ANALYTIC THEORY OF AUTOMORPHIC FORMS ON GL(2)

BINGRONG HUANG

ABSTRACT. In this mini-course, we will discuss (i) basic theories of automorphic forms, including modular forms, Maass forms and Eisenstein series; (ii) Hecke L-functions, Hecke operators, and the first moment of Fourier coefficients/Hecke eigenvalues; (iii) The Rankin-Selberg method and estimates of the second moment of Fourier coefficients; (iv) the sup norm problem of automorphic forms in the spectral aspect. We will focus on the analytic and arithmetic aspects of the theory of automorphic forms.

CONTENTS

Date: July 25, 2021.
Key words and phrases. Automorphic form, modular form, Eisenstein series, Maass form, L-function, Hecke operator, Rankin-Selberg method, Fourier coefficient, sup-norm.

0. Introduction

Automorphic forms are present in almost every area of modern number theory. They also appear in other areas of mathematics and in physics.

The concept of an automorphic function is a natural generalization of a periodic function. Let X be a locally compact space acted on discontinuously by a group Γ. Then a function $f : X \to \mathbb{C}$ is **automorphic** with respect to Γ if

$$f(\gamma x) = f(x) \quad \text{for all } \gamma \in \Gamma.$$

In other words, f lives on the quotient space $\Gamma\backslash X$ (the space of orbits).

Example 0.1. Take $X = \mathbb{R}$ and $\Gamma = \mathbb{Z}$, so $\mathbb{Z}\backslash\mathbb{R}$ is the circle. A function $f : \mathbb{R} \to \mathbb{C}$ is periodic of period 1 if $f(x+n) = f(x)$ for all $x \in \mathbb{R}$ and $n \in \mathbb{Z}$. We have the theory of Fourier series. For $f : \mathbb{R} \to \mathbb{C}$ a Schwartz function, we have (the Poisson summation formula)

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

where $\hat{f}(y) = \int_{\mathbb{R}} f(x)e(-xy)\mathrm{d}x$ is the Fourier transform of f. Here $e(z) = e^{2\pi i z}$.

Example 0.2. Take $X = \mathbb{R}^m$ and $\Gamma = \mathbb{Z}^m$, so $\mathbb{T}^m = \mathbb{Z}^m\backslash\mathbb{R}^m$ is the m-dimensional torus.

There are many interesting applications of the theory of automorphic forms, such as

- equidistribution of integral points on ellipsoids,
- equidistribution of quadratic roots,
- primes represented by $x^2 + y^4$, and so on.

Unfortunately, we won't have time to discuss those applications in this mini-course.

1. Modular forms

1.1. **The hyperbolic space.** As a model of the hyperbolic plane we will use the upper half of the plane $\mathbb{C}$ of complex numbers:

$$\mathbb{H} = \{z = x + iy : x \in \mathbb{R},\ y \in \mathbb{R}^+\}.$$

$\mathbb{H}$ is a riemannian manifold with the metric derived from the Poincare differential,

$$\mathrm{d}s^2 = y^{-2}(\mathrm{d}x^2 + \mathrm{d}y^2). \tag{1.1}$$

The distance function on $\mathbb{H}$ is given by

$$\rho(z, w) = \min_L \int_0^1 (x'(t)^2 + y'(t)^2)^{1/2} y(t)^{-1}\mathrm{d}t$$

where $L = \{(x(t), y(t)) : t \in [0, 1]\}$ ranges over smooth curves in $\mathbb{H}$ joining z and w. More explicitly we have

$$\rho(z, w) = \log \frac{|z - \bar{w}| + |z - w|}{|z - \bar{w}| - |z - w|}. \tag{1.2}$$

We have

$$\cosh \rho(z, w) = 1 + 2u(z, w), \tag{1.3}$$

where

$$u(z, w) = \frac{|z - w|^2}{4 \operatorname{Im} z \operatorname{Im} w}. \tag{1.4}$$

This function (a point-pair invariant) is more practical than the true distance function $\rho(z, w)$.

To describe the geometry of $\mathbb{H}$ we shall use well-known properties of the Möbius transformations

$$gz = \frac{az + b}{cz + d}, \quad a, b, c, d \in \mathbb{R},\ ad - bc = 1. \tag{1.5}$$

Observe that a Möbius transformation g determines the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ up to sign. In particular, both matrices $I = \begin{pmatrix} 1 & \\ & 1 \end{pmatrix}$ and $-I = \begin{pmatrix} -1 & \\ & -1 \end{pmatrix}$ give the identity transformation. We shall always take this distinction into account, but often without mention.

Throughout we denote $G = \mathrm{SL}_2(\mathbb{R})$, the group of real matrices of determinant 1. The group $\mathrm{PSL}_2(\mathbb{R}) = G/(\pm I)$ of all Möbius transformations acts on the whole compactified complex plane $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ (the Riemann sphere) as conformal mappings. A Möbius transformation g maps a euclidean circle onto a circle subject to the convention that the euclidean lines in $\hat{\mathbb{C}}$ are also circles. Of course, the centers may not be preserved, since g is not a euclidean isometry, save for $g = \pm \begin{pmatrix} 1 & * \\ & 1 \end{pmatrix}$, which represents a translation.

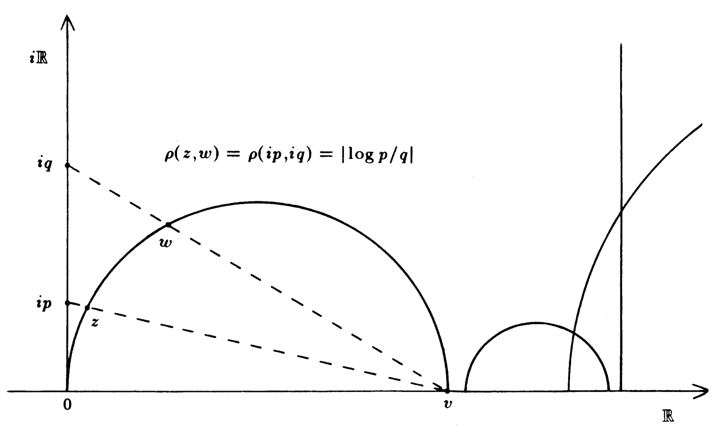

FIGURE 1. Geodesics in $\mathbb{H}$.

The riemannian measure derived from the Poincaré differential $\mathrm{d}s = y^{-1}|\mathrm{d}z|$ on $\mathbb{H}$ is expressed in terms of the Lebesgue measure simply by

$$\mathrm{d}\mu z = y^{-2}\mathrm{d}x\mathrm{d}y. \tag{1.6}$$

Note that

$$\frac{\mathrm{d}}{\mathrm{d}z} gz = \frac{a(cz + d) - (az + b)c}{(cz + d)^2} = \frac{1}{(cz + d)^2},$$

and

$$\operatorname{Im} gz = \frac{\operatorname{Im} z}{|cz + d|^2}.$$

Hence we have

$$(\operatorname{Im} gz)^{-1}|\mathrm{d}gz| = (\operatorname{Im} z)^{-1}|\mathrm{d}z|,$$

which shows that the differential $\mathrm{d}s$ on $\mathbb{H}$ is invariant under the group G.

Exercise 1.1. Show directly that the above measure is G-invariant.

Theorem 1.2. *The hyperbolic lines (geodesics in $\mathbb{H}$) are represented by the euclidean semicircles and half-lines orthogonal to $\mathbb{R}$.*

1.2. **The modular group and the fundamental domain.** The group $\mathrm{SL}_2(\mathbb{Z})$ is the first discrete subgroup of $\mathrm{SL}_2(\mathbb{R})$ which interests arithmeticians.

Definition 1.3. The group $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ is called **the modular group**.

Remark 1.4. Many authors may define Γ as the image of $\mathrm{SL}_2(\mathbb{Z})$ in $\mathrm{PSL}_2(\mathbb{R})$, *i.e.* $\Gamma = \mathrm{PSL}_2(\mathbb{Z}) = \mathrm{SL}_2(\mathbb{Z})/\{\pm I\}$.

Theorem 1.5. *The modular group is generated by two matrices*

$$T = \begin{pmatrix} 1 & 1 \\ & 1 \end{pmatrix}, \quad S = \begin{pmatrix} & -1 \\ 1 & \end{pmatrix}. \tag{1.7}$$

Proof. The action of S on $\mathbb{H}$ (the inversion) is involutary, more precisely $S^2 = -I$. We have

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

so S acts by interchanging the rows up to the sign. The matrix $T^n = \begin{pmatrix} 1 & n \\ & 1 \end{pmatrix}$ acts by translation,

$$T^n \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a + cn & b + dn \\ c & d \end{pmatrix}.$$

If $c \neq 0$ this has an effect of reducing the left upper entry to $0 \leq a < |c|$ by a suitable choice of $n \in \mathbb{Z}$. Applying both operations repeatedly, we end up with a matrix having $c = 0$, which must be of type $\pm \begin{pmatrix} 1 & m \\ & 1 \end{pmatrix}$. Then applying T^{-m} we arrive at $\pm I$. $\square$

Remark 1.6. The procedure described in the proof of Theorem 1.5 follows the steps of the continued fraction expansion of a/c.

Theorem 1.7. *The set*

$$D = \left\{ z = x + iy : |x| < \frac{1}{2}, |z| > 1 \right\} \tag{1.8}$$

is a ***fundamental domain*** *for the modular group* $\Gamma = \mathrm{SL}_2(\mathbb{Z})$, i.e. *it has the following properties:*

- *D is a domain in $\mathbb{H}$,*
- *every orbit of Γ has a point in D or on the boundary ∂D,*
- *distinct points in D are not in the same orbit of Γ.*

Proof. For any given $z \in \mathbb{H}$ and $B > 0$. We consider the set $\Sigma = \{\mathrm{Im}\, \gamma z : \gamma \in \Gamma, |cz+d| \leq B\}$. Since $|cz+d|^2 = (cx+d)^2+(cy)^2$, we have $|cy| \leq B$ and hence $|c| \leq B/y$. Note that $|cx+d| \leq B$, hence we have $|d| \leq B+|cx| \leq B+B|x|/y$. Thus we get $\#\Sigma \leq (2B/y+1)(B+B|x|/y+1)$, i.e., Σ is a finite set. So each orbit Γz has a point with largest height (the imaginary part). Such a point, say z, has the property that

$$|cz + d| \geq 1 \quad \text{for all } \gamma = \begin{pmatrix} * & * \\ c & d \end{pmatrix} \in \Gamma.$$

This maximal property is preserved by translations (which are represented by matrices with $c = 0$), so we can choose a maximal point of the orbit in the strip $|x| \leq \frac{1}{2}$. We shall show that the domain

$$D' = \left\{ z \in \mathbb{H} : |x| < \frac{1}{2},\ |cz + d| > 1 \text{ for all } c, d \text{ with } c \neq 0 \right\}$$

coincides with D. Indeed, $D' \subset D$ by choosing $c = 1, d = 0$. Conversely, if $z \in D$ and $c \neq 0$ then

$$|cz + d|^2 = c^2|z|^2 + 2cdx + d^2 > c^2 - |cd| + d^2 \geq 1,$$

so $D \subset D'$. From the construction of D' it follows that distinct points of D' are not equivalent and the closure $\bar{D}' = D' \cup \partial D'$ contains points of every orbit. Thus $D = D'$ is a fundamental domain. □

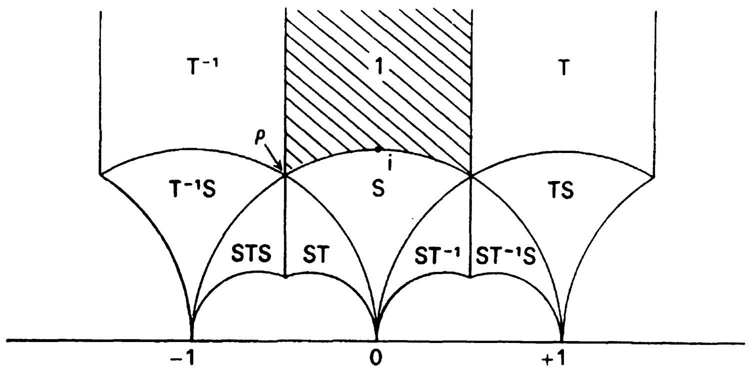

FIGURE 2. The fundamental domain D (shaded) and it's translates.

Let D be the standard fundamental polygon (1.8). We shall show that the equivalent points of the boundary ∂D are exactly those pairs of points which interchange upon reflection in the line $x = 0$, and they are identified by the transformations T and S. Indeed, if both z and γz are on ∂D then

$$1 = |cz + d|^2 = (cx + d)^2 + c^2y^2 \geq c^2 - |cd| + d^2 \geq 1;$$

hence either $c = 0, d = \pm 1$ or $c = \pm 1, d = 0$, giving T or S respectively. Observe that D has a parabolic vertex at ∞, an elliptic vertex at i of order $m(i) = 2$, and two equivalent elliptic vertices at $\rho = \frac{-1+i\sqrt{3}}{2}$ and $\rho' = \frac{1+i\sqrt{3}}{2}$ of order $m(\rho) = m(\rho') = 3$.

1.3. **Modular functions.**

Definition 1.8. A function $f : \mathbb{H} \to \mathbb{C}$ is called a **weakly modular function** of weight k if f is meromorphic on the half plane $\mathbb{H}$ and satisfies the transformation rule

$$f(\gamma z) = (cz + d)^k f(z) \quad \text{for all} \quad \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma. \tag{1.9}$$

Remark 1.9. (1) k must be even. If k is odd, then take $\gamma = -I$ we have

$$f\left(\begin{pmatrix} -1 & \\ & -1 \end{pmatrix} z \right) = (-1)^k f(z) = -f(z),$$

hence $f(z)$ vanishes identically. Hence from now on, we will assume that k is even.

(2) If f_1, f_2 are weakly modular functions of weight k_1, k_2 respectively, then $f_1 f_2$ is a weakly modular function of weight $k_1 + k_2$.

Let $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$. We have $\mathrm{d}(\gamma z)/\mathrm{d}z = (cz+d)^{-2}$. The relation (1.9) can then be written:

$$\frac{f(\gamma z)}{f(z)} = \left(\frac{\mathrm{d}(\gamma z)}{\mathrm{d}z}\right)^{-\frac{k}{2}}$$

or

$$f(\gamma z)\mathrm{d}(gz)^{k/2} = f(z)\mathrm{d}z^{k/2}.$$

It means that the "differential form of weight $k/2$" $f(z)\mathrm{d}z^{k/2}$ is invariant under Γ. Since Γ is generated by the elements S and T, it suffices to check the invariance by S and by T. This gives:

Theorem 1.10. *Let f be meromorphic on $\mathbb{H}$. The function f is a weakly modular function of weight k if and only if it satisfies the two relations:*

$$f(z+1) = f(z), \tag{1.10}$$

$$f(-1/z) = z^k f(z). \tag{1.11}$$

Proof. Let $j(\gamma, z) = cz + d$, where $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$. One can check

$$j(\gamma\gamma', z) = j(\gamma, \gamma' z) j(\gamma', z), \quad \text{for any } \gamma, \gamma' \in \Gamma.$$

Together with Theorem 1.5, we complete the proof. □

Suppose the relation (1.10) is verified. We can then express f as a function of $q = e^{2\pi i z}$, function which we will denote by $\tilde{f}$; it is meromorphic in the disk $|q| < 1$ with the origin removed. If $\tilde{f}$ extends to a meromorphic (resp. holomorphic) function at the origin, we say, by abuse of language, that f is **meromorphic** (resp. **holomorphic**) **at infinity**. This means that $\tilde{f}$ admits a Laurent expansion in a neighborhood of the origin

$$\tilde{f}(q) = \sum_{n=-\infty}^{+\infty} a_f(n) q^n$$

where the an are zero for n small enough (resp. for $n < 0$).

Definition 1.11. A weakly modular function is called a **modular function** if it is meromorphic at infinity.

When f is holomorphic at infinity, we set $f(\infty) = \tilde{f}(0)$. This is the value of f at infinity.

Definition 1.12. A modular function which is holomorphic everywhere (including infinity) is called a **modular form**; if such a function is zero at infinity, it is called a **cusp form** ("Spitzenform" in German and "forme parabolique" in French).

A modular form of weight k is thus given by a series

$$f(z) = \sum_{n=0}^{\infty} a_f(n) q^n = \sum_{n=0}^{\infty} a_f(n) e(nz) \tag{1.12}$$

which converges for $|q| < 1$ (i.e. for $\operatorname{Im}(z) > 0$), and which verifies the identity

$$f(-1/z) = z^k f(z).$$

It is a cusp form if $a_f(0) = 0$.

1.4. Eisenstein series and their Fourier expansion.

Definition 1.13. Let $k \geq 4$ be even. The kth **Eisenstein series** is defined by

$$G_k(z) = \sum_{(m,n)\neq(0,0)} \frac{1}{(mz+n)^k}, \quad z \in \mathbb{H}. \tag{1.13}$$

Theorem 1.14. *Let $k \geq 4$ be even.*

(i) *$G_k(z)$ is absolute convergent in $\mathbb{H}$ and is uniformly absolute convergent in any compact support;*
(ii) *$G_k(z)$ is a modular form of weight k;*
(iii) *the Fourier expansion of $G_k(z)$ is*

$$G_k(z) = 2\zeta(k) + 2\frac{(2\pi i)^k}{(k-1)!}\sum_{n=1}^{\infty} \sigma_{k-1}(n) e(nz), \tag{1.14}$$

where ζ is the Riemann zeta-function, and $\sigma_{k-1}(n) = \sum_{d|n} d^{k-1}$.

Proof. (i) Let $D_\varepsilon = \{z \in \mathbb{H} : -1/\varepsilon \leq x \leq 1/\varepsilon,\ \varepsilon \leq y \leq 1/\varepsilon\}$ for $0 < \varepsilon < 1$. We just need to prove that $G_k(z)$ is uniformly absolute convergent in D_ε for any given $0 < \varepsilon < 1$. Let $z = x + iy \in D_\varepsilon$, define

$$\lambda_z = \min_{(m,n)\neq(0,0)} |mz+n|$$

and

$$N_z(t) = \#\{(m,n) \in \mathbb{Z}^2 \setminus \{(0,0)\} : |mz+n| \leq t\}.$$

Then $\lambda_z = \min_{(m,n)\neq(0,0)} |mx+n+imy| \geq \min\{1,\varepsilon\} \geq \varepsilon$. If $|mz+n| \leq t$, then by $\varepsilon|m| \leq y|m| \leq |mx+n+imy| \leq t$ we have $|m| \leq t/\varepsilon$; and then by $|n| - |mx| \leq |mx+n| \leq |mx+n+imy| \leq t$ we have $|n| \leq t/\varepsilon^2 + t \leq \frac{2}{\varepsilon^2}t$. Hence we have $N_z(t) \leq \frac{8}{\varepsilon^3}t^2$. So

$$\begin{aligned}
|G_k(z)| &\leq \sum_{(m,n)\neq(0,0)} \frac{1}{|mz+n|^k} = \sum_{\substack{(m,n)\neq(0,0)\\ |mz+n|\leq 1}} \frac{1}{|mz+n|^k} + \sum_{|mz+n|>1} \frac{1}{|mz+n|^k} \\
&\leq \varepsilon^{-k}\frac{8}{\varepsilon^3} + \sum_{d\geq 1}\sum_{d<|mz+n|\leq d+1} \frac{1}{|mz+n|^k} \\
&\leq 8\varepsilon^{-k-3} + \sum_{d\geq 1} \frac{1}{d^k}(N_z(d+1) - N_z(d)) \\
&\leq 8\varepsilon^{-k-3} + \sum_{d\geq 2} \frac{N_z(d)}{(d-1)^k} - \sum_{d\geq 1} \frac{N_z(d)}{d^k} \\
&\ll_\varepsilon 1.
\end{aligned}$$

(ii) By Theorem 1.10, we just need to check that

$$G_k(z+1) = G_k(z), \quad \text{and} \quad G_k(-1/z) = z^k G_k(z).$$

(iii) We have

$$G_k(z)=\sum_{(m,n)\neq(0,0)}\frac{1}{(mz+n)^k}=2\zeta(k)+2\sum_{m=1}^{\infty}\sum_{n=-\infty}^{\infty}\frac{1}{(mz+n)^k}.$$

By the Poisson summation formula we have

$$\sum_{n=-\infty}^{\infty}\frac{1}{(mz+n)^k}=\sum_{n\in\mathbb{Z}}\int_{-\infty}^{\infty}\frac{1}{(mz+v)^k}e^{-2\pi inv}\mathrm{d}v.$$

Making a change of variable $w=u+iy=x+v/m+iy$ we have

$$\sum_{n=-\infty}^{\infty}\frac{1}{(mz+n)^k}=\frac{e(mnz)}{m^k}\sum_{n\in\mathbb{Z}}\int_{-\infty}^{\infty}\frac{1}{w^k}e^{-2\pi imnw}\mathrm{d}u.$$

By Cauchy integral theorem we have (for $k\geq 4$ and $m,y>0$)

$$\int_{-\infty+iy}^{\infty+iy}\frac{1}{w^k}e^{-2\pi imnw}\mathrm{d}w=\begin{cases}\frac{(-2\pi i)^k(mn)^{k-1}}{(k-1)!}, & \text{if } n\geq 1,\\ 0, & \text{otherwise.}\end{cases}$$

So we get

$$\sum_{n=-\infty}^{\infty}\frac{1}{(mz+n)^k}=\frac{(-2\pi i)^k}{(k-1)!}\sum_{n\geq 1}n^{k-1}e(mnz)$$

and (k is even)

$$\begin{aligned}G_k(z)&=2\zeta(k)+2\frac{(-2\pi i)^k}{(k-1)!}\sum_{m=1}^{\infty}\sum_{n\geq 1}n^{k-1}e(mnz)\\&=2\zeta(k)+2\frac{(2\pi i)^k}{(k-1)!}\sum_{n=1}^{\infty}\sigma_{k-1}(n)e(nz),\end{aligned}$$

as claimed.

Here is another proof of (iii): We begin by the well-known product representation for the sine function (which is easy to establish by comparing zeros and applying Liouville's theorem)

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

Take the logarithmic derivative

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

On the other hand, this is equal to

$$\pi\frac{\cos(\pi z)}{\sin(\pi z)}=\pi\frac{\frac{e^{i\pi z}+e^{-i\pi z}}{2}}{\frac{e^{i\pi z}-e^{-i\pi z}}{2i}}=\pi i\frac{e(z)+1}{e(z)-1}=\pi i+\frac{2\pi i}{e(z)-1}=\pi i-2\pi i\sum_{d=0}^{\infty}e(dz).$$

We differentiate these expansions $k-1$ times, getting

$$\begin{aligned} -(2\pi i)^k \sum_{d\ge 1} d^{k-1} e(dz) &= \frac{(-1)^{k-1}(k-1)!}{z^k} + \sum_{n=1}^{\infty}\left(\frac{(-1)^{k-1}(k-1)!}{(z+n)^k} + \frac{(-1)^{k-1}(k-1)!}{(z-n)^k}\right) \\ &= (-1)^{k-1}(k-1)! \sum_{n\in\mathbb{Z}} \frac{1}{(z+n)^k}, \end{aligned}$$

that is

$$\sum_{n\in\mathbb{Z}} \frac{1}{(z+n)^k} = \frac{(-2\pi i)^k}{(k-1)!} \sum_{d\ge 1} d^{k-1} e(dz) \tag{1.15}$$

for any $k \ge 2$. Hence we derive the Fourier expansion of $G_k(z)$ as follows

$$\begin{aligned} G_k(z) &= \sum_{(m,n)\neq(0,0)} \frac{1}{(mz+n)^k} = 2\zeta(k) + 2\sum_{m=1}^{\infty}\sum_{n=-\infty}^{\infty} \frac{1}{(mz+n)^k} \\ &= 2\zeta(k) + 2\frac{(2\pi i)^k}{(k-1)!}\sum_{m=1}^{\infty}\sum_{d=1}^{\infty} d^{k-1} e(dmz), \end{aligned}$$

and by collecting terms with $dm = n$ we arrive at

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

This completes the proof. □

Exercise 1.15. For $k \ge 4$ and $y > 0$, show that

$$\int_{-\infty}^{\infty} \frac{1}{w^k} e^{-2\pi i n w} \mathrm{d}w = \begin{cases} \frac{(-2\pi i)^k n^{k-1}}{(k-1)!}, & \text{if } n \ge 1, \\ 0, & \text{otherwise.} \end{cases}$$

Then we define the **normalized Eisenstein series** $E_k(z) = G_k(z)/2\zeta(k)$, which is

$$E_k(z) = \frac{1}{2} \sum_{\substack{m,n\in\mathbb{Z} \\ (m,n)=1}} \frac{1}{(mz+n)^k}. \tag{1.16}$$

By Theorem 1.14 we have

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

where B_k are the Bernoulli numbers and we have

$$\zeta(k) = -\frac{(2\pi i)^k}{2k!} B_k, \tag{1.17}$$

for $k \ge 4$ and $k \in 2\mathbb{Z}$.

1.5. **The Ramanujan τ function.** From the discriminant function we get a modular function of weight twelve:

$$\Delta(z) = g_2(z)^3 - 27g_3(z)^2$$

where $g_2(z) = 60G_4(z)$ and $g_3(z) = 140G_6(z)$.

Since

$$g_2(z) = \frac{(2\pi)^4}{12}\left[1 + 240\sum_{n=1}^{\infty}\sigma_3(n)e(nz)\right],$$

$$g_3(z) = \frac{(2\pi)^6}{216}\left[1 - 504\sum_{n=1}^{\infty}\sigma_5(n)e(nz)\right],$$

we find that in the Fourier expansion of $\Delta(z)$ the constant terms cancel out, and

$$\Delta(z) = (2\pi)^{12}\sum_{n=1}^{\infty}\tau(n)e(nz) \tag{1.18}$$

where $\tau(n) = 1$. The arithmetic function $\tau(n)$ is called the **Ramanujan function**. It possesses fascinating properties.

The first few values of $\tau(n)$ are given in the following table.

n	1	2	3	4	5	6	7	8
$\tau(n)$	1	-24	252	-1472	4830	-6048	-16744	84480
n	9	10	11	12	13	14	15	16
$\tau(n)$	-113643	-115920	534612	-370944	-577738	401856	1217160	987136

TABLE 1. Ramanujan τ function.

Theorem 1.16 (Mordell 1917)**.** *We have*

$$\tau(mn) = \tau(m)\tau(n) \quad \text{if } (m,n) = 1,$$

$$\tau(p^{n+1}) = \tau(p)\tau(p^n) - p^{11}\tau(p^{n-1}) \quad \text{for } p \text{ prime, } n \in \mathbb{N}.$$

Theorem 1.17 (Deligne 1973)**.** *We have*

$$|\tau(n)| \leq n^{11/2}d(n),$$

where $d(n)$ is the divisor function.

These was observed by Ramanujan in 1916, but did not prove.

Conjecture 1.18 (The Sato–Tate conjecture)**.** For p prime $\tau(p)/p^{11/2} =: 2\cos\theta_p$ with $\theta_p \in [0,\pi]$ is distributed like the trace of a random SU(2) matrix. That is, for $0 \leq \alpha < \beta \leq \pi$, we have

$$\lim_{x\to\infty}\frac{\#\{p \leq x : \alpha < \theta_p < \beta\}}{\#\{p \leq x\}} = \frac{2}{\pi}\int_{\alpha}^{\beta}(\sin\theta)^2 \mathrm{d}\theta.$$

This was proved by Barnet-Lamb, Geraghty, Harris, Taylor (2009).

Conjecture 1.19 (Lehmer's conjecture (1947))**.** $\tau(n) \neq 0$ for all $n \geq 1$.

This is still open.

1.6. **The linear space of modular forms.** Let k be an even number, $f : \mathbb{H} \to \mathbb{C}$ a modular function of weight k and $w \in \mathbb{H}$. Define $m_f(w)$ to be the **order** of f at w such that $h_f(z) = (z-w)^{-m_f(w)} f(z)$ is analytic at $z = w$ and $h_f(w) \neq 0$. Moreover we define $m_f(\infty)$ as the **order** for $q = 0$ of the function $\tilde{f}(q)$ associated to f.

Lemma 1.20. *Let k be an even number, $f : \mathbb{H} \to \mathbb{C}$ a modular function of weight k and $f \neq 0$. Then*
(i) $m_f(z) = m_f(\gamma z)$, *for any $\gamma \in \Gamma$ and any $z \in \mathbb{H}$;*
(ii) $m_f(z) = 0$ *except for finitely many points in $\overline{D}$.*

Theorem 1.21. *Let k be an even number, $f : \mathbb{H} \to \mathbb{C}$ a modular function of weight k and $f \neq 0$. Then*

$$\sum_{\substack{w \in \Gamma\backslash\mathbb{H} \\ w \neq i,\rho \pmod{\Gamma}}} m_f(w) + m_f(\infty) + \frac{1}{2} m_f(i) + \frac{1}{3} m_f(\rho) = \frac{k}{12}. \tag{1.19}$$

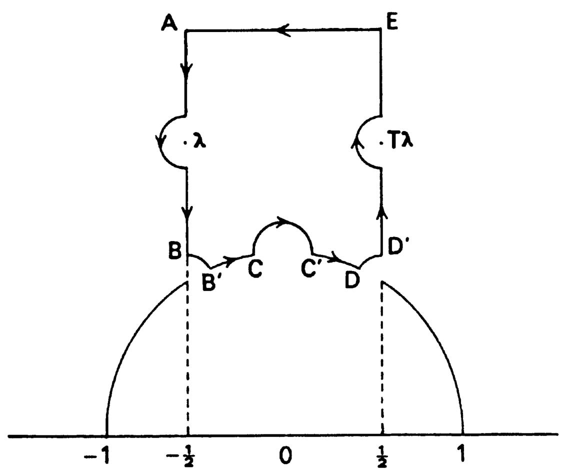

FIGURE 3.

Let $\mathcal{M}_k(\Gamma)$ denote the linear space of modular forms of weight k and $\mathcal{S}_k(\Gamma)$ the the linear space of cusp forms of weight k. Clearly modular forms of different weights are linearly independent over $\mathbb{C}$, so the space of all modular forms is the direct sum of the $\mathcal{M}_k(\Gamma)$,

$$\mathcal{M}(\Gamma) = \bigoplus_{k \geq 0} \mathcal{M}_k(\Gamma). \tag{1.20}$$

The whole space $\mathcal{M}(\Gamma)$ can also be considered as a graded algebra with respect to the inclusions

$$\mathcal{M}_k(\Gamma)\mathcal{M}_\ell(\Gamma) \subset \mathcal{M}_{k+\ell}(\Gamma).$$

We shall use the formula (1.19) to examine the structure of $\mathcal{M}(\Gamma)$ as an algebra over $\mathbb{C}$. Since $f \in \mathcal{M}(\Gamma)$ is holomorphic, all terms of (1.19) are nonnegative.

Theorem 1.22. (i) *Define*

$$\phi : \begin{cases} \mathcal{M}_{k-12}(\Gamma) & \to & \mathcal{S}_k(\Gamma), \\ f & \mapsto & \Delta f, \end{cases}$$

then ϕ is an isomorphism.

(ii) *If $k<0$ or $k=2$, then $\mathcal{M}_k(\Gamma)=\mathcal{S}_k(\Gamma)=\{0\}$.*

(iii) *If $k=0,4,6,8,10$, then $\mathcal{S}_k(\Gamma)=\{0\}$ and $\mathcal{M}_k(\Gamma)=\langle G_k\rangle$.*

(iv) *If $k\geq 12$, then*

$$\mathcal{M}_k(\Gamma)=\mathcal{S}_k(\Gamma)\oplus\mathbb{C}G_k,$$

that is, for any $f\in\mathcal{M}_k(\Gamma)$, there exist $f_0\in\mathcal{S}_k(\Gamma)$ and $c\in\mathbb{C}$ such that $f=f_0+cG_k$.

Corollary 1.23. *We have*

$$\dim\mathcal{M}_k(\Gamma)=\begin{cases}[k/12], & \text{if } k\equiv 2 \pmod{12},\ k\geq 0,\\ [k/12]+1, & \text{if } k\not\equiv 2 \pmod{12},\ k\geq 0,\end{cases} \tag{1.21}$$

and

$$\dim\mathcal{S}_k(\Gamma)=\begin{cases}[k/12]-1, & \text{if } k\equiv 2 \pmod{12},\ k\geq 0,\\ [k/12], & \text{if } k\not\equiv 2 \pmod{12},\ k\geq 0.\end{cases} \tag{1.22}$$

Exercise 1.24. Let $f,g\in\mathcal{M}_k(\Gamma)$, so that the first $[\frac{k}{12}]+1$ Fourier coefficients coincide (that is $f=\sum_{n\geq 0}a_f(n)q^n$, $g=\sum_{n\geq 0}a_g(n)q^n$, and $a_f(n)=a_g(n)$ for all $0\leq n\leq[\frac{k}{12}]$). Then $f=g$.

2. Bounds for Fourier coefficients, L-functions, and Hecke operators

2.1. Hecke's bound for Fourier coefficients of cusp forms. A key quantity associated to a modular form are its Fourier coefficients, that is the coefficients $a_f(n)$ of its q-expansion

$$f(z) = \sum_{n \geq 0} a_f(n) e(nz).$$

For instance, we computed the coefficients of the Eisenstein series $G_k(z)$, which for $n \neq 0$ are divisor sums $c_k \sigma_{k-1}(n)$. In particular, they are at least of size n^{k-1}. In fact, this is also an upper bound, since for $k-1>1$ we have

$$\sigma_{k-1}(n) = \sum_{d|n} d^{k-1} \ll n^{k-1}.$$

Exercise 2.1. Let $\sigma_s(n) = \sum_{d|n} d^s$ (the sum over all divisors of n).

a) Show that σ_s is multiplicative:

$$\sigma_s(mn) = \sigma_s(m)\sigma_s(n) \quad \text{if } (m,n)=1.$$

b) Show that for $s>1$, $\sigma_s(n) \ll n^s$.

As we shall see, the coefficients of cusp forms are much smaller, and that is a key input into the theory of modular forms. Those facts have important applications to the theory of representatives by quadratic forms.

Before beginning this study, we need some preparations.

Lemma 2.2. *Let $f \in \mathcal{S}_k$ be a cusp form. Then*

a) *$f(x+iy) \ll_f e^{-2\pi y}$ decays exponentially as $y \to +\infty$.*
b) *$F(z) = y^{k/2}|f(z)| \ll_f 1$ for any $z \in \mathbb{H}$.*

Proof. a) Using the Fourier expansion $f(z) = \sum_{n \geq 1} a_f(n) e(nz)$, which has no constant term when f is cuspidal. We see that $f(z) = O(|q|)$ as $q = e(z) \to 0$, which gives $f(x+iy) \ll_f e^{-2\pi y}$ as $y \to +\infty$.

b) Since $f \in \mathcal{S}_k$, we have

$$F(\gamma z) = (\operatorname{Im} \gamma z)^{k/2} |f(\gamma z)| = \left(\frac{\operatorname{Im} z}{|cz+d|^2} \right)^{k/2} |cz+d|^k |f(z)| = F(z).$$

So F is a Γ-invariant function. So

$$\sup_{z \in \mathbb{H}} F(z) = \sup_{z \in D} F(z),$$

where D is the standard fundamental domain of Γ. By a) we have $f(z) \ll e^{-2\pi y}$ for $y \geq \frac{\sqrt{3}}{2}$. Hence for any $z \in D$, we have $F(z) = y^{k/2}|f(z)| \ll_f 1$. This proves Lemma 2.2. □

For $f, g \in \mathcal{S}_k$, the expression $y^k f(z)\overline{g(z)}$ is Γ-invariant. Therefore we can define the inner product (due to H. Petersson)

$$\langle f, g \rangle = \int_{\Gamma \backslash \mathbb{H}} y^k f(z) \overline{g(z)} \mathrm{d}\mu z.$$

A cusp form f has exponential decay at cusp, so this inner product is absolutely convergent. For $f = g$ we set

$$\|f\|^2 = \langle f, g\rangle = \int_{\Gamma\backslash\mathbb{H}} y^k |f(z)|^2 \mathrm{d}\mu z < \infty.$$

The linear space of cusp forms $\mathcal{S}_k$ equipped with the Petersson inner product is a finite dimensional Hilbert space.

We can now present a bound for Fourier coefficients of cusp forms, due to Erich Hecke (1930's).

Theorem 2.3. *If f is cuspidal (and not identically zero), its Fourier coefficients satisfy $|a_f(n)| \leq Cn^{k/2}$ for some constant C independent of n.*

Proof. Since f is a cusp form, there exists a constant C_1 such that $|y^{k/2} f(z)| < C_1$ for all $z \in \mathbb{H}$. Now fixed $y > 0$, we have

$$a_f(n) e^{-2\pi n y} = \int_0^1 f(x+iy) e^{-2\pi i n x} \mathrm{d}x;$$

hence we get

$$|a_f(n)| < C_1 y^{-k/2} e^{2\pi n y}.$$

By taking $y = 1/n$, we obtain

$$|a_f(n)| < C_1 e^{2\pi} n^{k/2},$$

as required. □

This estimate, called the *trivial estimate*, is due to Hardy (1927) and (more simply) Hecke (1937). Using the theory of Poincaré series and estimates of Kloosterman sums, one can improve this esitmate. The correct estimate $a_f(n) \ll_{f,\varepsilon} n^{\frac{k-1}{2}+\varepsilon}$ for any $\varepsilon > 0$, was conjectured (for $f = \Delta$) by Ramanujan (1916); this famous statement, the *Ramanujan conjecture*, we finally proved around 1970 by Deligne using difficult techniques form algebraic geometry.

Corollary 2.4. *Let $f \in \mathcal{M}_k$ be a modular of weight $k > 2$, not necessarily cuspidal. Then the Fourier coefficients satisfy*

$$a_f(n) \ll_f n^{k-1}.$$

Proof. Writing $f = aG_k + g$ with $g \in \mathcal{S}_k$ cuspidal, we see that

$$a_f(n) = c\sigma_{k-1}(n) + a_g(n).$$

By Theorem 2.3 we have $a_g(n) \ll n^{k/2}$. The bound $\sigma_{k-1}(n) \ll n^{k-1}$ gives the result. □

2.2. **Sums of Fourier coefficients.** In this subsection, we consider some basic results on the average size of $\rho_f(n) = a_f(n)/n^{\frac{k-1}{2}}$. Note that we have Hecke's bound $\rho_f(n) \ll_f n^{1/2}$.

Theorem 2.5. *Let $f \in \mathcal{S}_k$. For any $N \geq 1$ we have*

$$\sum_{n \leq N} |\rho_f(n)|^2 \ll_f N. \tag{2.1}$$

Proof. Using the Fourier expansion $f(z) = \sum_{n\geq 1} \rho_f(n) n^{\frac{k-1}{2}} e(nz)$, we have

$$\int_0^1 |f(z)|^2 \mathrm{d}x = \int_0^1 \left| \sum_{n\geq 1} \rho_f(n) n^{\frac{k-1}{2}} e^{-2\pi ny} e(nx) \right|^2 \mathrm{d}x$$
$$= \sum_{n\geq 1} |\rho_f(n)|^2 n^{k-1} e^{-4\pi ny}.$$

By Lemma 2.2 we have $|f(z)| \ll_f y^{-k/2}$. Hence by taking $y = 1/N$, we have

$$\sum_{n\leq N} |\rho_f(n)|^2 n^{k-1} \ll \sum_{n\geq 1} |\rho_f(n)|^2 n^{k-1} e^{-4\pi n/N} \ll N^k.$$

By the partial summation formula we get

$$\sum_{n\leq N} |\rho_f(n)|^2 \ll N + \int_1^N \Big(\sum_{n\leq u} |\rho_f(n)|^2 n^{k-1} \Big) u^{-k} \mathrm{d}u \ll N.$$

This proves Theorem 2.2. □

Remark 2.6. This gives another proof of Hecke's bound $\rho_f(n) \ll_f n^{1/2}$.

Remark 2.7. The upper bound (2.1) is best possible, for one can prove by the Rankin–Selberg method in the next lecture, the more precise asymptotic formula

$$\sum_{n\leq N} |\rho_f(n)|^2 \sim c_f N,$$

as $N \to \infty$.

Theorem 2.8. *Let $f \in \mathcal{S}_k$. For any real α and $N \geq 1$, we have*

$$\sum_{n\leq N} \rho_f(n) e(n\alpha) \ll_f N^{1/2} \log 2N,$$

where the implied constant depends only on f (but not on α).

Proof. The Fourier coefficient are given by

$$a_f(n) = \int_0^1 f(z) e(-nz) \mathrm{d}x.$$

Hence the sum of coefficients twisted by the additive character $e(n\alpha)$ is equal to

$$\sum_{n\leq N} a_f(n) e(n\alpha) = \sum_{n\leq N} \int_0^1 f(z) e(-n(z-\alpha)) \mathrm{d}x = \int_0^1 f(z+\alpha) \sum_{n\leq N} e(-nz) \mathrm{d}x.$$

Note that

$$\sum_{n\leq N} e(-nz) = \frac{e(-Nz)-1}{1-e(z)} \ll \frac{e^{2\pi Ny}}{|1-e(z)|}.$$

Recall that $|f(z+\alpha)| \ll y^{-k/2}$. Note that

$$\int_0^1 \frac{1}{|1-e(z)|} \mathrm{d}x \ll \log(2+y^{-1}).$$

Indeed, this is obvious if $y \geq 1/10$. If $0 < y < 1/10$, then we have

- $|1-e(z)| \geq 1-e^{-2\pi y} \gg y$ for $0 \leq x \leq y$; and
- $|1-e(z)| > e^{-2\pi y}\sin(2\pi x) \gg x$ for $y < x \leq 1/2$.

Hence

$$\int_0^1 \frac{1}{|1-e(z)|}\mathrm{d}x = 2\int_0^{1/2} \frac{1}{|1-e(z)|}\mathrm{d}x \ll 1 + \int_y^{1/2} \frac{1}{x}\mathrm{d}x \ll \log(2+y^{-1}).$$

Thus we obtain

$$\sum_{n\leq N} a_f(n)e(n\alpha) \ll y^{-k/2}e^{2\pi N y}\log(2+y^{-1}),$$

where y is arbitrary positive number. For $y = 1/N$, we get

$$\sum_{n\leq N} a_f(n)e(n\alpha) \ll_f N^{k/2}\log 2N.$$

By the partial summation formula we get

$$\sum_{n\leq N} \rho_f(n)e(n\alpha) \ll_f N^{1/2}\log 2N + \int_2^N \Big|\sum_{n\leq u} a_f(n)e(n\alpha)\Big| u^{-\frac{k+1}{2}}\mathrm{d}u$$
$$\ll_f N^{1/2}\log 2N + \int_2^N u^{-\frac{1}{2}}\log u\,\mathrm{d}u \ll_f N^{1/2}\log 2N,$$

as claimed. □

Remark 2.9. This upper bound is almost sharp. Note that

$$\int_0^1 \Big|\sum_{n\leq N} \rho_f(n)e(n\alpha)\Big|^2 \mathrm{d}\alpha = \sum_{n\leq N} |\rho_f(n)|^2 \sim c_f N,$$

as in Remark 2.7. One has $c_f \neq 0$. Hence we have

$$\sum_{n\leq N} \rho_f(n)e(n\alpha) \gg \sqrt{N},$$

for some $\alpha \in [0,1]$.

We shall state the following beautiful result without giving a proof.

Theorem 2.10. *Let $f \in \mathcal{S}_k$. For any $N \geq 1$, we have*

$$\sum_{n\leq N} \rho_f(n) \ll_f N^{1/3},$$

where the implied constant depends on f.

This result shall be compared with the Gauss circle problem and Dirichlet divisor problem.

Exercise 2.11. Let $d(n)$ be the divisor function. Show that

$$\sum_{n\leq N} d(n) = N\log N + (2\gamma-1)N + O(N^{1/3}\log N),$$

where γ is Euler's constant.

2.3. **Hecke L-functions.** A connection between automorphic forms and L-functions can be traced in the celebrated memoir of B. Riemann (1860) on the zeta-function

$$\zeta(s)=\sum_{n=1}^{\infty}\frac{1}{n^s}=\prod_p\left(1-\frac{1}{p^s}\right)^{-1},\quad \mathrm{Re}(s)>1.$$

The gamma function $\Gamma(s)$ is defined as

$$\Gamma(s)=\int_0^{\infty}e^{-x}x^s\frac{\mathrm{d}x}{x},\quad \mathrm{Re}(s)>0.$$

Riemann established the functional equation

$$\xi(s):=\pi^{-s/2}\Gamma(s/2)\zeta(s)=\xi(1-s).$$

Let $f(z)=\sum_{n=0}^{\infty}\rho_f(n)n^{\frac{k-1}{2}}e(nz)$ be an element of $\mathcal{M}_k$. Let

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

This is known as the L-function of f. We need to know that this series is convergent for s sufficiently large. For this, the following eestimate is sufficient.

Theorem 2.12. *Let $f\in\mathcal{S}_k(\Gamma)$. The L-function $L(s,f)$ has analytic continuation to all s and satisfies a functional equation. In fact, if*

$$\Lambda(s,f)=(2\pi)^{-s}\Gamma\left(s+\frac{k-1}{2}\right)L(s,f)$$

then $\Lambda(s,f)$ extends to an analytic function of s. It satisfies

$$\Lambda(s,f)=i^k\Lambda(1-s,f).$$

Proof. Because f is cuspidal, $f(iy)\to 0$ very rapidly as $y\to\infty$. When $\gamma=S=\begin{pmatrix} & -1\\ 1 & \end{pmatrix}$, (1.9) implies that

$$f(iy)=i^{-k}y^{-k}f(i/y),\tag{2.2}$$

so $f(iy)\to 0$ very rapidly as $y\to 0$ also. Hence the integral

$$I(s)=\int_0^{\infty}f(iy)y^{s+\frac{k-1}{2}}\frac{\mathrm{d}y}{y}$$

is convergent for all s and clearly defines an analytic function of s. If $\mathrm{Re}(s)$ is large, we may substitute the Fourier expansion for f. Noting that

$$\int_0^{\infty}e^{-2\pi ny}y^{s+\frac{k-1}{2}}\frac{\mathrm{d}y}{y}=(2\pi)^{-s-\frac{k-1}{2}}n^{-s-\frac{k-1}{2}}\Gamma\left(s+\frac{k-1}{2}\right),$$

we get

$$I(s)=(2\pi)^{-\frac{k-1}{2}}\Lambda(s,f).\tag{2.3}$$

Now by (2.2) and making a change of variable $u=1/y$, we have

$$I(s)=\int_0^{\infty}i^{-k}y^{-k}f(i/y)y^{s+\frac{k-1}{2}}\frac{\mathrm{d}y}{y}=i^{-k}\int_0^{\infty}f(iu)u^{-s+\frac{k+1}{2}}\frac{\mathrm{d}u}{u}=i^kI(1-s).\tag{2.4}$$

Here we have used the fact $i^k=i^{-k}$ as k being even. By (2.3) and (2.4) we get $\Lambda(s,f)=i^k\Lambda(1-s,f)$. □

2.4. **Hecke operators.** It was observed long ago that the Fourier coefficients of basic classical modular forms have remarkable arithmetical properties. A fascinating example is the multiplicativity of the Ramanujan τ-function,

$$\tau(m)\tau(n) = \sum_{d|(m,n)} d^{11}\tau\left(\frac{mn}{d^2}\right). \tag{2.5}$$

This formula was first established by E. Hecke (1936) by means of certain self-adjoint operators. The theory of Hecke operators explains numerous other identities. More important, Hecke's original ideas proved to be vital for developments of modern fields such as Galois representations.

In a general setting the Hecke operators are averaging operators over a suitable finite collection of double cosets with respect to a group: therefore a great deal of the Hecke theory belongs to linear algebra. But when one considers the spectral analysis of these operators the problems become more delicate, and complete results are known only for arithmetic groups. In this chapter we shall present the theory of Hecke operators in the context of the modular group. For economy of exposition we replace the double coset constructions with specific representatives.

Throughout we assume that k is a fixed integer and $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ is the modular group. For any $A = \begin{pmatrix} * & * \\ c & d \end{pmatrix} \in \mathrm{GL}_2^+(\mathbb{R})$ we introduce the function $j(A,z) = cz+d$. The slash operator is defined on functions $f : \mathbb{H} \to \mathbb{C}$ by

$$f_{|A}(z) = (\det A)^{k/2} j(A,z)^{-k} f(Az).$$

Since k is an integer, the slash operator is associative:

$$f_{|AB} = (f_{|A})_{|B}.$$

Let n be a positive integer. The nth Hecke operator T_n is defined as

$$T_n f(z) = \frac{1}{\sqrt{n}} \sum_{\gamma \in \Gamma \backslash \Gamma_n} f_{|\gamma}(z),$$

where

$$\Gamma_n = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a,b,c,d \in \mathbb{Z},\ ad-bc = n \right\}.$$

Lemma 2.13. *The set*

$$S_n = \left\{ \begin{pmatrix} a & b \\ & d \end{pmatrix} : ad = m,\ 0 \le b < d \right\}$$

forms a complete set of right coset representatives of Γ_n *modulo* Γ, i.e. *we have the disjoint partition*

$$\Gamma_n = \bigcup_{\rho \in S_n} \Gamma\rho.$$

Proof. Let $\rho = \begin{pmatrix} a & * \\ c & * \end{pmatrix} \in \Gamma_n$. Let $\gamma = c/(a,c)$ and $\delta = -c/(a,c)$, we get $\gamma a + \delta c = 0$ and $(\gamma,\delta) = 1$. So there exists $\tau = \begin{pmatrix} * & * \\ \gamma & \delta \end{pmatrix} \in \Gamma$ which gives $\tau\rho = \begin{pmatrix} * & * \\ & * \end{pmatrix}$. Changing the sign of

τ if necessary, we get $\tau\rho = \begin{pmatrix} a & b \\ & d \end{pmatrix}$ with $ad = n$ and $d > 0$. Finally, multiplying on the left by $\begin{pmatrix} 1 & u \\ & 1 \end{pmatrix}$ with a suitable $u \in \mathbb{Z}$, we can take $0 \le b < d$. This proves that $\Gamma_n \subseteq \bigcup_{\rho \in S_n} \Gamma\rho$.

We now checke the cosets $\Gamma\rho$ are disjoint as ρ ranges over S_n. If

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} a & b \\ & d \end{pmatrix} = \begin{pmatrix} a' & b' \\ & d' \end{pmatrix},$$

we have $\gamma = 0$, $\alpha = \delta = 1$ and $\beta = 0$. This completes the proof of the lemma. □

By Lemma 2.13 we shall write T_n formally as

$$T_n = n^{-\frac{k+1}{2}} \sum_{ad=n} a^k \sum_{b \bmod d} \begin{pmatrix} a & b \\ & d \end{pmatrix}. \tag{2.6}$$

This expression means that for any function $f : \mathbb{H} \to \mathbb{C}$

$$(T_n f)(z) = n^{-\frac{k+1}{2}} \sum_{ad=n} a^k \sum_{b \bmod d} f\left(\frac{az+b}{d}\right).$$

Lemma 2.14. *There exists a one-to-one correspondence between $S_n \times \Gamma$ and $\Gamma \times S_n$,* i.e. *for any $\rho \in S_n$ and $\tau \in \Gamma$ there exist unique $\tau' \in \Gamma$ and $\rho' \in S_n$ such that*

$$\rho\tau = \tau'\rho'.$$

For a fixed $\tau \in \Gamma$, as ρ ranges over the whole set S_n, ρ' does also.

Proof. Exercise. □

Theorem 2.15. *The Hecke operator T_n maps a modular form to a modular form and a cusp form to a cusp form:*

$$T_n : \mathcal{M}_k(\Gamma) \to \mathcal{M}_k(\Gamma),$$
$$T_n : \mathcal{S}_k(\Gamma) \to \mathcal{S}_k(\Gamma).$$

Proof. By the correspondence $\rho\tau = \tau'\rho'$ we get

$$(T_n f)_{|\tau} = \frac{1}{\sqrt{n}} \sum_{\rho \in \Gamma\backslash\Gamma_n} f_{|\rho\tau} = \frac{1}{\sqrt{n}} \sum_{\rho' \in \Gamma\backslash\Gamma_n} f_{|\rho'} = T_n f,$$

since $f_{|\rho\tau} = f_{|\tau'\rho'} = f_{|\rho'}$. This proves $T_n : \mathcal{M}_k(\Gamma) \to \mathcal{M}_k(\Gamma)$.

If $f \in \mathcal{S}_k(\Gamma)$, then $f(\infty) = 0$. Since for $ad = n$, $d > 0$ and $b \in \mathbb{Z}$, we have $\begin{pmatrix} a & b \\ & d \end{pmatrix} \infty = \infty$. So $T_n f(\infty) = 0$. Hence $T_n : \mathcal{S}_k(\Gamma) \to \mathcal{S}_k(\Gamma)$. □

Theorem 2.16. *For any $m, n \ge 1$ we have*

$$T_m T_n = \sum_{d|(m,n)} T_{mn/d^2}.$$

Proof. By (2.6) we have

$$T_m T_n = (mn)^{-\frac{k+1}{2}} \sum_{\substack{a_1 d_1 = m \\ a_2 d_2 = n}} (a_1 a_2)^k \sum_{\substack{b_1 \bmod d_1 \\ b_2 \bmod d_2}} \begin{pmatrix} a_1 & b_1 \\ & d_1 \end{pmatrix} \begin{pmatrix} a_2 & b_2 \\ & d_2 \end{pmatrix},$$

and after multiplying we reduce the resulting matrix by $\delta = (a_1, d_2)$ to get

$$T_mT_n = (mn)^{-\frac{k+1}{2}} \sum_{\substack{\delta|(m,n)\\(a_1,d_2)=1}} \delta^k \sum_{\substack{a_1d_1=m/\delta\\a_2d_2=n/\delta}} (a_1a_2)^k \sum_{\substack{b_1 \bmod d_1\\b_2 \bmod d_2\delta}} \begin{pmatrix} a_1a_2 & a_1b_2+b_1d_2\\ & d_1d_2\end{pmatrix}.$$

Given a_1, a_2, d_1, d_2 as above, the upper-right entry $b = a_1b_2 + b_1d_2$ covers every class modulo d_1d_2 exactly δ times when b_1, b_2 range over all classes modulo $d_1, d_2\delta$ respectively. Indeed, b determines $b_2 \equiv b\bar{a}_1 \pmod{d_2}$ and $b_1 \equiv (b - a_1b_2)/d_2 \pmod{d_1}$. Therefore we have

$$T_mT_n = (mn)^{-\frac{k+1}{2}} \sum_{\substack{\delta|(m,n)\\(a_1,d_2)=1}} \delta^{k+1} \sum_{\substack{a_1d_1=m/\delta\\a_2d_2=n/\delta}} (a_1a_2)^k \sum_{b \bmod d_1d_2} \begin{pmatrix} a_1a_2 & b\\ & d_1d_2\end{pmatrix}.$$

Take $a = a_1a_2$ and $d = d_1d_2$, so $ad = mn/\delta^2$. Conversely, given a factorization $ad = mn/\delta^2$, there exist unique factorization $a = a_1a_2$, $d = d_1d_2$ with $(a_1, d_2) = 1$ $a_1d_1 = m/\delta$ and $a_2d_2 = n/\delta$; indeed $a_1 = m/(m, \delta d)$ and $d_2 = \delta d/(m, \delta d)$. Hence we can write

$$T_mT_n = \sum_{\delta|(m,n)} \left(\frac{mn}{\delta^2}\right)^{-\frac{k+1}{2}} \sum_{ad=mn/\delta^2} a^k \sum_{b \bmod d} \begin{pmatrix} a & b\\ & d\end{pmatrix}$$

which proves Theorem 2.16. □

Corollary 2.17. *The Hecke operators commute:*

$$T_mT_n = T_nT_m.$$

Theorem 2.18. *The Hecke operator T_n acting on the space of cusp forms for the modular group are self-adjoint,* i.e.

$$\langle T_nf, g\rangle = \langle f, T_ng\rangle, \quad \textit{for all } f, g \in \mathcal{S}_k(\Gamma).$$

Proof. One proof of this result uses Poincaré series which we won't introduce in these notes. See Iwaniec [4, §6.4] for details. □

Theorem 2.19 (Hecke)**.** *In the space $\mathcal{S}_k(\Gamma)$ of cusp forms for the modular group there exists an orthonormal basis H_k which consists of eigenfunctions of all the Hecke operators T_n.*

Proof. See Iwaniec [4, §6.4]. □

2.4.1. *Hecke eigenvalues vs. Fourier coefficients.*

Proposition 2.20. *Suppose f is given by the Fourier series*

$$f(z) = \sum_{m=0}^{\infty} \rho_f(m)m^{\frac{k-1}{2}}e(mz),$$

which converges absolutely in $\mathbb{H}$. Then T_nf is given by the series

$$(T_nf)(z) = \sum_{m=0}^{\infty} \rho_n(m)m^{\frac{k-1}{2}}e(mz),$$

whose coefficients are

$$\rho_n(m) = \sum_{d|(m,n)} \rho_f\left(\frac{mn}{d^2}\right).$$

Proof. Applying (2.6) we get

$$\begin{aligned}(T_nf)(z) &= n^{-\frac{k+1}{2}} \sum_{ad=n} a^k \sum_{b \bmod d} \sum_{m=0}^{\infty} \rho_f(m) m^{\frac{k-1}{2}} e\left(m\frac{az+b}{d}\right) \\ &= n^{-\frac{k+1}{2}} \sum_{ad=n} a^k \sum_{m=0}^{\infty} \rho_f(m) m^{\frac{k-1}{2}} e\left(m\frac{az}{d}\right) \sum_{b \bmod d} e\left(m\frac{b}{d}\right) \\ &= \sum_{ad=n} \sum_{\ell=0}^{\infty} \rho_f(d\ell)(a\ell)^{\frac{k-1}{2}} e\left(a\ell z\right) \\ &= \sum_{m=0}^{\infty} \left(\sum_{\substack{ad=n \\ a\ell=m}} \rho_f(d\ell) \right) m^{\frac{k-1}{2}} e(mz)\end{aligned}$$

as claimed. □

Let $f \in H_k$ be a Hecke eigenform, and

$$T_nf = \lambda_f(n)f \quad \text{for} \quad n \in \mathbb{N}. \tag{2.7}$$

Suppose f has the Fourier expansion

$$f(z) = \sum_{m=1}^{\infty} \rho_f(m) m^{\frac{k-1}{2}} e(mz).$$

Comparing the mth Fourier coefficients on the both sides of (2.7), together with Proposition 2.20, we get

$$\lambda_f(n)\rho_f(m) = \rho_n(m) = \sum_{d|(m,n)} \rho_f\left(\frac{mn}{d^2}\right).$$

For $m = 1$ this gives

$$\rho_f(n) = \rho_f(1)\lambda_f(n).$$

Hence $\rho_f(1) \neq 0$, as otherwise f would vanish identically. Therefore the Fourier coefficients of a Hecke cusp form are proportional to the eigenvalues of the Hecke operators.

As an example, consider the space $\mathcal{S}_{12}(\Gamma)$, which is one-dimensionally spanned by $\Delta(z)$; therefore $\Delta(z)$ is automatically a simultaneous eigenfunction of all the Hecke operators, namely

$$T_n\Delta(z) = \tau(n) n^{-\frac{11}{2}} \Delta(z),$$

where $\tau(n)$ is the Ramanujan function. By Theorem 2.16 we prove (2.5).

We may adjust any Hecke eigenform by setting the constant $\rho_f(1) = 1$. Such a Hecke eigenform will be called *(Hecke) normalized*. We see that $\mathcal{S}_k$ has a basis of normalized Hecke eigenforms. We now show that the L-function of a Hecke eigenform has an Euler product.

Theorem 2.21. *If $f \in H_k$ is a normalized Hecke eigenform, then*

$$L(s,f) = \sum_{n=1}^{\infty} \frac{\lambda_f(n)}{n^s} = \prod_{p=2}^{\infty} \left(1 - \lambda_f(p)p^{-s} + p^{-2s}\right)^{-1}.$$

Proof. It follows from the multiplicativity of the coefficients that

$$L(s,f) = \prod_{p=2}^{\infty} \left(\sum_{j=0}^{\infty} \frac{\lambda_f(p^j)}{p^{js}} \right).$$

By the Hecke relation, for $j \geq 1$ we have

$$\lambda_f(p^{j+1}) - \lambda_f(p)\lambda_f(p^j) + \lambda_f(p^{j-1}) = 0.$$

Hence

$$\begin{aligned}
&\left(1 - \lambda_f(p)p^{-s} + p^{-2s}\right) \left(\sum_{j=0}^{\infty} \frac{\lambda_f(p^j)}{p^{js}} \right) \\
&\qquad = 1 + \frac{\lambda_f(p)}{p^s} - \frac{\lambda_f(p)}{p^s} + \sum_{j=2}^{\infty} \frac{\lambda_f(p^j) - \lambda_f(p)\lambda_f(p^{j-1}) + \lambda_f(p^{j-2})}{p^{js}} = 1.
\end{aligned}$$

Hence we obtain

$$L(s,f) = \prod_{p=2}^{\infty} \left(1 - \lambda_f(p)p^{-s} + p^{-2s}\right)^{-1},$$

as claimed. □

3. The Rankin–Selberg method and the second moment of Fourier coefficients

The Rankin–Selberg method, which originated independently in the papers of Rankin (1939) and Selberg (1940), seeks to represent an L-function as an integral of one or more automorphic forms against an Eisenstein series, itself a type of automorphic form. The Eisenstein series itself has a functional equation, and so if the L-function can be represented as such an integral, it inherits this functional equation.

3.1. Eisenstein series.

Let

$$\Gamma_\infty = \left\{ \pm \begin{pmatrix} 1 & m \\ & 1 \end{pmatrix} : m \in \mathbb{Z} \right\}$$

be the subgroup of Γ which fixes $i\infty$.

Definition 3.1. Let $z \in \mathbb{H}$ and $\mathrm{Re}(s) > 1$. The *Eisenstein series* is defined as

$$E(z,s) = \sum_{\gamma \in \Gamma_\infty \backslash \Gamma} \mathrm{Im}(\gamma z)^s = \frac{1}{2} \sum_{\substack{c,d \in \mathbb{Z} \\ (c,d)=1}} \frac{y^s}{|cz+d|^{2s}}. \tag{3.1}$$

Let $\Delta = -y^2 \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)$ be the hyperbolic Laplacian.

Proposition 3.2. *We have $E(\gamma z, s) = E(z,s)$ for all $\gamma \in \Gamma$ and $\Delta E(z,s) = s(1-s)E(z,s)$.*

Proposition 3.3. *The Eisenstein series have the Fourier expansion*

$$E(z,s) = y^s + \phi(s) y^{1-s} + \frac{2\pi^s \sqrt{y}}{\Gamma(s)\zeta(2s)} \sum_{n \in \mathbb{Z}\backslash\{0\}} \sigma_{1-2s}(|n|)|n|^{s-1/2} K_{s-1/2}(2\pi|n|y) e(nx),$$

where

$$\phi(s) = \frac{\xi(2-2s)}{\xi(2s)}, \quad \xi(s) = \pi^{-s/2}\Gamma(s/2)\zeta(s),$$

$$\sigma_s(n) = \sum_{d|n} d^s,$$

and

$$K_s(y) = \frac{1}{2} \int_0^\infty \exp\left(-\frac{1}{2} y \Big(u + \frac{1}{u} \Big) \right) u^{s-1} \mathrm{d}u$$

the K-Bessel function.

Theorem 3.4. *$E(z,s)$ can be continued to meromorphic functions on $\mathbb{C}$. The modified Eisenstein series $E^*(z,s) = \xi(2s)E(z,s)$ is regular except for simple poles at $s = 0, 1$ and satisfies the functional equation*

$$E^*(z,s) = E^*(z,1-s).$$

We have

$$E(x+iy,s) \ll_s y^c \quad \text{as } y \to \infty,$$

where $c = \max(\mathrm{Re}(s), 1-\mathrm{Re}(s))$. Furthermore, the residue of the pole at $s=1$ is given by

$$\mathrm{Res}_{s=1} E(z,s) = \frac{3}{\pi}, \quad \text{for all } z \in \mathbb{H}.$$

The proofs of those results can be found in Goldfeld [3, §3.1] for example.

3.2. **The Rankin–Selberg method.** We are now ready to consider the Rankin–Selberg method. Let ϕ be an automorphic function on $\mathbb{H}$, that is, a smooth function satisfying $\phi(\gamma z) = \phi(z)$ for $\gamma \in \Gamma = \mathrm{SL}_2(\mathbb{Z})$ and $z \in \mathbb{H}$. Let us suppose that

$$\phi(x+iy) = O(y^{-N}) \quad \text{for all } N > 0 \text{ as } y \to \infty. \tag{3.2}$$

Because $\phi(z+1) = \phi(z)$, we have a Fourier expansion

$$\phi(z) = \sum_{n\in\mathbb{Z}} \phi_n(y) e(nx), \quad \phi_n(y) = \int_0^1 \phi(x+iy) e(-nx) \mathrm{d}x.$$

We naturally call ϕ_0 the constant term in the Fourier expansion of ϕ. Let

$$\widetilde{\phi}_0(s) = \int_0^\infty \phi_0(y) y^s \frac{\mathrm{d}y}{y} \tag{3.3}$$

be the Mellin transform of ϕ_0. With our assumption, ϕ is bounded on the fundamental domain; hence ϕ_0 is bounded as a function of y and decays rapidly as $y \to \infty$. Hence the integral in (3.3) is absolutely convergent if $\mathrm{Re}(s) > 0$. For $\mathrm{Re}(s) > 1$, we define

$$\Lambda(s;\phi) = \pi^{-s}\Gamma(s)\zeta(2s)\widetilde{\phi}_0(s-1).$$

Proposition 3.5. *With these hypotheses, we have*

$$\Lambda(s;\phi) = \pi^{-s}\Gamma(s)\zeta(2s) \int_{\Gamma\backslash\mathbb{H}} E(z,s)\phi(z) \frac{\mathrm{d}x\mathrm{d}y}{y^2}. \tag{3.4}$$

Then $\Lambda(s;\phi)$ has meromorphic continuation to all s, with at most simple poles at $s = 1$ and $s = 0$. We have

$$\mathrm{Res}_{s=1} \Lambda(s;\phi) = \frac{1}{2} \int_{\Gamma\backslash\mathbb{H}} \phi(z) \frac{\mathrm{d}x\mathrm{d}y}{y^2}.$$

It satisfies the functional equation $\Lambda(s;\phi) = \Lambda(1-s;\phi)$.

This statement, with its characteristic "unfolding" proof, contains the essence of the Rankin–Selberg method.

Proof. Assume $\mathrm{Re}(s) > 1$. By (3.1) we have

$$\begin{aligned}\int_{\Gamma\backslash\mathbb{H}} E(z,s)\phi(z)\frac{\mathrm{d}x\mathrm{d}y}{y^2} &= \int_{\Gamma\backslash\mathbb{H}} \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} \mathrm{Im}(\gamma z)^s \phi(\gamma z) \mathrm{d}\mu(\gamma z) \\ &= \int_{\Gamma_\infty\backslash\mathbb{H}} y^s \phi(z) \frac{\mathrm{d}x\mathrm{d}y}{y^2}.\end{aligned}$$

We may take the integration over any fundamental domain for Γ_∞. We choose the fundamental domain defined by $0 < x < 1, y > 0$. So we get

$$\int_{\Gamma\backslash\mathbb{H}} E(z,s)\phi(z)\frac{\mathrm{d}x\mathrm{d}y}{y^2} = \int_0^\infty y^{s-1} \int_0^1 \phi(z)\mathrm{d}x \frac{\mathrm{d}y}{y} = \widetilde{\phi}_0(s-1).$$

This proves (3.4).

By Theorem 3.4 and the assumption (3.2) we know that $\Lambda(s;\phi)$ has meromorphic continuation to all s, with at most simple poles at $s=1$ and $s=0$, and

$$\begin{aligned}\operatorname{Res}_{s=1}\Lambda(s;\phi) &= \operatorname{Res}_{s=1}\pi^{-s}\Gamma(s)\zeta(2s)\int_{\Gamma\backslash\mathbb{H}} E(z,s)\phi(z)\frac{\mathrm{d}x\mathrm{d}y}{y^2}\\ &= \pi^{-1}\frac{\pi^2}{6}\int_{\Gamma\backslash\mathbb{H}}\operatorname{Res}_{s=1}E(z,s)\phi(z)\frac{\mathrm{d}x\mathrm{d}y}{y^2} = \frac{1}{2}\int_{\Gamma\backslash\mathbb{H}}\phi(z)\frac{\mathrm{d}x\mathrm{d}y}{y^2}.\end{aligned}$$

This completes the proof of Proposition 3.5. □

The original application of the Rankin–Selberg method was the result that if $f(z) = \sum_{n\geq 1} a_f(n)e(nz)$ and $g(z) = \sum_{n\geq 1} a_g(n)e(nz)$ are cusp forms of weight k, then the Dirichlet series $\sum_{n\geq 1} a_f(n)\overline{a_g(n)}n^{-s}$ has analytic continuation and functional equation. This was discovered independently by Rankin (1939) and Selberg (1940).

We apply Proposition 3.5 with $\phi(z) = y^k f(z)\overline{g(z)}$, where $f, g\in\mathcal{S}_k$. Clearly ϕ satisfies the assumption (3.2); ϕ is automorphic with respect to Γ. Indeed

$$\phi(\gamma z) = \operatorname{Im}(\gamma z)^k f(\gamma z)\overline{g(\gamma z)} = y^k|cz+d|^{-2k}(cz+d)^k f(z)\overline{(cz+d)^k g(z)} = \phi(z).$$

We have

$$\begin{aligned}\phi_0(y) &= \int_0^1 y^k f(z)\overline{g(z)}\mathrm{d}x\\ &= y^k\sum_{m\geq 1}\rho_f(m)m^{\frac{k-1}{2}}e^{-2\pi my}\sum_{n\geq 1}\overline{\rho_g(n)}n^{\frac{k-1}{2}}e^{-2\pi ny}\int_0^1 e((m-n)x)\mathrm{d}x.\end{aligned}$$

Since $\int_0^1 e((m-n)x)\mathrm{d}x = o$ unless $m=n$, we get

$$\phi_0(y) = \sum_{n\geq 1}\rho_f(n)\overline{\rho_g(n)}n^{k-1}e^{-4\pi ny}y^k.$$

Note that

$$\int_0^\infty e^{-4\pi ny}y^{k+s-1}\frac{\mathrm{d}y}{y} = (4\pi n)^{-k-s+1}\int_0^\infty e^{-u}u^{s+k-1}\frac{\mathrm{d}u}{u} = (4\pi n)^{-k-s+1}\Gamma(s+k-1).$$

Hence we obtain

$$\begin{aligned}\widetilde{\phi}_0(s-1) &= \sum_{n\geq 1}\rho_f(n)\overline{\rho_g(n)}n^{k-1}\int_0^\infty e^{-4\pi ny}y^{k+s-1}\frac{\mathrm{d}y}{y}\\ &= (4\pi)^{-k-s+1}\Gamma(s+k-1)\sum_{n\geq 1}\rho_f(n)\overline{\rho_g(n)}n^{-s}.\end{aligned}$$

Thus

$$\begin{aligned}\Lambda(s;\phi) &= \pi^{-s}\Gamma(s)\zeta(2s)(4\pi)^{-k-s+1}\Gamma(s+k-1)\sum_{n\geq 1}\rho_f(n)\overline{\rho_g(n)}n^{-s}\\ &= (4\pi)^{-k+1}(2\pi)^{-2s}\Gamma(s)\Gamma(s+k-1)\zeta(2s)\sum_{n\geq 1}\rho_f(n)\overline{\rho_g(n)}n^{-s}.\end{aligned}$$

Now we define

$$L(s, f\times\bar{g}) = \zeta(2s)\sum_{n\geq 1}\rho_f(n)\overline{\rho_g(n)}n^{-s},$$

$$\Lambda(s, f \times \bar{g}) = (2\pi)^{-2s}\Gamma(s)\Gamma(s+k-1)L(s, f \times \bar{g}).$$

Note that

$$\Lambda(s, f \times \bar{g}) = (4\pi)^{k-1}\Lambda(s; \phi). \tag{3.5}$$

Theorem 3.6. *Let $f, g \in \mathcal{S}_k$. Then $\Lambda(s, f \times \bar{g})$, originally defined for $\mathrm{Re}(s)$ sufficiently large, has meromorphic continuation to all s, with holomorphic except for at most simple poles at $s = 0$ and $s = 1$. It satisfies a functional equation*

$$\Lambda(s, f \times \bar{g}) = \Lambda(1-s, f \times \bar{g}).$$

We have

$$\mathrm{Res}_{s=1}\,\Lambda(s, f \times \bar{g}) = \frac{1}{2}(4\pi)^{k-1}\langle f, g\rangle.$$

Proof. This is a simple consequence of (3.5) and Proposition 3.5. □

3.3. **The second moment of Fourier coefficients.** In this section we will prove the following asymptotic formula for the second moment of Fourier coefficients.

Theorem 3.7. *We have*

$$\sum_{n\leq X} |\rho_f(n)|^2 = C_f X + O(X^{3/5+\varepsilon}),$$

for any $\varepsilon > 0$.

Lemma 3.8 (Stirling's formula)**.** *For fixed $\sigma \in \mathbb{R}$ and real $t \geq 10$, we have*

$$\Gamma(\sigma+it) = e^{-\frac{\pi}{2}|t|}|t|^{\sigma-\frac{1}{2}}\exp\left(it\log\frac{|t|}{e}\right)(2\pi)^{1/2}i^{\sigma-1/2}\left\{1+O\left(\frac{1}{t}\right)\right\}.$$

Lemma 3.9. *We have*

$$L(-\varepsilon+it, f \times \bar{g}) = \left(\frac{i|t|}{2\pi}\right)^{2+4\varepsilon}\exp\left(-4it\log\frac{|t|}{2\pi e}\right)L(1+\varepsilon-it, f \times \bar{g})\left\{1+O\left(\frac{1}{t}\right)\right\}.$$

Proof. This follows from Theorem 3.6 and Stirling's formula. □

To prove Theorem 3.7, we consider the average of $\lambda_{f\times\bar{f}}(n) = \sum_{n=\ell^2 m}|\rho_f(m)|^2$.

Theorem 3.10. *We have*

$$\sum_{n\leq X}\lambda_{f\times\bar{f}}(n) = c_f X + O(X^{3/5+\varepsilon}),$$

for any $\varepsilon > 0$.

Proof. We first approximate $\sum_{n\leq X}\lambda_{f\times\bar{f}}(n)$ by a smooth sum. Let

$$Y = X^{\delta}, \quad \text{for some } \delta \in (1/2, 1).$$

Let W be a smooth function with support $\mathrm{supp}(W) \in [1/2-Y/X,\, 1+Y/X]$ such that $W(u) = 1, u \in [1/2,\, 1]$ and $W(u) \in [0,\, 1], u \in [1/2-Y/X,\, 1/2] \cup [1,\, 1+Y/X]$, and $W^{(m)}(u) \ll_m$

$(X/Y)^m$ for all $m \geq 1$. Therefore we have the approximating formula

$$\begin{aligned}\sum_{X/2<n\leq X} \lambda_{f\times\bar f}(n) = &\sum_{X/2-Y<n<X+Y} \lambda_{f\times\bar f}(n) W\Big(\frac{n}{X}\Big)\\ &+ O\bigg(\sum_{X/2-Y<n<X/2} |\lambda_{f\times\bar f}(n)| + \sum_{X<n<X+Y} |\lambda_{f\times\bar f}(n)|\bigg) \qquad (3.6)\\ &= \sum_{n\geq 1} \lambda_{f\times\bar f}(n) W\Big(\frac{n}{X}\Big) + O\big(Y^{1+\varepsilon}\big)\end{aligned}$$

where we have used Deligne's bound $\lambda_{f\times\bar f}(n) \ll \sum_{\ell^2 m=n} m^\varepsilon \ll n^\varepsilon$ when f is holomorphic for the error terms.

Next we only need to show

$$\sum_{n\geq 1} \lambda_{f\times\bar f}(n) W\Big(\frac{n}{X}\Big) = \mathrm{Res}_{s=1} L(s, f\times\bar f)\tilde W(1)X + O\big(X^{3/5+\varepsilon}\big) \qquad (3.7)$$

where $\tilde W(s) = \int_0^\infty W(x)x^{s-1}\mathrm{d}x$ is the Mellin transform of W. By breaking the sum into dyadic intervals and by inserting (3.7) into (3.6), we get

$$\sum_{n\leq X} \lambda_{f\times\bar f}(n) = 2\,\mathrm{Res}_{s=1} L(s, f\times\bar f)\tilde W(1)X + O\big(X^{3/5+\varepsilon}\big).$$

Then Theorem 3.10 follows immediately from the estimate $\tilde W(1) = 1/2 + O(Y/X)$.

Now we estimate the sum $\sum_{n\geq 1} \lambda_{f\times\bar f}(n) W\big(\frac{n}{X}\big)$ in (3.7). By the inverse Mellin transform

$$W(u) = \frac{1}{2\pi i}\int_{(2)} \tilde W(s) u^{-s}\mathrm{d}s,$$

we get

$$\sum_{n\geq 1} \lambda_{f\times\bar f}(n) W\Big(\frac{n}{X}\Big) = \frac{1}{2\pi i}\int_{(2)} \tilde W(s) L(s, f\times\bar f) X^s\mathrm{d}s.$$

We then move the integration to the parallel segment with $\mathrm{Re}\, s = \sigma = -\varepsilon$. We pass the pole at $s = 1$ with residue $\mathrm{Res}_{s=1} L(s, f\times\bar f)$. Hence we obtain

$$\sum_{n\geq 1} \lambda_{f\times\bar f}(n) W\Big(\frac{n}{X}\Big) = \mathrm{Res}_{s=1} L(s, f\times\bar f)\tilde W(1)X + \frac{1}{2\pi i}\int_{(-\varepsilon)} \tilde W(s) L(s,\ f\times\bar f) X^s\mathrm{d}s. \quad (3.8)$$

We denote by $I(X)$ the second term with integration on the right hand side of (3.8). Inserting a dyadic smooth partition of unity to the t-integral, we get

$$I(X) = \sum_{T\,\mathrm{dyadic}} I(X, T) \qquad (3.9)$$

where

$$I(X, T) := \frac{X^{-\varepsilon}}{2\pi}\int_{\mathbb{R}} X^{it}\tilde W(-\varepsilon+it) L(-\varepsilon+it,\ f\times\bar f) V\Big(\frac{t}{T}\Big)\mathrm{d}t$$

for some fixed V with compact support. For $\tilde{W}(s)$, by integration by parts, we have the estimate for any $m \geq 1$

$$\tilde{W}(s) = \frac{(-1)^m}{s(s+1)\cdots(s+m-1)} \int_0^\infty W^{(m)}(u) u^{s+m-1} \mathrm{d}u \ll_m \frac{1}{|s|^m} \Big(\frac{X}{Y}\Big)^{m-1}, \tag{3.10}$$

since $\mathrm{supp}(W^{(m)}) \in [1/2 - Y/X,\, 1/2] \cup [1,\, 1+Y/X]$. This estimate allows us to truncate the t-integral of $I(X,T)$ at $t \ll X^{1+\varepsilon}/Y$. In addition, by the upper bounds $L(-\varepsilon+it,\, f\times\bar{f}) \ll (1+|t|)^{2+\varepsilon}$ that follows from Lemma 3.9 and the Phragmén–Lindelöf principle and by (3.10) with $m=1$, we deduce that

$$I(X,T) \ll X^\varepsilon T^{2+\varepsilon} \ll Y$$

if $T \ll Y^{1/2-\varepsilon}$. Therefore, by the above arguments, we may impose a constraint $Y^{1/2-\varepsilon} \ll T \ll X^{1+\varepsilon}/Y$ in (3.9) with an admissible error term. We only consider positive T's, since negative T's can be handled similarly. Next, for $I(X,T)$, by the first equality in (3.10) with $m=1$, we get

$$\begin{aligned} I(X,T) &= -\frac{X^{-\varepsilon}}{2\pi} \int_{1/3}^3 W'(u) u^{-\varepsilon} \int_{\mathbb{R}} \frac{(Xu)^{it}}{-\varepsilon+it} L(-\varepsilon+it,\, f\times\bar{f}) V\Big(\frac{t}{T}\Big) \mathrm{d}t\, \mathrm{d}u \\ &\ll \frac{X^{-\varepsilon}}{T} \sup_{u\in[1/3,\,3]} \Big| \int_{\mathbb{R}} (Xu)^{it} L(-\varepsilon+it,\, f\times\bar{f}) V\Big(\frac{t}{T}\Big) \mathrm{d}t \Big|. \end{aligned} \tag{3.11}$$

Hence, in the following, we only need to consider $J(X,T)$ which is defined by

$$J(X,T) := \int_{\mathbb{R}} X^{it} L(-\varepsilon+it,\, f\times\bar{f}) V\Big(\frac{t}{T}\Big) \mathrm{d}t. \tag{3.12}$$

The trivial bound for $J(X,T)$ is $O(T^{3+\varepsilon})$. To get a better estimate for $J(X,T)$, we shall apply functional equation for $L(-\varepsilon+it,\, f\times\bar{f})$ to change the variable $s=-\varepsilon+it$ into $1-s = 1+\varepsilon-it$.

By inserting the functional equation (Lemma 3.9) into (3.12), it follows that

$$\begin{aligned} J(X,T) &= \int_{\mathbb{R}} X^{it} \left(\frac{i|t|}{2\pi}\right)^{2+4\varepsilon} \exp\left(-4it\log\frac{|t|}{2\pi e}\right) L(1+\varepsilon-it, f\times\bar{f}) V\Big(\frac{t}{T}\Big) \mathrm{d}t \\ &\quad + O\Big(\frac{1}{T}\cdot T^{2+\varepsilon}\cdot T\Big) \\ &\ll T^{2+\varepsilon} \Big| \int_{\mathbb{R}} \sum_{n\geq 1} \frac{\lambda_{f\times\bar{f}}(n)}{n^{1+\varepsilon-it}} X^{it} \Big(\frac{t}{2\pi \mathrm{e}}\Big)^{-4it} V_1\Big(\frac{t}{T}\Big) \mathrm{d}t \Big| + T^{2+\varepsilon}, \end{aligned}$$

for some smooth compactly supported function V_1.

Changing the order of the integral of summation above, and making a change of variable $t = T\xi$, we get

$$J(X,T) \ll T^{3+\varepsilon} \Big| \sum_{n\geq 1} \frac{\lambda_{f\times\bar{f}}(n)}{n^{1+\varepsilon}} \int_{\mathbb{R}} V_1(\xi) e^{ih(\xi)} \mathrm{d}\xi \Big| + T^{2+\varepsilon},$$

where

$$h(\xi) := T\xi \log\left(\frac{nX}{(\frac{T\xi}{2\pi e})^4}\right) \quad \text{and} \quad h'(\xi) = 4T\log\frac{2\pi (nX)^{1/4}}{T\xi}.$$

If the n's above are such that $2\pi(nX)^{1/4}/T \notin \mathrm{supp}(V_1)$, then it is not difficult to see that $h'(\xi) \gg T^{\varepsilon}$ which would imply that the integral over ξ is $O(T^{-2021})$ upon using repeated integration by parts. Now for the above integral over ξ, we consider the case where $2\pi(nX)^{1/4}/T \in \mathrm{supp}(V_1)$. By the second derivative test with

$$h'(\xi_0) = 0, \quad \xi_0 = \frac{2\pi(nX)^{1/4}}{T}, \quad h(\xi_0) = 4T\xi_0, \quad h''(\xi_0) = -\frac{4T}{\xi_0} \asymp T,$$

we get

$$\int_{\mathbb{R}} V_1(\xi) e^{iT\xi \log(nX(\frac{T\xi}{2\pi e})^{-4})} \mathrm{d}\xi \ll \frac{1}{\sqrt{T}},$$

and hence

$$\begin{aligned} J(X,T) &\ll T^{5/2+\varepsilon} \Bigg| \sum_{n \asymp T^4/X} \frac{\lambda_{f\times\bar{f}}(n)}{n^{1+\varepsilon}} \Bigg| + T^{2+\varepsilon} \\ &\ll T^{5/2+\varepsilon}. \end{aligned} \tag{3.13}$$

Hence $I(X,T) \ll T^{3/2+\varepsilon} \ll X^{3/2+\varepsilon}/Y^{3/2}$. Taking $Y = X^{3/5}$, we get $I(X,T) \ll X^{3/5+\varepsilon}$. This completes the proof. □

Proof of Theorem 3.7. Since $\lambda_{f\times\bar{f}}(n) = \sum_{n=\ell^2 m} |\rho_f(m)|^2$, we have

$$|\rho_f(n)|^2 = \sum_{\ell^2 m = n} \mu(\ell)\lambda_{f\times\bar{f}}(m).$$

Hence we obtain

$$\begin{aligned} \sum_{n\le X} |\rho_f(n)|^2 &= \sum\sum_{\ell^2 m \le X} \mu(\ell)\lambda_{f\times\bar{f}}(m) \\ &= \sum_{\ell \le X^{1/2}} \mu(\ell) \left(c_f \frac{X}{\ell^2} + O(X^{3/5+o(1)}\ell^{-6/5}) \right) \\ &= \frac{c_f}{\zeta(2)} X + O(X^{3/5+o(1)}), \end{aligned}$$

as claimed. □

4. The sup-norm problem of Maass forms

The eigenfunctions of the Laplace operator on a Riemannian manifold are of great interest for theoretical physicists working in quantum mechanics. The square-integrable eigenstates are particularly meaningful. How do they behave on high energy levels – that is, in the limit with respect to the eigenvalues? Do they concentrate onto specific submanifolds, or sets such as closed geodesics when being on distinguished energy levels, and if so, what is the distribution law for these levels? For physicists, if the individual eigenfunctions behave like random waves, this is a manifestation of quantum chaos.

A simpler question, yet not easy to answer, is how large the eigenfunctions can possibly be in terms of the spectrum. The case of the torus $\mathbb{Z}^2\backslash\mathbb{R}^2$ shows that all eigenfunctions of the standard basis are uniformly bounded. But this is not true on other manifolds such as the sphere S^2, on which the eigenfunctions given by the Legendre polynomials (spherical harmonics) take relatively large values at special points. However, this phenomenon seems to be much weaker if the manifold is negatively curved, such as the quotient space $\Gamma\backslash\mathbb{H}$ of the hyperbolic plane modulo a finite volume group.

More precisely, we consider the torus $\mathbb{Z}^2\backslash\mathbb{R}^2$. The Laplace operator on $\mathbb{Z}^2\backslash\mathbb{R}^2$ is $\Delta = -(\partial_x^2 + \partial_y^2)$. By the Fourier theory we know that $\psi_{m,n}(x,y) = e(mx+ny)$ with $m,n \in \mathbb{Z}$ form a standard basis of $L^2(\mathbb{Z}^2\backslash\mathbb{R}^2)$. Here $e(z) = e^{2\pi i z}$. The function $\psi_{m,n}$ is an eigenfunction of Δ with eigenvalue $4\pi^2(m^2+n^2)$. Clearly we have $\|\psi_{m,n}\|_\infty = 1$.

Berry [1] suggested that eigenfunctions for chaotic systems are modeled by random waves. In particular, one would like to compare the sup-norm $\|\phi\|_\infty$ of an L^2-normalized eigenfunction ϕ with the corresponding quantity for random waves, which grows very slowly, as $\sqrt{\log \lambda_\phi}$, if λ_ϕ is the corresponding eigenvalue (see Salem–Zygmund [8, Ch. IV]). The bound $\|\phi\|_\infty \ll \lambda_\phi^{1/4}$ is valid on any compact Riemannian surface (see Seeger–Sogge [10]), which is sharp for standard 2-sphere. However, this bound is not optimal for most surfaces. Especially, Sarnak [9] conjectured that, for compact surfaces of negative curvature, $\|\phi\|_\infty \ll \lambda_\phi^\varepsilon$ for all $\varepsilon > 0$.

In this lecture note, we will mainly follow Iwaniec [5, Ch. 13] to see how to break the standard upper bound for Maass cusp forms on the modular group.

4.1. Maass cusp forms.

We shall study the vector space $L^2(\Gamma\backslash\mathbb{H})$ (defined over $\mathbb{C}$) which is the completion of the subspace consisting of all smooth functions $f : \Gamma\backslash\mathbb{H} \to \mathbb{C}$ satisfying the L^2 condition

$$\int_{\Gamma\backslash\mathbb{H}} |f(z)|^2 \frac{\mathrm{d}x\mathrm{d}y}{y^2} < \infty.$$

The space $L^2(\Gamma\backslash\mathbb{H})$ is actually a Hilbert space with inner product given by

$$\langle f, g\rangle := \int_{\Gamma\backslash\mathbb{H}} f(z)\overline{g(z)} \frac{\mathrm{d}x\mathrm{d}y}{y^2}$$

for all $f, g \in L^2(\Gamma\backslash\mathbb{H})$. This inner product was first introduced by Petersson.

On the hyperbolic plane $\mathbb{H}$ the *Laplace operator* derived from the differential $\mathrm{d}s^2 = y^{-2}(\mathrm{d}x^2 + \mathrm{d}y^2)$ is given by

$$\Delta = -y^2\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right). \tag{4.1}$$

Definition 4.1. Let $\lambda \in \mathbb{C}$. A **Maass cusp form** for Γ with eigenvalue λ is a non-zero function $f \in L^2(\Gamma\backslash\mathbb{H})$ which satisfies

i) $f(\gamma z) = f(z)$, for all $\gamma \in \Gamma$ and $z \in \mathbb{H}$,
ii) $\Delta f = \lambda f$,
iii) $\int_0^1 f(z)\mathrm{d}x = 0$.

Since Δ is self-adjoint, so we have $\lambda \geq 0$. We write $\lambda = 1/4 + t_f^2$, where $t_f \in [0, \infty) \cup \{iv : v \in [0, 1/2]\}$ is called the spectral parameter of f. Let f be a Maass cusp form with the spectral parameter t_f for Γ. Note that we have $f(z+1) = f(z)$. Then we have the Fouier–Whittaker expansion

$$f(z) = \rho_f(1) \sum_{n \neq 0} \lambda_f(n) \sqrt{2\pi y} K_{it_f}(2\pi |n| y) e(nx).$$

Here $\{\lambda_f(n)\}$ are called the Fourier–Whittaker coefficients of f, normalized so that $\lambda_f(1) = 1$. We have the famous Ramanujan–Petersson conjecture for Maass forms: $|\lambda_f(n)| \ll d(n)$, where $d(n) = \sum_{d|n} 1$ denotes the number of divisors of n. We at least have the following bounds: $\lambda_f(n) \ll |n|^{1/2}$. We also have the Selberg eigenvalue conjecture: $\lambda_1 \geq 1/4$, or equivalently, $t_f \in [0, \infty)$ if f is not a constant. This is known for the modular group $\mathrm{SL}(2, \mathbb{Z})$, but not known in general (for congruence subgroups). Let V_λ be subspace of all $f \in L^2(\Gamma\backslash\mathbb{H})$ which are Maass forms of eigenvalue λ. Maass proved the for any $\lambda > 0$ the space V_λ is finite dimensional.

4.2. **Spectral decomposition.** Our main goal of this section is the Selberg spectral decomposition for $\mathrm{SL}(2, \mathbb{Z})$ which states that

$$L^2(\Gamma\backslash\mathbb{H}) = \mathbb{C} \oplus L^2_{\mathrm{cusp}}(\Gamma\backslash\mathbb{H}) \oplus L^2_{\mathrm{cont}}(\Gamma\backslash\mathbb{H}),$$

where $\mathbb{C}$ is the one-dimensional space of constant functions, $L^2_{\mathrm{cusp}}(\Gamma\backslash\mathbb{H})$ represents the Hilbert space of square integrable functions on $\mathbb{H}$ whose constant term is zero, and $L^2_{\mathrm{cont}}(\Gamma\backslash\mathbb{H})$ represents all square integrable functions on $\mathbb{H}$ which are representable as integrals of the Eisenstein series. The reason for the terminology L^2_{cusp}, L^2_{cont} is because the classical definition of cusp form, introduced by Hecke, requires that the constant term in the Fourier expansion around any cusp (a real number equivalent to $i\infty$ under the discrete group) be zero, and also because the Eisenstein series is in the continuous spectrum of the Laplace operator. The latter means that $\Delta E(z, s) = s(1-s)E(z, s)$, or that $s(1-s)$ is an eigenvalue of Δ for any complex number s. Let $u_j(z)$, $(j = 1, 2, \dots)$ be an orthonormal basis of Maass forms for Γ. We shall also adopt the convention that $u_0(z) = \sqrt{\frac{3}{\pi}}$ is the constant function of norm 1. The Selberg spectral decomposition is given in the following theorem.

Theorem 4.2 (Selberg spectral decomposition). *Let $f \in L^2(\Gamma\backslash\mathbb{H})$. Then we have*

$$f(z) = \sum_{j \geq 0} \langle f, u_j \rangle u_j(z) + \frac{1}{4\pi} \int_{\mathbb{R}} \langle f, E(\cdot, 1/2 + it) \rangle E(z, 1/2 + it) \mathrm{d}t.$$

We also have the following Weyl's law due to Selberg which shows the existence of Maass cusp forms.

Theorem 4.3 (Selberg Weyl's law). *Let $N(T) := \#\{j : |t_j| \leq T\}$. Then we have*

$$N(T) = \frac{|D|}{4\pi} T^2 + O(T \log T).$$

4.3. **Automorphic kernel and pre-trace formula.** A function $k(z,w)$ with this property

$$k(gz, gw) = k(z,w), \quad \text{for all } g \in G.$$

is called *point-pair invariant*; it depends solely on the hyperbolic distance between the points. Consequently, we can set

$$k(z,w) = k(u(z,w)),$$

where $k(u)$ is a function in one variable $u > 0$ and $u(z,w)$ is given by (1.4). The automorphic kernel $K(z,w)$ is defined by

$$K(z,w) = \sum_{\gamma\in\Gamma} k(z, \gamma w).$$

The Selberg/Harish-Chandra transform in the following three steps:

$$\begin{aligned} q(v) &= \int_v^{+\infty} k(u)(u-v)^{-1/2}\mathrm{d}u, \\ g(r) &= 2q\left(\left(\sinh\frac{r}{2}\right)^2\right), \\ h(t) &= \int_{-\infty}^{+\infty} e^{irt} g(r)\mathrm{d}r. \end{aligned} \tag{4.2}$$

It is simpler to express the sufficient conditions in terms of $h(t)$ rather than $k(u)$. These conditions are:

$$\begin{aligned} &h(t) \text{ is even,} \\ &h(t) \text{ is holomorphic in the strip } |\operatorname{Im} t| < \frac{1}{2} + \varepsilon, \\ &h(t) \ll (|t|+1)^{-2-\varepsilon} \text{ in the strip.} \end{aligned} \tag{4.3}$$

For any h having the above properties, one finds the inverse of the Selberg/ Harish-Chandra transform in the following three steps:

$$\begin{aligned} g(r) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{irt} h(t)\mathrm{d}t, \\ q(v) &= \frac{1}{2} g\left(2\log(\sqrt{v+1}+\sqrt{v})\right), \\ k(u) &= -\frac{1}{\pi}\int_u^{+\infty} (v-u)^{-1/2}\mathrm{d}q(v). \end{aligned} \tag{4.4}$$

The projection of $K(z,w)$ on the eigenfunctions are as follows:

$$\langle K(\cdot, w), u_j\rangle = h(t_j)\overline{u_j(w)},$$
$$\langle K(\cdot, w), E(\cdot, 1/2+it)\rangle = h(t)\overline{E(w, 1/2+it)}.$$

Theorem 4.4 (The pre-trace formula). *Let $K(z,w)$ be an automorphic kernel given by a point-pair invariant $k(z,w) = k(u(z,w))$ whose Selberg/Harish-Chandra transform $h(t)$ satisfies the conditions* (4.3)*. Then it has the spectral expansion*

$$K(z,w) = \sum_{j\geq 0} h(t_j)u_j(z)\overline{u_j(w)} + \frac{1}{4\pi}\int_{\mathbb{R}} h(t)E(z,1/2+it)\overline{E(w,1/2+it)}\mathrm{d}t, \tag{4.5}$$

which converges absolutely and uniformly on compacta.

4.4. **The sup-norm problem.** In this section, we will prove the following theorem.

Theorem 4.5. *We have*

$$|u_j(z)| \ll \lambda_j^{1/4}.$$

Here the implied constant may depend on z.

To prove this theorem we will need bounds for weight functions and lattice points counting.

Lemma 4.6. *Let $T \geq 10$ and*

$$h(t) = 4\pi^2 \frac{\cosh(\pi t/2)\cosh(\pi T/2)}{\cosh \pi t + \cosh \pi T}. \tag{4.6}$$

Then we have $h(t) > 0$ everywhere and $h(t) \asymp 1$ if $t = T + O(1)$. The Fourier transform of $h(t)$ is equal to

$$g(x) = 2\pi \frac{\cos xT}{\cosh x}.$$

The Selberg/Harish-Chandra transform satisfies

$$k(0) = T + O(1), \tag{4.7}$$

$$k(u) \ll T^{1/2} u^{-1/4} (1+u)^{-5/4}. \tag{4.8}$$

Lemma 4.7. *For any $X \geq 2$, we have*

$$\#\{\gamma \in \Gamma : 4u(\gamma z, z) + 2 \leq X\} \ll X,$$

where the implied constant depending on z.

We will not prove the above lemmas. See [7, Lemma 1.1 and Pages 317&318].

Proof of Theorem 4.5. We should prove this by restricting the spectral averaging to a short interval. For this purpose we need the complete spectral decomposition of an automorphic kernel

$$K(z,z) = \sum_{\gamma \in \Gamma} k(u(z, \gamma z)) = \sum_{j \geq 0} h(t_j)|u_j(z)|^2 + \frac{1}{4\pi} \int_{\mathbb{R}} h(t)|E(z, 1/2 + it)|^2 \mathrm{d}t.$$

By Lemmas 4.6 and 4.7 we obtain

$$K(z,z) = \nu k(0) + \sum_{\substack{\gamma \in \Gamma \\ \gamma z \neq z}} k(u(z, \gamma z)) = \nu T + O(T^{1/2}),$$

where $\nu = \#\{\gamma \in \Gamma : \gamma z = z\} \in \{1, 2, 3\}$. Thus we have

$$\sum_{T < t_j < T+1} |u_j(z)|^2 \ll T,$$

and hence

$$u_j(z) \ll T^{1/2},$$

as claimed. □

Remark 4.8. The above result is related to the local Weyl law. The same bound can be established for eigenfunctions on any compact Riemann surface, which is sharp for standard 2-sphere. The implied constant must depend on z. Indeed, for some z with $\operatorname{Im} z = t_j/2\pi + o(1)$, we have

$$u_j(z) \gg \lambda_j^{1/12-\varepsilon}.$$

Remark 4.9. From the above proof, we find that it is possible to improve the bound in this case, since the two upper bounds νT and $O(T^{1/2})$ are not the same. One may try to transfer some mass from the "diagonal" terms ($\gamma z = z$) to the "non-diagonal" terms ($\gamma z \neq z$). We will see that this can be done by the amplification method.

However, this bound is not optimal for most surfaces. We have the following conjecture.

Conjecture 4.10. We have

$$u_j(z) \ll \lambda_j^{\varepsilon},$$

for any $\varepsilon > 0$ and $z \in \mathbb{H}$.

The first breakthrough of the sup-norm problem for hyperbolic surfaces was achieved by Iwaniec–Sarnak [7], who improved the general bounds for certain arithmetic (compact) hyperbolic surfaces.

Theorem 4.11. *Let u_j be a Hecke–Maass cusp form. Then we have*

$$|u_j(z)| \ll_z \lambda_j^{5/24+\varepsilon},$$

for any $\varepsilon > 0$ and $z \in \mathbb{H}$.

This has important applications to the subconvexity problem of L-functions and nodal domains of Hecke–Maass cusp forms. We should not give a full proof of this interesting result. To simplify, we consider only the special point $z = i$ and will only give a weaker bound. However we can still see the main ideas in the proof.

4.5. **Amplification.** Recall that the nth Hecke operators T_n (of weight 0) is defined as

$$T_n f(z) = \frac{1}{\sqrt{n}} \sum_{\gamma \in \Gamma \backslash \Gamma_n} f(\gamma z),$$

where

$$\Gamma_n = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbb{Z}) : ad - bc = n \right\}.$$

We know that

$$\Gamma \backslash \Gamma_n = \left\{ \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} : ad = n, \ b \bmod d \right\}.$$

We have

$$T_m T_n = \sum_{d|(m,n)} T_{\frac{mn}{d^2}} = T_n T_m, \quad T_n \Delta = \Delta T_n, \quad T_n^* = T_n.$$

So we can take $\{u_j\}$ to be an orthonormal system of cusp forms for the modular group which are eigenfunctions of all the Hecke operators. In this case, the nth Fourier coefficient $\lambda_j(n)$ is also the eigenvalue of the nth Hecke operator T_n.

Applying the Hecke operator to both sides of (4.5) we get

$$T_n K(z,w) = \frac{1}{\sqrt{n}} \sum_{\gamma \in \Gamma_n} k(u(\gamma z, w)) = \sum_{j \geq 0} h(t_j) \lambda_j(n) u_j(z) \overline{u_j(w)} + cont.$$

Hence we have

$$\frac{1}{\sqrt{n}} \nu_n k(0) + \frac{1}{\sqrt{n}} \sum_{\substack{\gamma \in \Gamma_n \\ \gamma z \neq z}} k(u(\gamma z, z)) = \sum_{j \geq 0} h(t_j) \lambda_j(n) |u_j(z)|^2 + cont. \tag{4.9}$$

where $\nu_n := \#\{\gamma \in \Gamma_n : \gamma z = z\}$.

For $z = i$, we have

$$\nu_n = r_2(n) = 4 \sum_{d|n} \chi_{-4}(d) \ll n^{\varepsilon}.$$

Indeed, we have

$$\begin{aligned} \nu_n &= \# \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbb{Z}) : ad - bc = n, \gamma i = \frac{ai+b}{ci+d} = i \right\} \\ &= \{a, b \in \mathbb{Z} : a^2 + b^2 = n\} = r_2(n). \end{aligned}$$

Lemma 4.12. *For $z = i$ and $X \geq 2$ we have*

$$\#\{\gamma \in \Gamma_n : 4u(\gamma z, z) + 2 \leq X\} \ll (nX)^{1+\varepsilon}.$$

Proof. Recall that

$$u(z, w) = \frac{|z - w|^2}{4 \operatorname{Im} z \operatorname{Im} w}.$$

Since $\gamma i = \frac{ai+b}{ci+d} = \frac{(ac+bd)+ni}{c^2+d^2}$ for $\gamma \in \Gamma_n$, we have

$$\begin{aligned} 4n\, u(\gamma i, i) = 4n \frac{\frac{(ac+bd)^2 + (ad-bc-c^2-d^2)^2}{(c^2+d^2)^2}}{4\frac{n}{c^2+d^2}} &= \frac{(ac+bd)^2 + (ad-bc)^2 - 2n(c^2+d^2) + (c^2+d^2)^2}{c^2+d^2} \\ &= a^2 + b^2 + c^2 + d^2 - 2n. \end{aligned}$$

Hence

$$\#\{\gamma \in \Gamma_n : 4u(\gamma z, z) + 2 \leq X\} = \#\{\gamma \in \Gamma_n : a^2 + b^2 + c^2 + d^2 \leq Xn\}.$$

For any fixed pair (b, c) we have at most $(nX)^{\varepsilon}$ choices of (a, d) since $ad - bc = n$. Since $|b| \leq \sqrt{nX}$ and $|c| \leq \sqrt{nX}$, we have the bound $O((nX)^{1+\varepsilon})$ as claimed. □

Note that for any $n \geq 1$ we have

$$4n\, u(\gamma i, i) = a^2 + b^2 + c^2 + d^2 - 2n = (a - d)^2 + (b + c)^2 \geq 1,$$

if $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_n$ and $\gamma i \neq i$.

From the above estimates we infer an estimate for the geometric side of $K_n(z, z)$. Combining this estimate with the spectral decomposition, we get the following proposition.

Proposition 4.13. *For $z = i$ and $n \geq 1$ we have*

$$\sum_j h(t_j)\lambda_j(n)|u_j(z)|^2 + cont. = \frac{r_2(n)}{\sqrt{n}} T + O(T^{1/2} n^{3/4+\varepsilon}), \tag{4.10}$$

the implied constant depending only on $\varepsilon > 0$.

Proof. By (4.9) we have

$$\sum_j h(t_j)\lambda_j(n)|u_j(z)|^2 + cont. = \frac{r_2(n)}{\sqrt{n}} T + \frac{1}{\sqrt{n}} \sum_{\substack{\gamma \in \Gamma_n \\ \gamma z \neq z}} k(u(\gamma z, z)).$$

By (4.8) and Lemma 4.12, for $z = i$ we have

$$\frac{1}{\sqrt{n}} \sum_{\substack{\gamma\in\Gamma_n \\ \frac{1}{4n}\le u(\gamma z,z)\le 3}} k(u(\gamma z, z)) \ll T^{1/2} n^{3/4+\varepsilon}$$

and

$$\frac{1}{\sqrt{n}} \sum_{\substack{\gamma\in\Gamma_n \\ u(\gamma z,z)\ge 3}} k(u(\gamma z, z)) \ll \frac{T^{1/2}}{\sqrt{n}} \sum_{\substack{\gamma\in\Gamma_n \\ u(\gamma z,z)\ge 3}} u(\gamma z, z)^{-3/2}.$$

By the partial summation formula and Lemma 4.12 we have

$$\sum_{\substack{\gamma\in\Gamma_n \\ u(\gamma z,z)\ge 3}} u(\gamma z, z)^{-3/2} = \int_3^\infty v^{-3/2} \mathrm{d} \sum_{\substack{\gamma\in\Gamma_n \\ 0<u(\gamma z,z)\le v}} 1 \ll n^{1+\varepsilon}.$$

This completes the proof. □

Since $1 \le r_2(n) \le 4\tau(n) \ll n^\varepsilon$, the above formula shows that as n gets large there exists a considerable cancellation of spectral terms due to the variation in sign of the Hecke eigenvalue $\lambda_j(n)$. This variation is the key to improving the sup norm. Unfortunately, for the same reason, we cannot drop all but one term to obtain directly a good bound for the individual cusp form.

4.6. **Constructing an amplifier.** We shall overcome the lack of positivity on the spectral side of (4.10) by means of an amplifier. Recall that we have the Hecke multiplication rule for the eigenvalues

$$\lambda_j(m)\lambda_j(n) = \sum_{d|(m,n)} \lambda_j\left(\frac{mn}{d^2}\right).$$

We conclude the following proposition.

Proposition 4.14. *For any complex sequence a_n we have*

$$\sum_j h(t_j) \left|\sum_{n\le N} a_n\lambda_j(n)\right|^2 |u_j(z)|^2 \ll N^\varepsilon(T\|a\|_2^2 + T^{1/2}N^{3/2}\|a\|_1^2),$$

where $\|a\|_1 = \sum_{n\le N} |a_n|$ and $\|a\|_2 = (\sum_{n\le N} |a_n|^2)^{1/2}$.

Proof. From (4.10) we have that

$$\begin{aligned}
&\sum_j h(t_j) \left|\sum_{n\le N} a_n\lambda_j(n)\right|^2 |u_j(z)|^2 + cont. \\
&\quad = \sum_{m\le N}\sum_{n\le N} a_m\overline{a_n} \sum_j h(t_j)\lambda_j(m)\lambda_j(n)|u_j(z)|^2 + cont. \\
&\quad = \sum_{m\le N}\sum_{n\le N} a_m\overline{a_n} \sum_{d|(m,n)} \sum_j h(t_j)\lambda_j\left(\frac{mn}{d^2}\right)|u_j(z)|^2 + cont. \\
&\quad = \sum_{m\le N}\sum_{n\le N} \frac{a_m\overline{a_n}}{\sqrt{mn}} \sum_{d|(m,n)} r_2\left(\frac{mn}{d^2}\right) d\, T + O\left(\sum_{m\le N}\sum_{n\le N} |a_m a_n| \sum_{d|(m,n)} T^{1/2}\left(\frac{mn}{d^2}\right)^{3/4+\varepsilon}\right).
\end{aligned}$$

The above error term can be bounded by

$$O\left(T^{1/2}N^{3/2+\varepsilon}\|a\|_1^2\right).$$

For the sums in the first term we have

$$\begin{aligned}
&\sum_{m\le N}\sum_{n\le N}\frac{a_m\overline{a_n}}{\sqrt{mn}}\sum_{d|(m,n)} r_2\left(\frac{mn}{d^2}\right)d \\
&\quad\ll N^\varepsilon \sum_{m\le N}\sum_{\substack{n\le N\\(m,n)=1}}\sum_{\ell\le\min(N/m,N/n)}\frac{|a_{\ell m}a_{\ell n}|}{\sqrt{mn}}\sum_{d|\ell}\frac{d}{\ell} \\
&\quad\ll N^\varepsilon \sum_{m\le N}\sum_{n\le N}\sum_{\ell\le\min(N/m,N/n)}\left(\frac{|a_{\ell m}|^2}{n}+\frac{|a_{\ell n}|^2}{m}\right) \\
&\quad\ll N^\varepsilon\|a\|_2^2.
\end{aligned}$$

This proves the proposition. □

On the left side the terms are non-negative, so we can drop all but one, getting

$$|L_j u_j(z)|^2 \ll N^\varepsilon(t_j\|a\|_2^2 + t_j^{1/2}N^{3/2}\|a\|_1^2), \tag{4.11}$$

where $z=i$ and

$$L_j = \sum_{n\le N} a_n\lambda_j(n).$$

Remark 4.15. If $\sum_{n\le N}\lambda_j(n)^2 \gg t_j^{-\varepsilon}N$, then we can take $a_n=\lambda_j(n)$, getting

$$L_j \gg N t_j^{-\varepsilon}.$$

Note that we have (due to Iwaniec)

$$\|a\|_2^2 = \sum_{n\le N}\lambda_j(n)^2 \ll N t_j^\varepsilon,$$

and

$$\|a\|_1^2 = (\sum_{n\le N}|\lambda_j(n)|)^2 \ll N^2 t_j^\varepsilon.$$

We conclude that

$$|u_j(z)|^2 \ll (Nt_j)^\varepsilon(t_j/N + t_j^{1/2}N^{3/2}).$$

The best choice is $N=t_j^{1/5}$, and we get

$$u_j(z) \ll \lambda_j^{1/5+\varepsilon}, \quad z=i.$$

Note that this is better than Theorem 4.11, however it is conditional.

Note that $\lambda_j(p)^2-\lambda_j(p^2)=1$. Without any conjecture we have still a good choice, namely

$$a_n = \begin{cases} \lambda_j(p), & \text{if } n=p\le\sqrt{N}, \\ -1, & \text{if } n=p^2\le N, \\ 0, & \text{othewise.}\end{cases} \tag{4.12}$$

By the prime number theorem, we have

$$L_j = \sum_{p \le \sqrt{N}} (\lambda_j(p)^2 - \lambda_j(p^2)) \sim \frac{\sqrt{N}}{\log \sqrt{N}}.$$

Note that

$$\|a\|_2^2 = \sum_{p \le \sqrt{N}} (\lambda_j(p)^2 + 1) \ll N^{1/2} t_j^{\varepsilon},$$

and

$$\|a\|_1^2 = \left(\sum_{p \le \sqrt{N}} (|\lambda_j(p)| + 1) \right)^2 \ll N t_j^{\varepsilon}.$$

Hence by (4.11) we get

$$|u_j(z)|^2 \ll N^{\varepsilon} (t_j/\sqrt{N} + t_j^{1/2} N^{3/2}).$$

The best choice is $N = t_j^{1/4}$, and we get

$$u_j(z) \ll \lambda_j^{7/32+\varepsilon}, \quad z = i.$$

Note that this is worse than Theorem 4.11. In any case, we obtain the following unconditional result.

Theorem 4.16. *For $z = i$ we have*

$$|u_j(z)| \ll \lambda_j^{7/32+\varepsilon}$$

the implied constant depending only on $\varepsilon > 0$.

Remark 4.17. Using more refined estimates for $\#\{\gamma \in \Gamma_n : u(\gamma z, z) \le \delta\}$, one can get a bound with the exponent $5/24$ in place of $7/32$. Also the result holds true for any $z \in \mathbb{H}$, so it yields a bound for the L^{∞}-norm:

$$\|u_j\|_{\infty} \ll \lambda_j^{5/24+\varepsilon}.$$

The same estimate has been established for the eigenfunctions with respect to the quaternion group by Iwaniec–Sarnak.

Remark 4.18. For Eisenstein series and dihedral Maass forms, we can improve Iwaniec–Sarnak's bound (getting $3/16$ instead of $5/24$). In those cases we can prove the lower bound $\sum_{n \sim N} |\lambda_f(n)|^2 \gg N t_f^{-\varepsilon}$ where f is an Eisenstein series or a dihedral Maass form. The main tool here is Korobov–Vinogradov type zero free region of L-functions.

Remark 4.19. There are many generalizations of this result, such as the level aspect estimates, hybrid bounds, higher rank cases, and so on.

References

[1] M.V. Berry, Regular and irregular semiclassical wavefunctions, *J. Phys. A* 10(12), 2083–2091, 1977. 30

[2] D. Bump, Automorphic forms and representations. Cambridge Studies in Advanced Mathematics, 55. *Cambridge University Press, Cambridge*, 1997. xiv+574 pp.

[3] D. Goldfeld, *Automorphic forms and L-functions for the group* $\mathrm{GL}(n, \mathbb{R})$. With an appendix by Kevin A. Broughan. Cambridge Studies in Advanced Mathematics, 99. Cambridge University Press, Cambridge, 2006. xiv+493 pp. 23

[4] H. Iwaniec, Topics in classical automorphic forms. Graduate Studies in Mathematics, 17. *American Mathematical Society, Providence, RI*, 1997. xii+259 pp. 20

[5] H. Iwaniec, Spectral methods of automorphic forms. Second edition. Graduate Studies in Mathematics, 53. *American Mathematical Society, Providence, RI; Revista Matemática Iberoamericana, Madrid,* 2002. xii+220 pp. 30

[6] H. Iwaniec and E. Kowalski, *Analytic number theory*, volume 53 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2004.

[7] H. Iwaniec and P. Sarnak, L^∞ norms of eigenfunctions of arithmetic surfaces. *Ann. of Math. (2)* 141 (1995), no. 2, 301–320. 33, 34

[8] R. Salem and A. Zygmund, Some properties of trigonometric series whose terms have random signs, *Acta. Math.* 91 (1954), 245–301. 30

[9] P. Sarnak, Arithmetical quantum chaos, in: *The Schur Lectures* (Tel Aviv 1992), Israel Math. Conf. Proc. 8, Bar-Ilan University, Ramat Gan (1995), 183–236. 30

[10] A. Seeger and C.D. Sogge, Bounds for eigenfunctions of differential operators, *Indiana Univ. Math. J.* 38 (1989), no. 3, 669–682. 30

[11] J.-P. Serre, A course in arithmetic. Translated from the French. Graduate Texts in Mathematics, No. 7. *Springer-Verlag, New York-Heidelberg*, 1973. viii+115 pp.

Data Science Institute and School of Mathematics, Shandong University, Jinan, Shandong 250100, China

E-mail address: brhuang@sdu.edu.cn