} \\ 0 \end{pmatrix}, v_3 = \begin{pmatrix} 1 \\ 2-\sqrt{2} \\ 3 \\ 3-\sqrt{2} \\ 0 \end{pmatrix}, v_4 = \begin{pmatrix} 2 \\ -2 \\ 0 \\ 1 \\ 1 \end{pmatrix}, v_5 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

For both T_2 and T_3, v_1, v_2 and v_3 generate distinct eigenspaces, but v_4 and v_5 span a 2-dimensional eigenspace. In this case, T_7 only has 2 eigenspaces. Specifically, with respect to this basis of eigenvectors, these Hecke matrices become

$$T_2' = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1+\sqrt{2} & 0 & 0 & 0 \\ 0 & 0 & 1-\sqrt{2} & 0 & 0 \\ 0 & 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & 0 & -2 \end{pmatrix}, \quad T_3' = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & -2\sqrt{2} & 0 & 0 & 0 \\ 0 & 0 & 2\sqrt{2} & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{pmatrix},$$

and

$$T_7' = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & 0 & -2 \end{pmatrix},$$

From looking at the eigenvalues for T_2 and T_3, we know the first three eigenvectors must correspond to newforms (because Hecke eigenvalues oldforms will appear multiple times), and we might suspect that the 2-dimensional space $\langle v_4, v_5 \rangle$ corresponds to a 2-dimensional oldspace $\langle f \rangle_p^{55}$ where $p = 5$ or $p = 11$.

Indeed, $S_2(1) = S_2(5) = 0$ but $\dim S_2(11) = 1$. So the decomposition theorem says

$$S_2(55) = S_2^{\text{new}}(55) \oplus \langle f_4 \rangle_{11}^{55},$$

with the latter space being 2-dimensional, and

$$f_4(z) = q - 2q^2 - q^3 + 2q^4 + q^5 + 2q^6 - 2q^7 - 2q^9 + \cdots$$

(Here we know the Fourier coefficients a_2, a_3 and a_7 by the above calculations. One can separately check T_5 on $S_2(11)$, and get the other listed coefficients by multiplicativity and recurrence relations.) The above calculations, together with the calculation of T_5, tells us

$$S_2^{\text{new}}(55) = \mathbb{C}f_1 \oplus \mathbb{C}f_2 \oplus \mathbb{C}f_3,$$

where

$$\begin{aligned} f_1(z) &= q + q^2 - q^4 + q^5 - 3q^8 - 3q^9 + \cdots \\ f_2(z) &= q + (1+\sqrt{2})q^2 - 2\sqrt{2}q^3(1-2\sqrt{2})q^5 - q^5 - (4+2\sqrt{2})q^6 - 2q^7 + \cdots \\ f_3(z) &= q + (1-\sqrt{2})q^2 + 2\sqrt{2}q^3(1+2\sqrt{2})q^5 - q^5 - (4-2\sqrt{2})q^6 - 2q^7 + \cdots \end{aligned}$$

(Here for $i = 1, 2, 3$, f_i is the eigennewform with Fourier coeffcient a_p equal to the eigenvalue for v_i under the action of T_p.)

Let us end with some remarks on the original question about whether $M_k(N)$ has a basis consisting of forms whose Fourier coefficients are multiplicative for all n (i.e., $a_{mn} = a_m a_n$ whenever $\gcd(m,n) = 1$).

By (8.4.2), for the space of cusp forms $S_k(N)$, this boils down to determining whether, for every newform $f \in S_k^{\text{new}}(d)$, the subspace $V_{f,m} := \langle \iota_m(f) : m|\frac{N}{d} \rangle$ has a basis consisting of forms with multiplicative coefficients. As in Example 8.4.11, we expect this to be true for $S_k(p)$ because conjecturally there is a basis of complete eigenforms.

On the other hand, the Atkin–Lehner decomposition easily gives a positive answer to a more generalized notion of multiplicative coefficients:

Proposition 8.4.16. *The space $S_k(N)$ has a basis of forms $f(z) = \sum a_n q^n$ whose Fourier coefficients are essentially multiplicative in the following sense: for each such f, there exists $m \in \mathbb{N}$ (in fact $m|N$) such that $a_n = 0$ if $m \nmid n$ and $a_{mnn'} = a_{mn} a_{mn'}$ for all $n, n' \in \mathbb{N}$ such that $\gcd(n,n') = 1$.*

Note that the essential multiplicativity condition is simply multiplicativity when $m = 1$.

Proof. Note $\iota_m(f)$ satisfies this property for a newform $f \in S_k(d)$, and apply the decomposition (8.4.2). □

If one wants to extend this to $M_k(N)$, one needs a newform theory for Eisenstein series. This was not done by Atkin and Lehner, but worked out in the thesis of Weisinger (Harvard, 1977). We won't explain this, but just mention it is possible. The obstruction to defining newforms in $M_k(N)$ in the same way as for $S_k(N)$ is that the Petersson inner product is not defined on all of $M_k(N)$.

Chapter 9

Hilbert modular forms

The modular forms we have been studying up until now are often called **elliptic modular forms**, due to the connection with classical elliptic functions (and elliptic curves). In this chapter, we give a brief exposition of the theory of Hilbert modular forms, which is one kind of generalization of our usual elliptic modular forms. We will assume some familiarity with basic algebraic number theory in this chapter.

We motivated the theory of modular forms in the introduction via quadratic forms and theta series. Namely, the number of ways $r_k(n)$ is a sum of k-squares is just the n-th coefficient in the q-expansion of $\vartheta^k(z)$. Observing that ϑ satisfies some transformation properties, we consider the space of functions with similar properties, and are led to the the definition of a modular form of weight $k/2$.

One can similarly motivate the theory of Hilbert modular forms by considering quadratic forms over more general number fields. For simplicity, suppose $F = \mathbb{Q}(\sqrt{d})$ is with $d > 0$ the discriminant and having class number one. This has ring of integers $\mathcal{O}_F = \mathbb{Z}[\sqrt{d}]$ if $d \equiv 0 \bmod 4$ and $\mathcal{O}_F = \mathbb{Z}[\frac{1+\sqrt{d}}{2}]$ if $d \equiv 1 \bmod 4$. Here the right analogue of Jacobi's theta function from (4.4.1) is

$$\vartheta_{\mathcal{O}_F}(z_1, z_2) = \sum_{a+b\sqrt{d} \in \mathcal{O}_F} e^{2\pi i(z_1(a+b\sqrt{d})^2 + z_2(a-b\sqrt{d})^2)} = \sum_{\alpha \in \mathcal{O}_F} q_1^{\alpha^2} q_2^{\overline{\alpha}^2},$$

where $q_i = e^{2\pi i z_i}$. The reason one considers a theta function of 2 variables is to account for the two embeddings of F into $\mathbb{R}$. Note this contains the more naive generalization

$$\sum_{\alpha \in \mathcal{O}_F} q^{\alpha^2} = \vartheta_{\mathcal{O}_F}(z, 0)$$

of Jacobi's theta function, where, as usual $q = e^{2\pi i z}$. However, one gets a nicer theory by accounting for the different embedding of F into $\mathbb{R}$ (just as one does in algebraic number theory). Using $\vartheta_{\mathcal{O}_F}$, one can study the number of representations of n in $\mathcal{O}_F$ as a sum of k squares. Again, one can proceed from the transformation properties of $\vartheta_{\mathcal{O}_F}$ to define Hilbert modular forms over F. Notice that our "Fourier expansion" is also more complicated in this case—it runs not over $\mathbb{Z}$ but over $\mathcal{O}_F$.

While quadratic forms provided our main motivating problem to study modular forms, we built up the definition more geometrically, by considering the surfaces $Y_0(N) = \Gamma_0(N)\backslash\mathfrak{H}$

(or rather their compactifications $X_0(N)$) and studying functions on them. Recall that while there are no nonconstant holomorphic functions on $X_0(N)$, the derivatives of modular functions (weak modular forms) have different transformation properties and for most even weights (just exclude $k = 2$ when $N = 1$), there are non-constant holomorphic forms. Consequently one can construct modular functions by constructing holomorphic modular forms and taking appropriate products and ratios to get something of weight 0 (a modular function). This gives a construction of the famous j-invariant modular function in Exercise 4.2.13, which is important in the theory of elliptic curves.[1]

In our real quadratic case $F = \mathbb{Q}(\sqrt{d})$, the analogue will be to look at functions on a quotient $\Gamma\backslash(\mathfrak{H} \times \mathfrak{H})$, where Γ is a group of isometries of $\mathfrak{H} \times \mathfrak{H}$. Note that since $\mathrm{PSL}_2(\mathbb{R})$ is the isometry group of $\mathfrak{H}$, $\mathrm{PSL}_2(\mathbb{R}) \times \mathrm{PSL}_2(\mathbb{R})$ acts by isometries on $\mathfrak{H} \times \mathfrak{H}$. The embedding $\alpha \mapsto (\alpha, \overline{\alpha})$ of F into $R \times \mathbb{R}$ that one usually considers in algebraic number theory induces an embedding of $\mathrm{PSL}_2(F)$ into $\mathrm{PSL}_2(\mathbb{R}) \times \mathrm{PSL}_2(\mathbb{R})$. The analogue of looking at congruence subgroups of $\mathrm{PSL}_2(\mathbb{Z})$ in the case of modular forms is to look at congruence subgroups

$$\Gamma \subset \mathrm{PSL}_2(\mathcal{O}_F) \subset \mathrm{PSL}_2(\mathbb{R}) \times \mathrm{PSL}_2(\mathbb{R}).$$

In fact one often replaces $\mathrm{PSL}_2(\mathcal{O}_F)$ by slightly more general arithmetic groups.

Either of these approaches lead one to define Hilbert modular forms over a *totally real* number field F (i.e., all embeddings of F into $\mathbb{C}$ lie in $\mathbb{R}$, so $\mathbb{Q}(\sqrt[3]{2})$ is real, but not totally real).

Hilbert's motivation for studying modular forms over number fields was as follows. The Kronecker–Weber theorem says that any abelian extension of $\mathbb{Q}$ can be generated by roots of unity, i.e., special values of the function $f(x) = e^{ix}$. The theory of complex multiplication does something analogous for imaginary quadratic extensions $K/\mathbb{Q}$, realizing *Kronecker's Jugendtraum* ("dream of youth"). Specifically, one can generate any abelian extension of K using special values of the j-invariant special values of the Weierstrass $\wp$ function. Hilbert's twelfth problem asks more generally for a determination of the the abelian extensions of a general number field K. Consequently, one wants analogues of special functions like exponential functions, the modular function $j(\tau)$ and the $\wp$ function. Hilbert's idea was to consider modular forms over number fields to find analogues of these special functions.

Some references are [Fre90] for full level, [Bru08] or [vdG88] for F real quadratic, or [Gar90] or [Shi78] in full generality.

9.1 Basic definitions and results

Let F be a totally real number field of degree r, i.e., $[F : \mathbb{Q}] = r$ and there are r embeddings, $\iota_1, \ldots, \iota_r$, of F into $\mathbb{R}$. Denote the ring of integers of F by $\mathcal{O}_F$. We consider all of these embeddings at once via

$$\iota : F \to \mathbb{R}^r, \quad \iota(x) = (\iota_1(x), \ldots, \iota_r(x)).$$

[1] In retrospect, I suppose this was a bit incoherent—our motivation for modular forms was to study quadratic forms, and we use modular functions as motivation for the definition, but then we essentially drop the problem of constructing modular functions once we have the definition of modular forms apart from that one exercise, and just study modular forms with applications to quadratic forms. If I ever add a chapter on elliptic curves, I will hopefully say more about the j-invariant so that there is some payoff of spending all that time on modular functions.

The advantage of doing this is that then $\iota(\mathcal{O}_F)$ is a lattice (discrete $\mathbb{Z}$-module) in $\mathbb{R}^r$, which is analogous to $\mathbb{Z}$ being a discrete subgroup of $\mathbb{R}$. (Note if we just consider a single embedding, e.g., $\mathbb{Z}[\sqrt{5}] \subset \mathbb{R}$, we do not get a discrete subset of $\mathbb{R}$ when $r > 1$.)

Elliptic modular forms arose from looking at the modular curves $\Gamma\backslash\mathfrak{H}$, where Γ is a finite-index subgroup of $\mathrm{PSL}_2(\mathbb{Z})$, which acts discretely on $\mathfrak{H}$ (i.e., the orbits under this action are discrete subsets of $\mathfrak{H}$). The analogue should be to look at finite-index subgroups of $\mathrm{PSL}_2(\mathcal{O}_F)$. But since $\mathrm{PSL}_2(\mathcal{O}_F) \subset \mathrm{PSL}_2(\mathbb{R})$ does not act discretely on $\mathfrak{H}$. Just as $\mathcal{O}_F$ can be viewed as a lattice in $\mathbb{R}^r$, we can view $\mathrm{PSL}_2(\mathcal{O}_F)$ acting discretely on the r-fold product of upper-half planes $\mathfrak{H}^r$.

For $g = (g_1, \dots, g_r) \in \mathrm{PSL}_2(\mathbb{R})^r$, we let g act on $\mathfrak{H}^r$ via

$$g \cdot z = (g_1 z_1, \dots, g_r z_r), \quad z = (z_1, \dots, z_r).$$

Then we extend

$$\iota : \mathrm{GL}_2(F) \to \mathrm{GL}_2(\mathbb{R})^r, \quad \iota(\gamma) = (\iota_1(\gamma), \dots, \iota_r(\gamma)),$$

where$\iota_i(\gamma) : \mathrm{GL}_2(F) \to \mathrm{GL}_2(\mathbb{R})$ is just given by coordinate-wise application of ι_i. Restricting to $\mathrm{SL}_2(F)$ and quotienting out by $\pm I$ gives gives an embedding $\iota : \mathrm{PSL}_2(F) \to \mathrm{PSL}_2(\mathbb{R})^r$. So we let $\mathrm{PSL}_2(F)$ act on $\mathfrak{H}^r$ via composition with ι. Notationally, if $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PSL}_2(F)$ and $z = (z_1, \dots, z_r)$, we will denote this action by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} z = \frac{az+b}{cz+d} := \left(\frac{\iota_1(a)z_1 + \iota_1(b)}{\iota_1(c)z_1 + \iota_1(d)}, \dots, \frac{\iota_r(a)z_1 + \iota_r(b)}{\iota_r(c)z_1 + \iota_r(d)} \right).$$

Proposition 9.1.1. $\mathrm{PSL}_2(\mathcal{O}_F)$ *acts discretely on* $\mathfrak{H}^r$.

We call $\mathrm{PSL}_2(\mathcal{O}_F)$ the **(full) Hilbert modular group**. The analogue of the congruence subgroups $\Gamma_0(N)$ inside $\mathrm{PSL}_2(\mathbb{Z})$ are

$$\Gamma_0(\mathfrak{N}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathcal{O}_F) : c \in \mathfrak{N} \right\} / \{\pm I\},$$

where $\mathfrak{N}$ is a nonzero integral ideal of $\mathcal{O}_F$. We call these **congruence subgroups** of $\mathrm{PSL}_2(\mathcal{O}_F)$. (As when $r = 1$, there are more congruence subgroups than just the $\Gamma_0(\mathfrak{N})$'s, but these are all we will consirer.) Note when $F = \mathbb{Q}$, $\Gamma_0(N\mathbb{Z}) = \Gamma_0(N)$ for $N \in \mathbb{N}$. In general, one can check that $\Gamma_0(\mathfrak{N})$ has finite index in $\mathrm{PSL}_2(\mathcal{O}_F)$.

As in the case of $r = 1$, one adjoins cusps to $\mathfrak{H}^r$. Namely, we define the extended r-fold product of upper half-planes $\overline{\mathfrak{H}^r}$ to be the union of $\mathfrak{H}^r$ together with the boundary points $\iota(F) \subset \mathbb{R}^r$ and the point at infinity $i\infty = (i\infty, \dots, i\infty)$. Then $\overline{\mathfrak{H}^r} \subset \{\mathfrak{H} \cup \mathbb{R} \cup \{i\infty\}\}^r$. The elements of $\overline{\mathfrak{H}^r} - \mathfrak{H}^r$ are the **cusps** of $\mathfrak{H}^r$. There is a natural way to put a topology on $\overline{\mathfrak{H}^r}$ as well as extend the action of $\mathrm{PSL}_2(F)$ to $\overline{\mathfrak{H}^r}$. Thus for a subgroup Γ of $\mathrm{PSL}_2(\mathcal{O}_F)$, we can talk about the cusps of the quotient $\Gamma\backslash\mathfrak{H}^r$, which are the Γ-orbits of the cusps of $\mathfrak{H}^r$.

Unlike the case of $F = \mathbb{Q}$, there may be many cusps for $\mathrm{PSL}_2(\mathcal{O}_F)\backslash\mathfrak{H}^r$, but one can show there are finitely many:

Theorem 9.1.2. *The number of cusps of* $\mathrm{PSL}_2(\mathcal{O}_F)\backslash\mathfrak{H}^r$ *is the class number* h_F *of* F.

For $c, d \in F$, $\mathbf{k} \in \mathbb{Z}^r$ and $z = (z_1, \ldots, z_r) \in \mathfrak{H}^r$, we denote by

$$(cz + d)^{\mathbf{k}} = (\iota_1(c)z_1 + \iota_1(d))^{k_1} \cdots (\iota_r(c)z_1 + \iota_r(d))^{k_r}.$$

Definition 9.1.3. *Let $\mathbf{k} = (k_1, \ldots, k_r) \in \mathbb{Z}^r$ and Γ be a finite index subgroup of $\mathrm{PSL}_2(\mathcal{O}_F)$. A* **(holomorphic) Hilbert modular form** *of* **weight k** *on $\Gamma\backslash\mathfrak{H}^r$ is holomorphic* function $f : \overline{\mathfrak{H}^r} \to \mathbb{C}$ satisfying*

$$f(\gamma z) = (cz + d)^{\mathbf{k}} f(z), \quad \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(\mathfrak{N}). \tag{9.1.1}$$

If also f vanishes at the cusps, we say f is a **cusp form**. *Denote the space of weight $\mathbf{k}$ Hilbert modular forms and cusp forms on $\Gamma\backslash\mathfrak{H}^r$ respectively by $M_{\mathbf{k}}(\Gamma)$ and $S_{\mathbf{k}}(\Gamma)$.*

If $\Gamma = \Gamma_0(\mathfrak{N})$, then we write $M_{\mathbf{k}}(\mathfrak{N}) = M_{\mathbf{k}}(\Gamma)$ and $S_{\mathbf{k}}(\mathfrak{N}) = S_{\mathbf{k}}(\Gamma)$. A form in $M_{\mathbf{k}}(\mathfrak{N})$ is said to have **level $\mathfrak{N}$**.

For $F = \mathbb{Q}$, we have $M_k(N\mathbb{Z}) = M_k(N)$ and $S_k(N\mathbb{Z}) = S_k(N)$. When $F \neq \mathbb{Q}$, by a holomorphic function of $\overline{\mathfrak{H}^r}$ we mean a function $f : \mathfrak{H}^r \to \mathbb{C}$ which is holomorphic in each of the r complex variables $z_1, \ldots, z_r$. It should also extend to be "holomorphic at the cusps," which is a notion one can define precisely, and once one defines it one can show it's not actually needed! (That is, it's not needed for the definition of holomorphic Hilbert modular forms, but it is used in the theory.)

Theorem 9.1.4. (Koecher's principle) *Suppose $[F : \mathbb{Q}] > 1$. If $f : \mathfrak{H}^r \to \mathbb{C}$ is holomorphic and satisfies (9.1.1), then f is holomorphic at the cusps.*

The proof of Koecher's principle uses Fourier expansions, which we describe next (that is, we describe Fourier expansions, but not the proof of Koecher's principle).

The **inverse different** of F is the fractional ideal

$$\mathcal{O}_F^{\perp} = \left\{ x \in F : \mathrm{tr}_{F/\mathbb{Q}}(x\mathcal{O}_F) \subset \mathcal{O}_F \right\}.$$

Here the **trace** from F to $\mathbb{Q}$ is defined as the sum of the conjugates, i.e., $\mathrm{tr}_{F/\mathbb{Q}}(x) = \sum \iota_i(x)$. So if we define $\mathrm{tr} : \mathbb{C}^r \to \mathbb{C}$ via $\mathrm{tr}(z_1, \ldots, z_r) = \sum z_i$, then $\mathrm{tr}(\iota(x)) = \mathrm{tr}_{F/\mathbb{Q}}(x)$, which we also denote simply as $\mathrm{tr}(x)$.

Proposition 9.1.5. *Given $f \in M_{\mathbf{k}}(\mathfrak{N})$, we have the* **Fourier expansion (at ∞)**

$$f(z) = \sum_{\xi \in \mathcal{O}_F^{\perp}} c_\xi e^{2\pi i \mathrm{tr}(\xi z)} = \sum_{\xi \in \mathcal{O}_F^{\perp}} c_\xi e^{2\pi i(\iota_1(\xi)z_1 + \cdots + \iota_r(\xi)z_r)},$$

for some (uniquely determined) collection of **Fourier coefficients** $c_\xi \in \mathbb{C}$.

As when $F = \mathbb{Q}$, there are also Fourier expansions around other cusps.

To specify f, we sometimes denote the Fourier coefficient c_ξ by $c_\xi(f)$. The idea of the proof is to use that fact that f is invariant under action by $\begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}$ for all $x \in \mathcal{O}_F$.

We call $x \in F$ **totally positive** if $\iota_i(x) > 0$ for all x. Koecher's principle in fact tells us:

Theorem 9.1.6. (Koecher's principle, 2nd version) *For $f \in M_{\mathbf{k}}(\mathfrak{N})$, $c_\xi(f) = 0$ unless $\xi = 0$ or ξ is totally positive.*

As in the case of $F = \mathbb{Q}$, a necessary condition for f to be a cusp form is $c_0(f) = 0$.

Corollary 9.1.7. *Let $f \in M_{\mathbf{k}}(\mathfrak{N})$, where $\mathbf{k} = (k_1, \ldots, k_r)$. If some $k_i \neq k_j$, then $f \in S_{\mathbf{k}}(\mathfrak{N})$.*

We call the weight of the form $\mathbf{k} = (k, \ldots, k)$ **parallel weight** k. Thus the corollary tells us that we can only have non-cuspidal holomorphic Hilbert modular forms (e.g., the Eisenstein series discussed below) in parallel weights. Another nice thing about parallel weights is restriction to the diagonal yields elliptic modular forms:

Exercise 9.1.8. *Let $f \in M_{\mathbf{k}}(N\mathcal{O}_F)$ where $N \in \mathbb{N}$ and $\mathbf{k} = (k, \ldots, k)$. Show $g : \mathfrak{H} \to \mathbb{C}$ given by $g(z) = f(z, \ldots, z)$ is an elliptic modular form in $M_k(N)$.*

Theorem 9.1.9. $\dim M_{\mathbf{k}}(\mathfrak{N}) < \infty$.

Bibliography

[Ahl78] Lars V. Ahlfors, *Complex analysis*, third ed., McGraw-Hill Book Co., New York, 1978, An introduction to the theory of analytic functions of one complex variable, International Series in Pure and Applied Mathematics. MR 510197 (80c:30001)

[AL70] A. O. L. Atkin and J. Lehner, *Hecke operators on* $\Gamma_0(m)$, Math. Ann. **185** (1970), 134–160. MR 0268123 (42 #3022)

[And05] James W. Anderson, *Hyperbolic geometry*, second ed., Springer Undergraduate Mathematics Series, Springer-Verlag London Ltd., London, 2005. MR 2161463 (2006b:51001)

[Apo90] Tom M. Apostol, *Modular functions and Dirichlet series in number theory*, second ed., Graduate Texts in Mathematics, vol. 41, Springer-Verlag, New York, 1990. MR 1027834 (90j:11001)

[Boy01] Matthew Boylan, *Swinnerton-Dyer type congruences for certain Eisenstein series*, q-series with applications to combinatorics, number theory, and physics (Urbana, IL, 2000), Contemp. Math., vol. 291, Amer. Math. Soc., Providence, RI, 2001, pp. 93–108. MR 1874523 (2002k:11063)

[Bru08] Jan Hendrik Bruinier, *Hilbert modular forms and their applications*, The 1-2-3 of modular forms, Universitext, Springer, Berlin, 2008, pp. 105–179. MR 2447162

[Bum97] Daniel Bump, *Automorphic forms and representations*, Cambridge Studies in Advanced Mathematics, vol. 55, Cambridge University Press, Cambridge, 1997. MR 1431508 (97k:11080)

[CSS97] Gary Cornell, Joseph H. Silverman, and Glenn Stevens (eds.), *Modular forms and Fermat's last theorem*, Springer-Verlag, New York, 1997, Papers from the Instructional Conference on Number Theory and Arithmetic Geometry held at Boston University, Boston, MA, August 9–18, 1995. MR 1638473 (99k:11004)

[DS05] Fred Diamond and Jerry Shurman, *A first course in modular forms*, Graduate Texts in Mathematics, vol. 228, Springer-Verlag, New York, 2005. MR 2112196 (2006f:11045)

[Els06] Jürgen Elstrodt, *A very simple proof of the eta transformation formula*, Manuscripta Math. **121** (2006), no. 4, 457–459. MR 2283473 (2007g:11051)

[FB09] Eberhard Freitag and Rolf Busam, *Complex analysis*, second ed., Universitext, Springer-Verlag, Berlin, 2009. MR 2513384

[Fre90] Eberhard Freitag, *Hilbert modular forms*, Springer-Verlag, Berlin, 1990. MR 1050763

[Gar90] Paul B. Garrett, *Holomorphic Hilbert modular forms*, The Wadsworth & Brooks/Cole Mathematics Series, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1990. MR 1008244

[Iwa97] Henryk Iwaniec, *Topics in classical automorphic forms*, Graduate Studies in Mathematics, vol. 17, American Mathematical Society, Providence, RI, 1997. MR 1474964 (98e:11051)

[Kat92] Svetlana Katok, *Fuchsian groups*, Chicago Lectures in Mathematics, University of Chicago Press, Chicago, IL, 1992. MR 1177168 (93d:20088)

[Kil08] L. J. P. Kilford, *Modular forms*, Imperial College Press, London, 2008, A classical and computational introduction. MR 2441106 (2009m:11001)

[KL06] Andrew Knightly and Charles Li, *Traces of Hecke operators*, Mathematical Surveys and Monographs, vol. 133, American Mathematical Society, Providence, RI, 2006. MR 2273356

[Kob93] Neal Koblitz, *Introduction to elliptic curves and modular forms*, second ed., Graduate Texts in Mathematics, vol. 97, Springer-Verlag, New York, 1993. MR 1216136 (94a:11078)

[Lan95] Serge Lang, *Introduction to modular forms*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 222, Springer-Verlag, Berlin, 1995, With appendixes by D. Zagier and Walter Feit, Corrected reprint of the 1976 original. MR 1363488 (96g:11037)

[Lan99] ———, *Complex analysis*, fourth ed., Graduate Texts in Mathematics, vol. 103, Springer-Verlag, New York, 1999. MR 1659317 (99i:30001)

[Mar05] Greg Martin, *Dimensions of the spaces of cusp forms and newforms on* $\Gamma_0(N)$ *and* $\Gamma_1(N)$, J. Number Theory **112** (2005), no. 2, 298–331. MR 2141534

[Mil] J.S. Milne, *Modular functions and modular forms*, Online course notes. *http://www.jmilne.org/math/CourseNotes/mf.html*.

[Miy06] Toshitsune Miyake, *Modular forms*, english ed., Springer Monographs in Mathematics, Springer-Verlag, Berlin, 2006, Translated from the 1976 Japanese original by Yoshitaka Maeda. MR 2194815 (2006g:11084)

[Sch74] Bruno Schoeneberg, *Elliptic modular functions: an introduction*, Springer-Verlag, New York, 1974, Translated from the German by J. R. Smart and E. A. Schwandt, Die Grundlehren der mathematischen Wissenschaften, Band 203. MR 0412107 (54 #236)

[Ser73] J.-P. Serre, *A course in arithmetic*, Springer-Verlag, New York, 1973, Translated from the French, Graduate Texts in Mathematics, No. 7. MR 0344216 (49 #8956)

[Shi78] Goro Shimura, *The special values of the zeta functions associated with Hilbert modular forms*, Duke Math. J. **45** (1978), no. 3, 637–679. MR 507462

[Sta09] John Stalker, *Complex analysis*, Modern Birkhäuser Classics, Birkhäuser Boston Inc., Boston, MA, 2009, Fundamentals of the classical theory of functions, Reprint of the 1998 edition. MR 2547082 (2010i:30001)

[Ste07] William Stein, *Modular forms, a computational approach*, Graduate Studies in Mathematics, vol. 79, American Mathematical Society, Providence, RI, 2007, With an appendix by Paul E. Gunnells. MR 2289048 (2008d:11037)

[vdG88] Gerard van der Geer, *Hilbert modular surfaces*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)], vol. 16, Springer-Verlag, Berlin, 1988. MR 930101

[Zag08] Don Zagier, *Elliptic modular forms and their applications*, The 1-2-3 of modular forms, Universitext, Springer, Berlin, 2008, pp. 1–103. MR 2409678 (2010b:11047)

Index