An introduction to analytic number theory on GL_2

October 27, 2020

Nate Gillman

Contents

1 L-functions and modular forms

(Lecture 1: September 10, 2020)

1.1 Bureaucracy

These informal notes cover the first fifteen lectures of Junehyuk Jung's topics in number theory course, taught at[1] Brown University in Fall 2020. These lectures contained an introduction to the GL_2 aspects of analytic number theory, and some connections to other fields. All errors in these notes are my own, feel free to send corrections to ngillman@brown.edu.

1.2 Periodic functions

Definition 1.1

A function $f : \mathbb{R} \to \mathbb{C}$ is *periodic* of period 1 if $f(x+n) = f(x)$ for all $n \in \mathbb{Z}$. In other words, such an f is invariant under the $\mathbb{Z}$-action.

A periodic function can be identified with a function on the circle $\mathbb{R}/\mathbb{Z} \cong S^1$. How does one construct a periodic function? Say g is a function on $[0,1)$. Then we can extend g periodically to $\mathbb{R}$, by setting $f(x) := g(\{x\})$, where $\{x\} := x - \lfloor x \rfloor$ is the fractional part of x. More generally, if $g \in S(\mathbb{R})$ (Schwartz space, which is the space of rapidly decreasing smooth functions on $\mathbb{R}$; namely, for any fixed $A > 0$, we have $|g(x)| < |x|^{-A}$ for $|x|$ sufficiently large) then we can define

$$f(x) := \sum_{n \in \mathbb{Z}} g(x+n),$$

which is an (absolutely convergent) periodic function.

Notation 1.2

Throughout this course, we'll take $e(x) := e^{2\pi i x}$.

Now we'll discuss the Fourier transform. Define

$$S_N(f) := \sum_{|n| \le N} a_n e(nx), \qquad \text{where} \qquad a_n := \hat{f}(n) := \int_0^1 f(x) e(-nx) dx.$$

[1] This course was conducted entirely via Zoom.

Theorem 1.3

1. If $f \in L^2(S^1)$, then $S_N(f) \to f$ in L^2.
2. If $f \in L^p(S^1)$ for $p > 1$, then $S_N(f) \to f$ almost everywhere.
3. There exists an $f \in L^1(S^1)$ such that $S_N(f)$ does not converge to f almost everywhere.

Theorem 1.4

If $f \in L^1$ satisfies

$$\int_0^1 \left| \frac{f(x_0+t)+f(x_0-t)}{2} - \ell \right| \frac{dt}{t} < \infty,$$

then $S_N(f)(x_0) \to \ell$.

The above finiteness condition clearly implies that the integrand converges to 0 at a rate $O(t)$, as $t \to 0$. In particular, the average value of f around x_0 is ℓ. Therefore, loosely speaking, this result says that "if f behaves like ℓ (with some regularity) around x_0, then the Fourier series of f at x_0 indeed converges to ℓ." This theorem yields the following results:

Corollary 1.5

1. If $f \in C^{0,\alpha}$ for some $\alpha > 0$, then $S_N(f) \to f$ uniformly.
2. If $f \in C$ and $\sum_{n\in\mathbb{Z}} |a_n| < \infty$, then $S_N(f) \to f$ uniformly.
3. If f has bounded variation, then $S_N(f) \to f$ pointwise.
4. There exists a continuous f such that $S_N(f)(x_0) \not\to f(x_0)$.

Next, one of the most important theorems in analytic number theory:

Theorem 1.6: Poisson summation formula

For $g \in S(\mathbb{R})$, the following identity holds:

$$\sum_{n\in\mathbb{Z}} g(n) = \sum_{n\in\mathbb{Z}} \hat{g}(n).$$

Proof. Define $f(x) := \sum_{n\in\mathbb{Z}} g(x+n)$. If we denote $a_n := \int_0^1 f(x)e(-nx)dx$, then we can compute that

$$a_n = \int_0^1 \sum_{m\in\mathbb{Z}} g(x+m)e(-nx)dx = \int_{-\infty}^{\infty} g(x)e(-nx)dx = \hat{g}(n),$$

so the n'th Fourier coefficient of f is the Fourier transform of g evaluated at n. Since $f \in S(\mathbb{R})$ implies f converges uniformly to its Fourier series, the above implies that

$$\sum_{n\in\mathbb{Z}} g(x+n) = f(x) = \sum_{n\in\mathbb{Z}} a_n e(nx) = \sum_{n\in\mathbb{Z}} \hat{g}(n)e(nx).$$

Specializing this to $x = 0$ yields the Poisson summation formula, as needed. □

As a first application, we'll see how Poisson summation can help us prove the analytic continuation and functional equation for $\zeta(s)$.

Theorem 1.7

The Riemann zeta function

$$\zeta(s) := \sum_{n \geq 1} \frac{1}{n^s},$$

which converges absolutely for $\Re(s) > 1$, can be meromorphically continued to all of $\mathbb{C}$.

Proof. Set $g(y) := e^{-\alpha y^2}$, so that $\hat{g}(t) = \sqrt{\frac{\pi}{\alpha}} e^{-\frac{\pi^2 t^2}{\alpha}}$ (see Lemma 1.2 below.) Applying Poisson summation to g, with $\alpha = \pi x$, yields

$$\sum_{n \in \mathbb{Z}} e^{-n^2 \pi x} = \frac{1}{\sqrt{x}} \sum_{n \in \mathbb{Z}} e^{-\frac{n^2 \pi}{x}}.$$

The salient feature of this formula is that the x in the exponent on the LHS has been moved to the denominator in the RHS. Set $\psi(x) := \sum_{n \geq 1} e^{-n^2 \pi x}$, so the above equation reads

$$1 + 2\psi(x) = \frac{1}{\sqrt{x}} \left(2\psi(1/x) + 1\right). \tag{1.1}$$

Now, apply to $\Gamma(s/2) := \int_0^\infty y^{\frac{s}{2}-1} e^{-y} dy$ the variable transformation $y = n^2 \pi x$, yielding

$$\int_0^\infty x^{\frac{s}{2}-1} e^{-n^2 x \pi} dx = \frac{\Gamma(s/2)}{n^s \pi^{s/2}}.$$

Summing this formula over $n \in \mathbb{Z}$ yields a ψ on the LHS and a ζ on the RHS, namely,

$$\Gamma(s/2)\pi^{-s/2}\zeta(s) = \int_0^\infty x^{\frac{s}{2}-1} \psi(x) dx.$$

Using (1.1), we can continue

$$\begin{aligned}
\int_0^\infty x^{\frac{s}{2}-1}\psi(x)dx &= \int_0^1 x^{\frac{s}{2}-1}\psi(x)dx + \int_1^\infty x^{\frac{s}{2}-1}\psi(x)dx \\
&= \int_0^1 x^{\frac{s}{2}-1}\left(\frac{1}{\sqrt{x}}\psi(1/x) + \frac{1}{2\sqrt{x}} - \frac{1}{2}\right)dx + \int_1^\infty x^{\frac{s}{2}-1}\psi(x)dx \\
&= \int_0^1 \frac{x^{\frac{s}{2}-\frac{3}{2}} - x^{\frac{s}{2}-1}}{2} dx + \int_0^1 x^{\frac{s}{2}-\frac{3}{2}}\psi(1/x)dx + \int_1^\infty x^{\frac{s}{2}-1}\psi(x)dx \\
&= -\frac{1}{s(1-s)} + \int_1^\infty y^{\frac{3}{2}-\frac{s}{2}}\psi(y)(-y^{-2}dy) + \int_1^\infty x^{\frac{s}{2}-1}\psi(x)dx \\
&= -\frac{1}{s(1-s)} + \int_1^\infty \left(x^{\frac{1-s}{2}} + x^{\frac{s}{2}}\right)\psi(x)\frac{dx}{x}.
\end{aligned}$$

In summary, we've shown that

$$\Gamma(s/2)\pi^{-s/2}\zeta(s) = -\frac{1}{s(1-s)} + \int_1^\infty \left(x^{\frac{1-s}{2}} + x^{\frac{s}{2}}\right)\psi(x)\frac{dx}{x}. \tag{1.2}$$

The integral is a holomorphic function in s since $\psi(x)$ converges rapidly (in particular, it converges faster than any x^{-A}). And the rational function out front implies that the RHS has a simple pole at $s = 0, 1$. This formula is invariant under $s \to 1 - s$, which implies that

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

This relation tells us that $\zeta(s)$ is meromorphic on $\mathbb{C}$, and has a simple pole only at $s = 1$ (note that there's no pole at $s = 0$ because of the pole of $\Gamma(s)$ at $s = 0$.) □

Lemma 1.8

The Fourier transform of the Gaussian $g(y) := e^{-\alpha y^2}$ is

$$\hat{g}(t) = \sqrt{\frac{\pi}{\alpha}} e^{-\frac{\pi^2 t^2}{\alpha}}.$$

Proof. We compute

$$\hat{g}(t) = \int_{-\infty}^{\infty} e^{-\alpha y^2} e(-ty) dy = \int_{-\infty}^{\infty} e^{-\alpha y^2 - 2\pi i t y} dy.$$

We complete the square and change variables to turn this into the standard Gaussian integral. Namely, since

$$-\alpha y^2 - 2\pi i t y = -\Big[(\sqrt{\alpha} y + \frac{i\pi t}{\sqrt{\alpha}})^2 - \left(\frac{i\pi t}{\sqrt{\alpha}}\right)^2\Big],$$

we have

$$\hat{g}(t) = e^{-\frac{\pi^2 t^2}{\alpha}} \int_{-\infty}^{\infty} e^{-(\sqrt{\alpha} y + \frac{i\pi t}{\sqrt{\alpha}})^2} dy.$$

Applying the variable transformation $x = \sqrt{\alpha} y + i\pi t/\sqrt{\alpha}$ yields

$$\hat{g}(t) = \frac{1}{\sqrt{\alpha}} e^{-\frac{\pi^2 t^2}{\alpha}} \int_{-\infty + \frac{i\pi t}{\sqrt{\alpha}}}^{\infty + \frac{i\pi t}{\sqrt{\alpha}}} e^{-x^2} dx.$$

By Cauchy's integral theorem, we can shift this contour down to the real line because the integral is small over the segments $[\pm T + i\pi t/\sqrt{\alpha}]$ for large T. And the Gaussian integral has value $\sqrt{\pi}$, which one can show by integrating two Gaussians against each other in polar coorinates. This finishes the proof. □

Next, we'll briefly prove other fundamental facts about zeros of ζ.

Proposition 1.9

$\zeta(s)$ doesn't vanish when $\Re(s) > 1$.

Proof. Since $\mu * 1 = \epsilon$ (see below two results) we know that

$$\sum_{n \geq 1} \frac{\mu(n)}{n^s} \sum_{m \geq 1} \frac{1}{m^s} = 1.$$

Both series converge absolutely for $\Re(s) > 1$ (observe that the first is bounded termwise by the second), therefore neither Dirichlet series can have a pole in this half-plane. □

Theorem 1.10

If the Dirichlet series

$$F(s) = \sum_{n \geq 1} \frac{f(n)}{n^s}, \qquad G(s) = \sum_{n \geq 1} \frac{g(n)}{n^s}$$

both converge absolutely at s, then so does

$$H(s) = \sum_{n \geq 1} \frac{(f * g)(n)}{n^s}.$$

Furthermore, in this case we have that $F(s)G(s) = H(s)$.

Proof. See Theorem 4.1 here. □

Lemma 1.11

We have that $\mu * 1 = \epsilon$, where

$$\epsilon(n) = \begin{cases} 1 & \text{if } n = 1 \\ 0 & \text{if } n > 1 \end{cases}.$$

Proof. We can compute

$$(\mu * 1)(n) = \sum_{d|n} \mu(n) \cdot 1(n/d) = \sum_{d|n} \mu(n).$$

If $n > 1$, then we can write $n = p_1^{e_1} \cdots p_m^{e_m}$, so

$$\sum_{d|n} \mu(n) = \sum_{d|p_1 \cdots p_m} \mu(n) = 0,$$

since half of the subsets of $\{p_1, \ldots, p_m\}$ have even cardinality and half have odd cardinality (e.g. one can define an involution on this set by removing or adding p_1, and such an involution has no fixed points.) □

Corollary 1.12

For $\Re(s) < 0$, the only zeros of $\zeta(s)$ are $s = -2n$, for $n \in \mathbb{Z}$.

Proof. By the Proposition, $\Gamma(s/2)\pi^{-s/2}\zeta(s)$ has no zero or pole for $\Re(s) > 1$. Therefore the same is true for $\Gamma((1-s)/2)\pi^{-(1-s)/2}\zeta(1-s)$. By change of variable from s to $1-s$, this implies $\Gamma(s/2)\pi^{-s/2}\zeta(s)$ has no zero or pole for $\Re(s) < 0$. We know that Γ has simple poles precisely at the non-positive integers. So whenever Γ hits a pole in this region, ζ has to hit a zero to compensate. It follows that for $\Re(s) < 0$, $\zeta(s)$ has simple zeros precisely at the negative even integers. □

Those zeros are referred to as "the trivial zeros" of $\zeta(s)$. Riemann showed that the prime number theorem follows from $\zeta(1 + it) \neq 0$ for $0 \neq t \in \mathbb{R}$.

Theorem 1.13: Prime number theorem

$\#\{p < x : p \text{ prime}\} \sim x/\log x$.

This is an example of how analytic properties of ζ tell you things about prime numbers. Riemann conjectured:

Conjecture 1.14: Riemann Hypothesis

Every non-trivial zero of $\zeta(s)$ is of the form $\frac{1}{2} + it$ for some $t \in \mathbb{R}$.

Finally we state without proof some fundamental results about Γ that will be needed throughout this course:

Proposition 1.15

Facts about the Γ function:

1. Γ is meromorphic on $\mathbb{C}$ with simple poles at the non-positive integers.
2. $\Gamma(n+1) = n!$, for $n \in \mathbb{Z}_{\geq 0}$.
3. $z\Gamma(z) = \Gamma(z+1)$.
4. $\Gamma(1-z)\Gamma(z) = \pi/\sin(\pi z)$, which implies the first point..

(Lecture 2: September 15, 2020)

1.3 Elliptic functions

Elliptic functions are a natural source of modular forms. Let $\Lambda \subseteq \mathbb{C}$ be a lattice, i.e., Λ is a discrete subgroup of $\mathbb{C}$ of rank 2. This means $\Lambda = \omega_1\mathbb{Z} + \omega_2\mathbb{Z}$ for some $\mathbb{R}$-linearly independent complex numbers ω_1, ω_2. For example, taking $\omega_1 = 1$ and $\omega_2 = i$, we get $\Lambda = \mathbb{Z}[i]$.

Definition 1.16

A function $f : \mathbb{C} \to \mathbb{C}$ is *elliptic* with respect to Λ if and only if

1. f is meromorphic on $\mathbb{C}$
2. f is periodic with periods in Λ, i.e. $f(u+\omega) = f(u)$ for all $\omega \in \Lambda$ and any $u \in \mathbb{C}$.

Proposition 1.17

Fix a lattice Λ. Then the set $E(\Lambda)$ of elliptic functions with respect to Λ is a field, and this field is isomorphic to the field of meromorphic functions on the torus C/Λ.

Proof. The product/sum/difference of meromorphic functions is clearly meromorphic, and the reciprocal of a meromorphic functions is also meromorphic. □

Proposition 1.18

If f is an elliptic function with no poles, then f is a constant function.

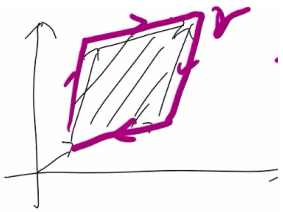
Proof. If f has no poles, then by Liouville's theorem, f is a bounded function on the entire plane (since its values in $\mathbb{C}$ are determined by its values in a compact neighborhood of a fundamental parallelogram.) □

Proposition 1.19

Let P be the set of poles in a fundamental parallelogram. Then,

$$\sum_{\omega \in P} \operatorname{Res}_f(\omega) = 0.$$

Proof. We can shift the fundamental parallelogram so that the edges intersect no poles/zeros (since the set of poles/zeros is discrete, or else f is constant.)



On one hand, we know that $\frac{1}{2\pi i}\int_\gamma f(z)dz = 0$ by periodicity (namely, the integral along opposite edges cancel each other.) But on the other hand, this integral is the sum of the residues. □

Corollary 1.20

No $f \in E(\Lambda)$ has exactly one simple pole in the fundamental parallelogram.

Proof. If f had one simple pole, then $\sum_{\omega \in P} \operatorname{Res}_f(\omega)$ would be the (nonzero) residue of that pole. □

Definition 1.21

Let $f \in E(\Lambda)$, and let $\omega \in \mathbb{C}$. The order $m := m_f(\omega)$ of f at ω is the unique $m \in \mathbb{Z}$ such that $f(u)(u-\omega)^{-m}$ has no zero or pole at $u = \omega$.

For example, if f has a simple pole at ω, then $f(u)(u-\omega)$ has no zero or pole at $u = \omega$, which implies that the order of f at ω is $m_f(\omega) = -1$.

Proposition 1.22

Fix a lattice Λ, and let S be a fundamental parallelogram for C/Λ. Then

$$\sum_{\omega \pmod{\Lambda}} m_f(\omega) = \sum_{\omega \in S} m_f(\omega) = 0.$$

Proof. Apply the previous proposition to the elliptic function f'/f.

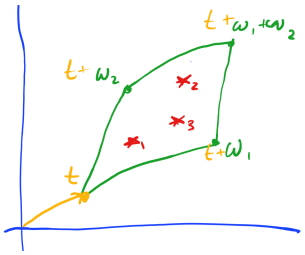
□

Proposition 1.23

For any $\omega \in \mathbb{C}$, we have

$$\sum_{\omega \pmod{\Lambda}} \omega \cdot m_f(\omega) \equiv 0 \pmod{\Lambda}.$$

Proof. Let us consider a fundamental parellelogram S which has no poles or zeros on its boundary.



It is sufficient to show the following:

1. $\frac{1}{2\pi i}\int_{\partial S} \frac{uf'(u)}{f(u)}du = \sum_{\omega \in S} \omega \cdot m_f(\omega)$
2. $\frac{1}{2\pi i}\int_{\partial S} \frac{uf'(u)}{f(u)}du = 0$

The first point is clear, as every (nonzero) residue of $uf'(u)/f(u)$ corresponds to a pole or zero ω of $f(u)$, and its residue is $\omega \cdot m_f(\omega)$. For the second point, by symmetry it suffices to show that

$$\frac{1}{2\pi i}\int_{[t,t+\omega_1]} \frac{uf'(u)}{f(u)}du + \frac{1}{2\pi i}\int_{[t+\omega_1+\omega_2,t+\omega_2]} \frac{uf'(u)}{f(u)}du \in \Lambda.$$

Applying the transformation $v = u - \omega_2$ to the second integral shows that this sum is precisely

$$\begin{aligned}
&\frac{1}{2\pi i}\int_{[t,t+\omega_1]} \frac{uf'(u)}{f(u)}du + \frac{1}{2\pi i}\int_{[t+\omega_1+\omega_2,t+\omega_2]} \frac{uf'(u)}{f(u)}du \\
&\quad = \frac{1}{2\pi i}\int_{[t,t+\omega_1]} \frac{uf'(u)}{f(u)}du + \frac{1}{2\pi i}\int_{[t+\omega_1,t]} \frac{(v+\omega_2)f'(v+\omega_1)}{f(v+\omega_1)}dv \\
&\quad = -\frac{\omega_2}{2\pi i}\int_{[t,t+\omega_1]} \frac{f'(u)}{f(u)}du \\
&\quad = -\frac{\omega_2}{2\pi i}\left(\log f(t+\omega_1) - \log f(t)\right) \\
&\quad = 0.
\end{aligned}$$

The salient point here is that we applied periodicity repeatedly. Note that this computation returns 0, rather than some other element of the lattice, because we summed over a set of representatives for $\omega \pmod{\Lambda}$ from a connected parallelogram. □

Definition 1.24

Let $f \in E(\Lambda)$. The *order* of f, denoted r_f, is the sum of the orders of zeros modulo Λ. That is,

$$r_f := \sum_{\omega \pmod{\Lambda}} \max\{m_f(\omega), 0\} = \sum_{\omega \pmod{\Lambda}} -\min\{m_f(\omega), 0\}$$

The second equality follows from the fact that $\sum_{\omega \pmod{\Lambda}} m_f(\omega) = 0$.

Proposition 1.25

No $f \in E(\Lambda)$ has order 1.

Proof. We argued above that no $f \in E(\Lambda)$ has exactly one simple pole in the fundamental parallelogram, so the result follows from $\sum_{\omega \pmod{\Lambda}} m_f(\omega) = 0$. □

The simplest elliptic function that one can come up with is the Weierstrass $\wp$-function. As we'll see, this function is fundamental to the theory.

Definition 1.26

For a lattice Λ, the Weierstrass $\wp$-function is defined to be

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

The prime indicates that the sum is restricted to nonzero elements of the lattice.

(Why do we subtract $1/\omega^2$? One can show that, without this term, the sum doesn't converge absolutely. See Apostol for a proof that this series converges absolutely.) One might expect from this expression that $\wp$ has a pole at 0 of order 2. Since the series converges absolutely, it follows that this is the only pole, hence the order of $\wp$ is 2.

Here we collect some results about the $\wp$-function.

1. **Proposition 1.27**

   $r_\wp = 2$, and in particular, for any $c \in \mathbb{C}$ we have $r_{\wp - c} = 2$.

   Proof. We already saw that $r_\wp = 2$, so $\wp$ has two poles in any fundamental domain. This implies that for any $c \in \mathbb{C}$, $\wp - c$ also has two poles in any fundamental domain. Since $\sum_{\omega \pmod{\Lambda}} m_f(\omega) = 0$, we conclude that $\wp - c$ has two zeros counted with multiplicity, so by definition $r_{\wp - c} = 2$. □

Proposition 1.28

2. $\wp$ is an even function, i.e. $\wp(u) = \wp(-u)$.

Proof. This follows immediately from the definition of $\wp$ (rearrange the series via $\omega \mapsto -\omega$). □

Proposition 1.29

3. $\wp(u) = \wp(w)$ if and only if $u = \pm w \pmod{\Lambda}$.

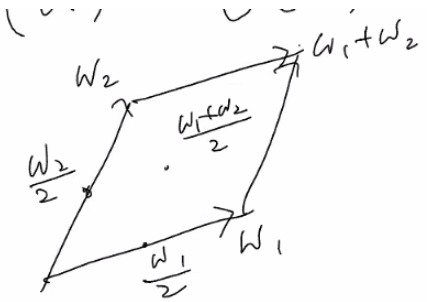
Proof. We combine the previous two results. Specifically, fix w. Then $r_{\wp(*)-\wp(w)} = 2$, so $\wp(u) - \wp(w) = 0$ only at two possible values. Clearly w works, but also $\wp(w) = \wp(-w)$, hence $-w$ works as well. □

Proposition 1.30

4. Fix $w \in \mathbb{C}/\Lambda$. Then, $\wp(u) - \wp(w)$ has a double zero at $u = w$ if and only if $w \equiv -w \pmod{\Lambda}$, which happens if and only if $2w \equiv 0 \pmod{\Lambda}$.

Proof. By the previous proposition, $\wp(u) - \wp(w) = 0$ if and only if $u = \pm w$. These two zeros are actually a double zero if and only if these are the same, i.e., if $w \equiv -w \pmod{\Lambda}$. □

Observe that the only the only 2-torsion points on the torus $\mathbb{C}/\Lambda$ are the three points $\omega_1/2, \omega_2/2$, and $(\omega_1 + \omega_2)/2$.



Proposition 1.31

5. The values of $\wp$ at the 2-torsion points of Λ,

$$e_1 = \wp(\omega_1/2), \qquad e_2 = \wp(\omega_2/2), \qquad e_3 = \wp((\omega_1 + \omega_2)/2),$$

are distinct. In particular, the discriminant is nonzero:

$$\Delta := 16(e_1 - e_2)^2(e_2 - e_3)^2(e_3 - e_1)^2 \neq 0.$$

Proof. We'll argue that $e_1 \neq e_2$. By the previous proposition, on one hand $\wp(u) - e_1$ has a double zero at $\omega_1/2$, and on the other hand $\wp(u) - e_2$ has a double zero at $\omega_2/2$. So if $e_1 = e_2$, this implies $\wp(u) - e_1 = \wp(u) - e_2$ has two double poles, since $\omega_1/2$ and $\omega_2/2$ are distinct points of the torus $\mathbb{C}/\Lambda$. This contradicts $r_{\wp - e_1} = 2$. The other two cases follow similarly. □

It turns out $\wp$ and its derivative generate all elliptic functions on C/Λ.

Theorem 1.32

For any lattice Λ, we have $E(\Lambda) = \mathbb{C}(\wp, \wp')$.

Proof. If $f \in E(\Lambda)$, then $f(-u) \in E(\Lambda)$ as well. So we may decompose f into even and odd elliptic functions as follows:

$$f(u) = \frac{f(u) + f(-u)}{2} + \frac{f(u) - f(-u)}{2} =: g(u) + h(u).$$

We'll show that the even part is a rational function in $\wp$, and the odd part a rational function in $\wp'$. Consider the product

$$g(u) \prod_{w \pmod{\Lambda}} (\wp(u) - \wp(w))^{-m_f(w)},$$

which is a finite product since $m_f(w)$ is nonzero only for finitely many w. One can check that this is an elliptic function without poles (since the poles of $g(u)$ are killed by the product), which means it has to be a constant function, by Liouville's theorem. Therefore,

$$g(u) = c \prod_{\omega \pmod{\Lambda}} (\wp(u) - \wp(w))^{m_f(\omega)},$$

which is a polynomial in $\wp(u)$. A similar argument gives that $h(u)$ is a rational function in the $\wp'$, which is odd. □

The previous result shows that $\wp$ and $\wp'$ are related in that they together generate all ellptic functions for a given lattice. The next theorem says that, in fact, $\wp$ and $\wp'$ are algebraically related.

Theorem 1.33

Fix a lattice Λ. Then, $\wp'$ is algebraically related to $\wp$ via

$$(\wp')^2 = 4(\wp - e_1)(\wp - e_2)(\wp - e_3).$$

Proof. The RHS has double zeros at $\omega_1/2, \omega_2/2, (\omega_1 + \omega_2)/2$, so

$$\wp(u) - \wp(\omega_1/2) = (z - \omega_1/2)^2 h(z),$$

for some $h(z)$ which doesn't vanish at $\omega_1/2$. So taking the derivative of both sides implies

$$\wp'(z) = (z - \omega_1/2) \cdot \text{something}.$$

Therefore, the LHS has double zeros at these three points as well. Now, consider

$$f = \frac{(\wp')^2}{4(\wp - e_1)(\wp - e_2)(\wp - e_3)}.$$

We just verified that the zeros of the RHS don't create any poles when we divide the LHS by the RHS, hence f may only have a pole at $u = 0$, coming from the numerator. But one can show that, as $u \to 0$, $\wp'(u) \sim -2u^{-3}$ (this follows directly from differentiating $\wp$ term by term) so in fact f has a triple pole at 0, whereas $\wp(u) \sim u^{-2}$, and therefore $f(u) \sim 1$ as $u \to 0$. Thus f is an elliptic function with no poles, and is therefore holomorphic on all of $\mathbb{C}$. By Liouville's theorem, it has to be constant, hence, $f(u) = 1$. □

Next, we'll talk about the power series expansion of $\wp$. First, recall the expansion

$$\frac{1}{u - w} = -\frac{\frac{1}{w}}{1 - \frac{u}{w}} = -\frac{1}{w} - \frac{u}{w^2} - \frac{u^2}{w^3} - \cdots.$$

Differentiating with respect to w on both sides yields

$$\frac{1}{(u - w)^2} = \frac{1}{w^2} + \frac{2u}{w^3} + \frac{3u^2}{w^4} + \cdots.$$

Plugging this into the definition of $\wp$ as the lattice average, we get a power series expansion for $\wp$:

$$\begin{aligned}
\wp(u) &= \frac{1}{u^2} + \sum_{\omega \in \Lambda}{}' \left(\frac{1}{(u - \omega)^2} - \frac{1}{\omega^2} \right) \\
&= \frac{1}{u^2} + \sum_{\omega \in \Lambda}{}' \left(\frac{2u}{\omega^3} + \frac{3u^2}{\omega^4} + \frac{4u^3}{\omega^5} + \cdots \right) \\
&= \frac{1}{u^2} + \sum_{m \geq 1} (m+1) G_{m+2} u^m,
\end{aligned}$$

where $G_k = \sum'_{\omega \in \Lambda} \omega^{-k}$. (The point: collecting the coefficients of u^m picks up a term of the form $(m+1)/\omega^{m+2}$ for every $\omega \in \Lambda^*$.) Note that when k is odd, G_k vanishes by symmetry in the sum. Also, in the expansion, we won't see G_2 since the expansion starts from G_3. We summarize:

Proposition 1.34

Fix a lattice Λ. The power series expansion of $\wp$ at 0 is

$$\wp(u) = \frac{1}{u^2} + \sum_{m \geq 1} (m+1) G_{m+2} u^m,$$

where $G_k = G_k(\Lambda) := \sum'_{\omega \in \Lambda} \omega^{-k}$.

If we differentiate this with respect to u, we get

$$\wp'(u) = -\frac{2}{u^3} + \sum_{m \geq 1} m(m+1) G_{m+2} u^{m-1}.$$

Proposition 1.35

Fix a lattice Λ, and set $g_2 := 60G_4$ and $g_3 := 140G_6$. Then

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

Proof. One can check explicitly that the power series expansion for $(\wp'(u))^2 - 4\wp(u)^3 + g_2\wp(u) + g_3$ has only positive powers of u, so it must vanish. □

Corollary 1.36

Fix a lattice Λ. Then the collection $\{(\wp(u), \wp'(u)) : u \in \mathbb{C}/\Lambda\}$ lies on the elliptic curve given by

$$y^2 = 4x^3 - g_2x - g_3 = 4(x - e_1)(x - e_2)(x - e_3).$$

Moreover, this gives a complete parameterization. Further, the discriminant of this elliptic curve is $\Delta = g_2^3 - 27g_3^2$.

1.4 Modular functions

One can regard an elliptic function $f(u)$ as a function of lattices as well. Write

$$f(u) := f(u; \omega_1, \omega_2),$$

where ω_1, ω_2 are the generators of Λ.

Proposition 1.37

For any $\lambda \neq 0$, we have $\wp(\lambda u; \lambda\omega_1, \lambda\omega_2) = \lambda^{-2}\wp(u; \omega_1, \omega_2)$ and $\wp'(\lambda u; \lambda\omega_1, \lambda\omega_2) = \lambda^{-3}\wp(u; \omega_1, \omega_2)$

Proof. The first equality follows from the definition of $\wp$, and the second taking from the derivative term by term. □

Definition 1.38

We say $f \in E(\Lambda)$ is *homogeneous* in u and Λ of degree $-k$ if, for every $\lambda \neq 0$,

$$f(\lambda u; \lambda\omega_1, \lambda\omega_2) = \lambda^{-k} f(u, \omega_1, \omega_2).$$

Note: the functions $\wp$ and $\wp'$ natural building blocks for such homogeneous lattice functions, since $(2, 3) = 1$.

Let $f \in E(\Lambda)$ be homogeneous of degree $-k$, and let

$$f(u) = \sum_m F_m(\omega_1, \omega_2)u^{m-k}$$

be the power series expansion in u around 0. Without loss of generality, assume $\Im(\omega_1/\omega_2) > 0 > \Im(\omega_2/\omega_1)$. Then,

$$\sum_m F_m(\omega_1,\omega_2)u^{m-k} = \lambda^k f(\lambda u, \lambda\omega_1, \lambda\omega_2) = \sum_m \lambda^m F_m(\lambda\omega_1,\lambda\omega_2)u^{m-k}.$$

Because the power series expansion is unique, these coefficients must match, meaning

$$F_m(\omega_1,\omega_2) = \lambda^m F_m(\lambda\omega_1,\lambda\omega_2) = \omega_2^{-m} F_m(\omega_1/\omega_2, 1) = \omega_2^{-m} F_m(z),$$

where we specialized to $\lambda = \omega_2^{-1}$ and wrote $z = \omega_1/\omega_2$. Next, we note that, by homogeneity of F_m, we have $F_m(\omega_1',\omega_2') = F_m(\omega_1,\omega_2)$ if $\omega_1' = a\omega_1 + b\omega_2$ and $\omega_2' = c\omega_1 + d\omega_2$ with $ad - bc = \pm 1$ (we'll show on the homework that choosing a lattice basis is well-defined up to a $\mathrm{PSL}_2(\mathbb{Z})$-action.) If $\Im(\omega_1/\omega_2) > 0$ and $\Im(\omega_1'/\omega_2') > 0$, then one can show that $\gamma z = z'$ for some $\gamma \in \mathrm{SL}_2(\mathbb{Z})$, where $z = \omega_1/\omega_2$ and $z' = \omega_1'/\omega_2'$. Here, $\gamma z := \frac{az+b}{cz+d}$. From this, it's clear that

$$F_m(z)\omega_2^{-m} = F_m(z')\omega_2'^{-m} = F_m(\gamma z)\omega_2'^{-m}.$$

This implies that

$$F_m(z) = F_m(\gamma z)(\omega_2'/\omega_2)^{-m} = F_m(\gamma z)((c\omega_1 + d\omega_2)/\omega_2)^{-m} = F_m(\gamma z)(cz+d)^{-m}.$$

These are the modular functions!! **And notice how such F_m generalize the Eisenstein series, which are the power series coefficients of $\wp$ and satisfy the same homogeneity property that leads to the slighly messier modular transformation law above.**

Definition 1.39

A function $F : \mathbb{H} \to \mathbb{C}$ satisfying

$$F_m(\gamma z) = (cz+d)^m F_m(z)$$

is called a *modular function* of weight m.

Proposition 1.40

The weight m must be even for F to not vanish.

Proof. Taking $\gamma = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$, we get $F_m(z) = F_m(z)(-1)^{-m}$, which forces F_m to vanish everywhere. □

(Note: that proposition also generalizes the Eisenstein case.)

Example: the Eisenstein series. For a lattice Λ generated by ω_1 and ω_2, we defined

$$G_k(\omega_1,\omega_2) = \sum_{(m,n)\neq(0,0)} \frac{1}{(m\omega_1 + n\omega_2)^k},$$

which leads us to define

$$G_k(z) := \sum_{(m,n)\neq(0,0)} \frac{1}{(mz+n)^k} = \left(\frac{1}{1^k} + \frac{1}{2^k} + \frac{1}{3^k} + \dots\right) \sum_{(m,n)=1} \frac{1}{(mz+n)^k} = 2\zeta(k)E_k(z),$$

where the $1/\ell^k$ terms comes from factoring out $\ell = (m, n)$ from the lattice sum.

Definition 1.41

G_k is the *Eisenstein series*, and E_k is the *normalized Eisenstein series*.

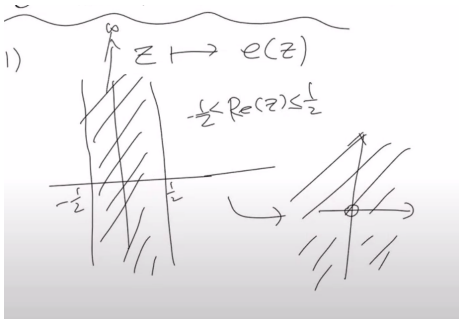
As we've seen, $E_k(z)$ is a modular function of weight k, hence $\Delta(z)$ is a modular function of weight 12.

1.5 Modular forms

Definition 1.42

A modular function which is meromorphic on $\mathbb{H}$ and at ∞ is called a *modular form*.

Unpacking this definition: if we take $\gamma = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, then $\gamma z = z + 1$, so the modular transformation law says that $F(z) = F(z+1)$. Now, consider the map that sends $z \mapsto e(z)$. This maps the strip $\Re(z) \in (1/2, 1/2]$ to the punctured complex plane. In order to achieve the point 0, you'd need to go to "all the way to $i\infty$".



If we write $F(z) = G(e(z))$ for some function G on $\mathbb{C}^*$, then what we mean by "F is meromorphic at $i\infty$" is just that G is meromorphic at 0, which means that G has a power series expansion

$$G(q) = \sum_{n \geq -M} a_n q^n$$

which starts at some *finite* lower index. In this case, we have a Fourier expansion of $F(z)$

$$F(z) = \sum_{n \geq -M} a_n e(nz)$$

because $q = e(z)$ here.

(Lecture 3: September 17, 2020)

Proposition 1.43

We have

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

We'll use this fact from complex analysis to find a Fourier expansion of Eisenstein series. By taking the logarithmic derivative of this equation, we find that

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

We can continue

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

which implies

$$\frac{1}{z} + \sum_{n \geq 1} \left(\frac{1}{z-n} + \frac{1}{z+n} \right) = \pi i - 2\pi i \sum_{d \geq 0} e(dz).$$

Differentiating this $k-1$ times, with $k \geq 2$, yields

$$\sum_{n=-\infty}^{\infty} (z-n)^{-k} = \frac{(-2\pi i)^k}{(k-1)!} \sum_{d \geq 1} d^{k-1} e(dz)$$

because of the identities

$$\frac{d^{k-1}}{dz^{k-1}} (z \pm n)^{-1} = (k-1)!(-1)^{k-1}(z \pm n)^{-k}, \qquad \frac{d^{k-1}}{dz^{k-1}} e^{2\pi i d z} = (2\pi i d)^{k-1} e^{2\pi i d z}.$$

Using this, we compute

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

where in the final step, we made the substitution $dm = n$, so the summation turns to $\sum_{d|n} d^{k-1} =: \sigma_{k-1}(n)$ (Remark: you can also compute the Fourier expansion just from the definition.)

Proposition 1.44

For k even,

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

where B_k is the kth Bernoulli number.

Note: understanding the case where k is *odd* is a notoriously hard problem.

Corollary 1.45

The normalized Eisenstein series $E_k = G_k/2\zeta(k)$ has Fourier expansion

$$E_k(z) = 1 - 2kB_k^{-1} \sum_{n \geq 1} \sigma_{k-1}(n) e(nz).$$

Recall that the modular discriminant corresponding to the lattice Λ is $\Delta(z) = g_2(z)^3 - 27g_3(z)^2$. In a homework problem, we'll show that this has a Fourier expansion of the form

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

with the $\tau(n)$ integers and $\tau(1) = 1$; to verify this, we'll apply the expansion of $E_k(z)$.

1.6 Modular surface

Definition 1.46

We define $\mathrm{PSL}_2(\mathbb{Z}) := \mathrm{SL}_2(\mathbb{Z})/\{\pm I\}$.

Theorem 1.47

$\mathrm{PSL}_2(\mathbb{Z}) = \langle T, S \rangle$, with $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$.

Proof. We compute

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

If $c \neq 0$, then for a suitable choice of $n_1 \in \mathbb{Z}$, applying T^{n_1} on the left has the effect of reducing the left upper entry to $0 \leq a < |c|$. Then apply S to swap the top and bottom rows, and apply another suitable T^{n_2}, etc. This eventually yields a matrix having $c = 0$, which must be of the form $\pm \begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix}$. Applying T^{-m} then yields $\pm I$, and we're done once we solve for the original matrix in the resulting equation. □

Next we'll discuss fundamental domains. As a motivating example, recall that we identified $\mathbb{R}/\mathbb{Z} \cong S^1$. But if we want to realize $\mathbb{R}/\mathbb{Z}$ on the real line, then we could take for example $[0, 1)$. We'll do a similar thing for the $\mathrm{SL}_2(\mathbb{R})$ action on $\mathbb{H}$. Recall that $\mathrm{SL}_2(\mathbb{R})$ acts on $\mathbb{H}$ by fractional transformations,

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

Furthermore, $\mathrm{SL}_2(\mathbb{Z})$ acts on $\mathbb{H}$ discontinuously, just like how $\mathbb{Z}$ acts on $\mathbb{R}$ discontinuously. So it makes sense to talk about the fundamental domain of $\mathbb{H}/\mathrm{SL}_2(\mathbb{Z})$.

Definition 1.48

Let Γ be a discrete subgroup acting on $\mathbb{H}$. A *fundamental domain* of $\Gamma\backslash\mathbb{H}$ is an open region $\mathcal{F} \subseteq \mathbb{H}$ such that, for every $z \in \mathbb{H}$, there exists a $\gamma \in \Gamma$ such that $\gamma z \in \overline{\mathcal{F}}$, and for any $z_1 \neq z_2 \in \mathcal{F}$, $\Gamma z_1 \not\ni z_2$.

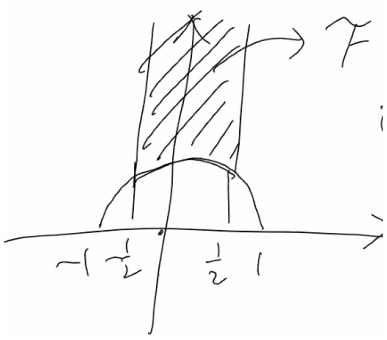
Theorem 1.49

The region

$$\mathcal{F} := \{z : |\Re z| < 1/2, |z| > 1\}$$

is a fundamental domain for $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$.

(See the proof in Iwaniec 1.5.) An illustration of the fundamental domain for $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$:



A priori, $\mathrm{PSL}_2(\mathbb{R})$ acts as bijections on $\mathbb{H}$. But we can say more if we allow ourselves to use the language of differential geometry. Namely, $\mathrm{PSL}_2(\mathbb{R})$ is an isometry group of $\mathbb{H}$ when it's equipped with the hyperbolic metric, which makes the constant curvature of $\mathbb{H}$ equal to -1. We can classify the different elements of $\mathrm{PSL}_2(\mathbb{R})$ according to their geometric idiosyncrasies:

Definition 1.50

A transformation $\gamma \in \mathrm{PSL}_2(\mathbb{R})$ is called:

1. *elliptic* if $|\operatorname{Tr}\gamma| < 2$;
2. *parabolic* if $|\operatorname{Tr}\gamma| = 2$;
3. *hyperbolic* if $|\operatorname{Tr}\gamma| > 2$.

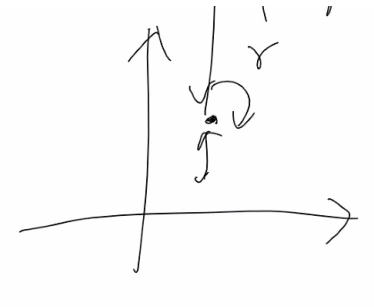
Proposition 1.51

The transformation $\gamma \in \mathrm{PSL}_2(\mathbb{Z})$ is:

1. elliptic iff γ has a fixed point on $\mathbb{H}$;
2. parabolic iff γ has exactly one fixed point on $\partial H = \mathbb{R} \cup \{i\infty\}$;
3. hyperbolic iff γ has two fixed points on $\partial\mathbb{H}$.

We won't prove this, but we will illustrate it.

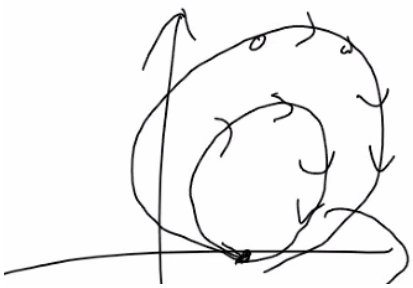
1. *Elliptic case*: an isometry with a fixed point necessarily rotates everything around that fixed point.



One can show that any elliptic element is conjugate to one of $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ and $\begin{pmatrix} 1 & 1 \\ -1 & 0 \end{pmatrix}$.

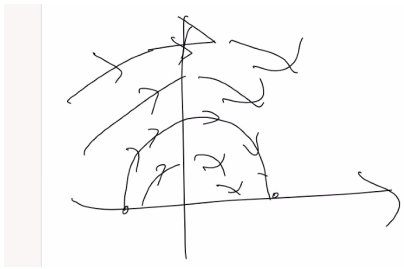
2. *Parabolic case:* suppose $i\infty$ is fixed, as is the case for $\begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix}$. Then the whole plane is just translated horizontally:

But what if the fixed point is located on $\mathbb{R}$? If we draw a circle touching that point, then γ is going to move any point on the circle along the same direction on the circle.



Essentially, these two illustrations are the same, from different reference points. These circles touching the boundary are *horocycles*; in the former picture, these are conceptualized as "horizontal lines touching $i\infty$." The point is that *parabolic transformations move points along horocycles.*

3. *Hyperbolic case:* Suppose γ fixes two points on $\mathbb{R}$. Draw a semicircle between those points; things will move along that direction. Other points will move "parallel" to it, but not along a half-circle. These half-circles are exactly the geodesics on $\mathbb{H}$ equipped with the hyperbolic metric.



One can compute that any $\mathrm{PSL}_2(\mathbb{R})$ action sends these half-circles to one another, which justifies calling them isometries.

Proposition 1.52

On $\overline{\mathcal{F}}$, there are three elliptic fixed points, given by

$$\begin{cases} \rho = \frac{-1+i\sqrt{3}}{2} \\ \rho' = \frac{1+i\sqrt{3}}{2} \\ i \end{cases} \quad \text{fixed by} \quad \begin{cases} ST \\ ST^{-1} \\ S \end{cases}$$

Proof. First, observe that elliptic fixed points would necessarily lie on the boundary of $\mathcal{F}$ (otherwise, by the definition of fundamental domain, they would necessarily be mapped outside of the fundamental domain.) Next, if $|cz+d| = 1$ for $z \neq 0$, so $c, d \in \mathbb{Z}$ with $c \neq 0$, then:

1. First case: $z = \pm\frac{1}{2} + iy$ with $y \geq \sqrt{3}/2$. In this case, we have $1 = |d \pm c/2|^2 + c^2y^2 \geq 1/4 + 3/4 = 1$, so for equality, we need $y = \sqrt{3}/2$. This falls in to one of the first two cases.

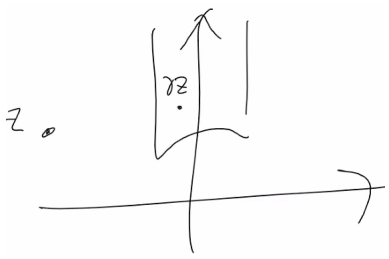
2. Second case: $|z| = 1$ implies $1 = |cz+d|^2 = c^2+2cd+d^2 \geq c^2-|cd|+d^2$. If $cd = 0$, then by assumption $d = 0$, so we have

$$z = \begin{pmatrix} a & -1 \\ 1 & 0 \end{pmatrix} z = \frac{az-1}{z} = a - \frac{1}{z},$$

which implies $z^2 - az + 1 = 0$, and $\Re z = a/2$ has modulus $\leq 1/2$, hence $a = 1, 0, -1$. This falls into one of the three claimed cases. If $cd \neq 0$, then it must be $x = \pm 1/2$. This case falls in to one of the three cases.

Hence we have only three elliptic fixed points on the boundary of $\mathscr{F}$, and this classifies all the elliptic fixed points. □

Suppose we have an elliptic element which fixes z, and get some γz that moves it into the fundamental domain.



If $\beta z = z$, then $(\gamma\beta\gamma^{-1})\gamma z = \gamma z$, so finding all the elliptic points is equivalent (via conjugation) to finding all the elliptic points inside the fundamental domain. So we're done with classifying elliptic fixed points!

(Lecture 4: September 22, 2020)

1.7 Zeros of modular forms

Theorem 1.53

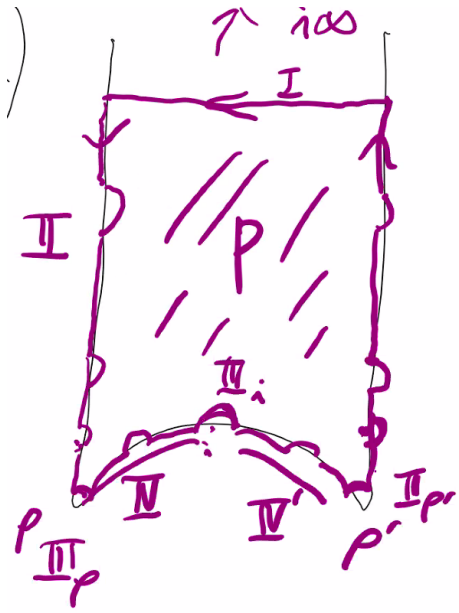
Let $f \neq 0$ be a modular form of weight $k \geq 0$, and write $m_f(w)$ to be the order of f at w. Define

$$\begin{cases} m(w) = 1 & \text{if } w \in \mathbb{H} \text{ is not an elliptic fixed point} \\ m(w) = 2 & \text{if } w = i \\ m(w) = 3 & \text{if } w = \pm\frac{1}{2} + i\frac{\sqrt{3}}{2}. \end{cases}.$$

Then, the following identity holds:

$$\sum_{\omega \pmod{\mathrm{SL}_2(\mathbb{Z})}} \frac{m_f(w)}{m(w)} = \frac{k}{12}.$$

Proof. Consider our fundamental domain $\mathcal{F}$ for $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$, and draw a horizontal line high enough so that the only zero above it is at $i\infty$. (Why is this possible? We assumed meromorphcity at $i\infty$, which means zeros and poles can't accumulate at $i\infty$.) Draw a line down the left side, diverting around it when we hit a zero at the boundary. When we hit the elliptic point ρ, wrap around it in an arc. We do the same thing at i, and then also at ρ'. This process yields a counterclockwise contour. Furthermore, we can do this so that the countours on the LHS and RHS are identical (by $x \mapsto x+1$ periodicity, the zeros are the same on left and right boundary of $\mathcal{F}$.) Label the contours $I, II, III_\rho, IV, IV_i, IV', III_{\rho'}$, and call this region P.



Then

$$\frac{1}{2\pi i}\int_{\partial P}\frac{f'}{f}(z)dz = \sum_{w\in P} m_f(w),$$

so we can write the Fourier expansions

$$f(z) = \sum_{n=m_f(i\infty)}^{\infty} a_n e(nz), \qquad f'(z) = \sum_{n=m_f(i\infty)}^{\infty} 2\pi i n a_n e(nz),$$

which implies that

$$\frac{f'}{f}(z) = \sum_{\ell\geq 0} b_\ell e(\ell z)$$

with $b_0 = 2\pi i m_f(\infty)$, which follows from just checking the power series expansion.

- *Computing integral over contour I:* using this expansion, we compute that

$$\frac{1}{2\pi i}\int_I \frac{f'}{f}(z)dz = -m_f(i\infty)$$

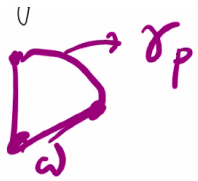
since all the $b_{\ell>0}$ vanish in the integral.

- *Computing integral over contour II:* Next, we compute

$$\frac{1}{2\pi i}\int_{II}\frac{f'}{f}(z)dz = 0$$

since these contours down the left and right side are the same contour in the opposite direction (recall automorphy factor $(cz+d)^k$ is trivial for $Tx = x+1$.)

- *Computing integral over contour III_ρ:* Here we're integrating over the arc γ_ρ:



Around w we can write $f(z) = (z-w)^{m_f(w)}h(z)$, where h is holomorphic with no zeros at $z = w$. As $f'(z) = (z-w)^{m_f(w)}h'(z) + m(z-w)^{m_f(w)-1}h(z)$, this gives

$$\frac{f'}{f}(z) = \frac{m_f(w)}{z-w} + \frac{h'}{h}(z),$$

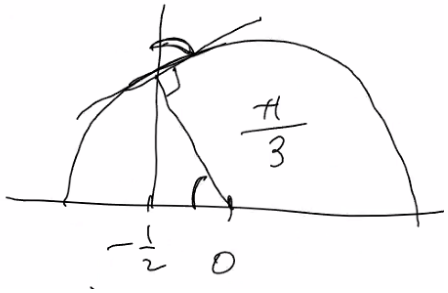
where (h'/h) is holomorphic near $z = w$. Therefore,

$$\frac{1}{2\pi i}\int_{\gamma_p}\frac{f'}{f}(z) = \frac{\mu(\gamma_p)}{2\pi}m_f(w)$$

where $\mu(\gamma_p)$ is the arc angle of the circle. This implies

$$\begin{aligned}\frac{1}{2\pi i}\int_{III_\rho \cup IV_i \cup III_{\rho'}}\frac{f'}{f}(z)dz &= \frac{m_f(\rho)}{2m(\rho)} + \frac{m_f(\rho')}{2m(\rho')} + \frac{m_f(i)}{m(i)} \\ &= \frac{m_f(\rho)}{m(\rho)} + \frac{m_f(i)}{m(i)}\end{aligned}$$

since $\mu(i) = 1/m(i)$ and $\mu(\rho) = \mu(\rho') = 1/2m(\rho)$. Note: from a quick sketch, we can see that the arc angles are $\pi/3$:



- *Computing integral over contours IV and IV':* By automorphy, $f(Sz) = z^k f(z)$ and $f'(Sz)z^{-2} = kz^{k-1}f(z) + z^k f'(z)$, hence

$$\frac{f'}{f}(Sz)z^{-2} = \frac{k}{z} + \frac{f'}{f}(z).$$

Finally we get

$$\frac{1}{2\pi i}\int_{IV'}\frac{f'}{f}(z)dz = -\frac{1}{2\pi i}\int_{IV}\frac{f'}{f}(z)(Sz)d(Sz)$$

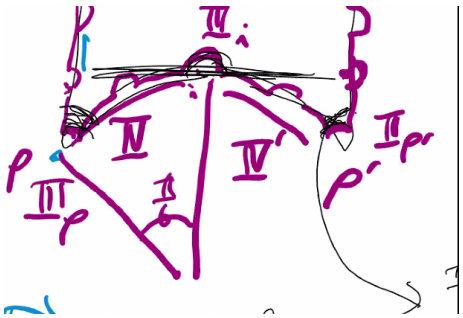
thinking about the action of S geometrically (or just doing change of variable) which yields

$$-\frac{1}{2\pi i}\int_{IV}\left(\frac{k}{z}+\frac{f'}{f}(z)dz\right).$$

Therefore,

$$\frac{1}{2\pi i}\int_{IV\cup IV'}\frac{f'}{f} = \frac{1}{2\pi i}\int_{IV\cup IV'}\frac{k}{z}dz = \frac{k}{12}.$$

as the angle for this arc of integration is $\pi/6$:



The result follows from combining all these computations. □

Remark: this theorem is a simple consequence of Riemann-Roch for the Riemann surface $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$. But the proof we provided has nothing to do with Riemann-Roch, since we're trying to keep this course self-contained.

Remark 1.54

Recall that the three elliptic fixed points are i, ρ, ρ', and in the theorem above, we're defining $m(w)$ to be the order of the elliptic fixed point. Recall that S fixes i, and $\#\langle S\rangle = 2$. In this case, i has order two. And we can check that other fixed points are of order three. **In general, w is an elliptic fixed point of Γ means there is a subgroup of Γ which fixes w, and the order of this subgroup is defined to be the order of w.** So $m(w)$ is just the order of the elliptic fixed point on $\mathbb{H}$.

Here's the takeaway from this: the total quantity of zeros of a modular form of weight k on the full modular surface is exactly $k/12$.

The grand definition of modular forms:

Definition 1.55

f is a modular form of weight k on $\Gamma\backslash\mathbb{H}$ (in our case, $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$) if

1. f is holomorphic on $\mathbb{H}$ and $i\infty$
2. $f(\gamma z) = (cz+d)^k f(z)$

Digression: f isn't properly defined on $\Gamma\backslash\mathbb{H}$, but we say f is a modular form on $\Gamma\backslash\mathbb{H}$ because its values on the fundamental domain determine its values everywhere else. But given a modular form f on $\Gamma\backslash\mathbb{H}$, we can use the Iwasawa decomposition to realize a function truly defined on $\Gamma\backslash \mathrm{PSL}_2(\mathbb{R})$, namely, we consider the function $F(h) := f(z)y^{k/2}e^{ik\theta}$; then, $F(h)$ is actually invariant on $\mathrm{SL}_2(\mathbb{Z})$, so this is the "real" automorphic form living on $\Gamma\backslash\mathbb{H}$.

1.8 The space of modular forms

Definition 1.56

1. M_k is the linear space of modular forms of weight $k \geq 0$.
2. $M = \oplus_{k \text{ even}} M_k$ is the graded algebra of modular forms.

Proposition 1.57

$\Delta(z) \neq 0$ if $z \in \mathbb{H}$, and Δ has a simple zero at $i\infty$.

Proof. We constructed Δ as the discriminant of the elliptic curve $y^2 = 4(x-e_1)(x-e_2)(x-e_3)$ where the e_i's are the values of the Weierstrass $\wp$ function at its half-periods, and we argued (using the fact that $\wp$ has only two zeros and poles on any fundamental parallelogram) that this function has nonzero discriminant. □

Proposition 1.58

$M_k M_l \subseteq M_{k+l}$, which justifies calling M a graded algebra.

Now we'll compute the dimension of M_k, for each k, using our theorem from earlier.

1. If $k = 0$, them $m_f(w) = 0$ for all $w \in \mathbb{H}$ (since modular forms can't have poles in the fundamental domain) which means that f has to be identically zero.
2. If $k = 2$, then since $1/2, 1/3$, and 1 can't add up in any combination to get the $1/6$ in the theorem (note: no negative terms are allowed since there are no poles for modular forms as we've defined them), then no f can satisfy the equation, hence $M_2 = 0$.
3. If $k = 4$, then by
$$\sum_{\omega \pmod{\mathrm{SL}_2(\mathbb{Z})}} \frac{m_f(w)}{m(w)} = \frac{1}{3},$$
the only combinations of $1/2, 1/3$, and 1 that can sum to $1/3$ is just $1/3$. This implies that $m_f(\rho) = 1$, and $m_f(z) = 0$ for $z \not\equiv \rho \pmod{\Gamma}$. We already know E_4 is a modular form of weight 4, which implies that E_4 will satisfy this condition too, by the theorem. We know from the Fourier expansion that $E_4(i\infty) = 1$, which implies that for any $f \in M_4$, there exists some c so that $(f - cE_4)(i\infty) = 0$, hence

$f - cE_4$ is identically zero (because by the theorem, $f - cE_4$ can have a zero at ρ and nowhere else.) This means $f = cE_4$, so $M_4 = \langle E_4 \rangle$.

4. If $k = 6$, then by
$$\sum_{\omega \pmod{\mathrm{SL}_2(\mathbb{Z})}} \frac{m_f(w)}{m(w)} = \frac{1}{2},$$
$m_f(i) = 1$ (since everything has to add up to $1/2$, and the only way to get that is using i because $m(i) = 2$) and every other $m_f(z) = 0$. So by the same reasoning as before, $M_6 = G_6\mathbb{C}$.

5. If $k = 8$, then from the sum needing to equal $2/3$, we get $m_f(\rho) = 2$ and $m_f(z) = 0$ otherwise. In fact, using the same reasoning as before, we get $M_8 = G_4(z)^2\mathbb{C} = G_8(z)\mathbb{C}$. So by considering the constant term, one can actually show $E_4(z)^2 = E_8(z)$ (which yields some cool identities about summations of powers of divisors.)

6. If $k = 10$, then the sum needs to equal $5/6$, and the only way this can happen is via $1/2 + 1/3$, which implies that $m_f(i) = 1$ and $m_f(\rho) = 1$, and $m_f(z) = 0$ otherwise. By the same reasoning as above, this implies M_{10} is one-dimensional, hence $M_{10} = G_4G_6\mathbb{C}$.

7. If $k \geq 12$, then for all $f \in M_k$, there exists a unique $c \in \mathbb{C}$ such that $f(z) - cG_k(z)$ vanishes at $i\infty$. It follows that the quotient
$$\frac{f(z) - cG_k(z)}{\Delta(z)} \in M_{k-12}$$
must be a holomorphic function with weight $k - 12$. This implies that every $f \in M_k$ can be written as $f(z) = \tilde{f}(z)\Delta(z) + cG_k$ where $\tilde{f}(z) \in M_{k-12}$. Therefore,
$$M_k = \Delta(z)M_{k-12} \oplus G_k\mathbb{C}.$$
Using this, we can now compute the dimension of M_k for every k by induction.

Theorem 1.59

For $k \geq 2$ even, we have
$$\dim M_k = \begin{cases} \lfloor k/12 \rfloor & k \equiv 2 \pmod{12} \\ \lfloor k/12 \rfloor + 1 & \text{otherwise.} \end{cases}$$

Proof. For the base cases, we computed above
$$\dim M_0 = 0, \qquad \dim M_2 = 0, \qquad \dim M_4 = 1, \qquad \dim M_6 = 1, \qquad \dim M_8 = 1, \qquad \dim M_{10} = 1.$$
We showed that for $k \geq 12$, we have
$$\dim M_k = \dim M_{k-12} + 1.$$
This finishes the proof. □

Proposition 1.60

M is generated by G_4 and G_6.

Proof. Use $G_4^a G_6^b$ instead of G_k in our direct sum decomposition for G_k above. I.e., we know Δ is a polynomial in G_4, G_6, now just apply $M_k = \Delta(z) M_{k-12} \oplus G_k \mathbb{C}$, and the base cases that $M_0, M_2, M_4, M_6, M_8, M_{10}$ are generated by G_4 and G_6 (as we computed above.) □

Proposition 1.61

G_4 and G_6 are algebraically independent.

Proof. Assume $P(G_4, G_6) = 0$ with $\deg P$ minimal. Then we have either

$$G_4^m + G_6 Q(G_4, G_6) = 0, \qquad \text{or} \qquad G_6^m + G_4 (Q(G_4, G_6) = 0,$$

for if there is no leading term like this, then we can divide out by the Eisenstein series and constant term out front. But $G_6(i) = 0$, which means the first option is impossible since $G_4(i) \neq 0$; as for the second option, $G_4(\rho) = 0 \neq G_6(\rho)$. □

Corollary 1.62

$M = \mathbb{C}[G_4, G_6]$

Definition 1.63

$f \in M_k$ is a *cusp form* if it vanishes at $i\infty$. So we can decompose

$$M_k = E_k \mathbb{C} \oplus S_k,$$

where S_k is the space of cusp forms.

Proposition 1.64

There is no cusp form of weight ≤ 10 on $\mathrm{SL}_2(\mathbb{Z}) \backslash \mathbb{H}$.

Proof. We proved this using our case work above. □

Using 1-dimensionality, it's clear that:

Corollary 1.65

$E_4^2 = E_8$ and $E_4 E_6 = E_{10}$ and $E_6 E_8 = E_4 E_{10} = E_{14}$ and $E_6^2 - E_{12} = c\Delta$.

If we compute what c has to be by comparing the first coefficients in this last relation, we get that

Corollary 1.66

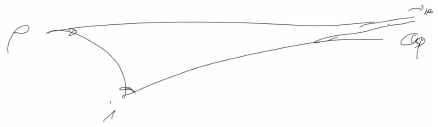
$$\tau(n) = \frac{65}{756}\sigma_{11}(n) + \frac{691}{756}\sigma_5(n) - \frac{691}{3}\sum_{0<m<n}\sigma_5(m)\sigma_5(n-m).$$

Theorem 1.67: (Niebar)

$$\tau(n) = n^4\sigma(n) - 24\sum_{0<m<n} m^2 f(m,n)\sigma(n)\sigma(n-m),$$

where $f(mn) = 35m^2 - 52mn + 18n^2$.

Why the name "cusp form"? If we draw the fundamental domain like a geometer, then it looks like a funnel, since the higher you go up the closer points get to each other (which provides intuition for why geodesics are those strange semicirles that go perpendicular to $\mathbb{R}$! Namely, traveling horizontally close to $\mathbb{R}$ is very expensiive, whereas traveling horizontally far away is much cheaper) so that the "point" at the very end is just a cusp.



(Lecture 5: September 24, 2020)

1.9 Modular forms on congruence subgroups

If you're doing number theory, then most modular forms you concern yourself live on a congruence subgroup of $\mathrm{SL}_2(\mathbb{Z})$.

Definition 1.68

Let N be a positive integer. The *principal congruence subgroup of level* N is

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

i.e., $\Gamma(N)$ is the kernel of the projection map $\mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/N\mathbb{Z})$.

Definition 1.69

$\Gamma \subseteq \mathrm{SL}_2(\mathbb{Z})$ is a *congruence subgroup* if it contains $\Gamma(N)$ for some N. The least such N is the *level* of Γ.

Examples:

1. $\Gamma(1) = \mathrm{SL}_2(\mathbb{Z})$.

2. $\Gamma_0(N) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma(1) : c \equiv 0 \pmod{N} \right\}$. For some analytic number theorists, this is maybe the only congruence subgroup they'll ever see in their research.
3. $\Gamma_1(N) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma(1) : \begin{pmatrix} a & b \\ c & d \end{pmatrix} \equiv \begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix} \pmod{N} \right\}$.
4. So $\Gamma(N) \subseteq \Gamma_1(N) \subseteq \Gamma_0(N) \subseteq \mathrm{SL}_2(\mathbb{Z})$

Note: we only care about subgroups $\Gamma \subseteq \mathrm{SL}_2(\mathbb{Z})$ such that $\mathrm{vol}(\Gamma\backslash\mathbb{H}) < \infty$, or equivalently, finite index subgroups; this equivalence is due to the identity

$$\mathrm{vol}(\Gamma\backslash\mathbb{H}) = [\mathrm{SL}_2(\mathbb{Z}) : \Gamma)] \,\mathrm{vol}(\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}).$$

The volume here is integration of the fundamental domain against the hyperbolic metric. (Remark: since $\mathrm{SL}_2(\mathbb{R})$ is the isometry group of $\mathbb{H}$, every fundamental domain has the same volume.)

Proposition 1.70

$\Gamma(N) \trianglelefteq \Gamma_1(N)$ and $\Gamma_1(N) \trianglelefteq \Gamma_0(N)$.

Proof. Consider the map

$$\Gamma_1(N) \to \mathbb{Z}/n\mathbb{Z} : \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto b \pmod{N}.$$

This has kernel $\Gamma(N)$. Next, consider the map

$$\Gamma_0(N) \to (\mathbb{Z}/n\mathbb{Z})^* : \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto d \pmod{N}.$$

This has kernel $\Gamma_1(N)$. □

Grand definition number two of modular forms:

Definition 1.71

Let Γ be a congruence subgroup of $\mathrm{SL}_2(\mathbb{Z})$. Then $f : \mathbb{H} \to \mathbb{C}$ a *modular form of weight k with respect to Γ* if:

1. f is holomorphic;
2. $f(\gamma z) = (cz + d)^k f(z)$ for any $\gamma \in \Gamma$;
3. $f(\alpha z)$ is holomorphic at $z = i\infty$ for any $\alpha \in \mathrm{SL}_2(\mathbb{Z})$.

We say f is a *cusp form* if, in addition,

4 $\int_0^{n_\alpha} f(\alpha z)dx = 0$ for all $\alpha \in \mathrm{SL}_2(\mathbb{Z})$, where n_α is the width of the cusp.

Why the third condition? In the full modular surface, the only cusp was at $i\infty$, but these congruence subgroups can come with more cusps! For example, on $\Gamma_0(2)\backslash\mathbb{H}$, there is a cusp at $i\infty$ and another at 0. For example, taking $\alpha = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$, then $\alpha : i\infty \to 0$, so asking for $f(\alpha z)$ to be holomorphic at $i\infty$ is the same thing as asking for f to be holomorphic at 0. In other words, this third condition is equivalent to saying there are *no negative terms* in the Fourier expansion of f at any cusp (compare to our go-to expansion, at the cusp $i\infty$.) And the fourth requirement is equivalent to saying that there is *no constant term* in the Fourier expansion at each cusp.

1.10 Theta series

We are concerned with the problem of counting the number of representations of n as a sum of k squares. Towards this, we define

$$r(n,k) := \#\left\{(r_1,\dots,r_k) \in \mathbb{Z}^k : n = r_1^2 + \cdots + r_k^2\right\}.$$

A combinatorial correspondence:

Proposition 1.72

We have

$$r(n, i+j) = \sum_{\ell+m=n} r(\ell,i)r(m,j).$$

Proof. Representing ℓ as a sum of i squares and m as a sum of j squares, adding these together yields $\ell+m$ as a sum of $i+j$ squares. This gives all representations. □

Definition 1.73

The *theta series* is

$$\vartheta(\tau,k) := \sum_{n\geq 0} r(n,k)q^n,$$

where $q = e(\tau)$ and $\tau \in \mathbb{H}$.

Fact 1.74

For $\tau \in \mathbb{H}$, $\vartheta(\tau,k)$ converges absolutely.

Proof. If $r_1^2 + \cdots + r_k^2 = n$ then each $|r_i| \leq \sqrt{n}$, which gives an upper bound

$$r(n,k) \leq (2\sqrt{n}+1)^k \leq (3\sqrt{n})^k \leq 3^k n^{k/2},$$

so writing $\tau = x + iy$, we have

$$\sum_{n\geq 0} r(n,k)|q^n| \leq 3^k \sum_{n\geq 0} n^{k/2}(e^{-2\pi y})^n.$$

Finally, observe that the summands are eventually bounded termwise by R^n for any fixed $R \in (e^{-2\pi y}, 1)$, since

$$n^{k/2}(e^{-2\pi y})^n \leq R^n \iff n^{k/2} \leq \left(\frac{R}{e^{-2\pi y}}\right)^n,$$

which eventually holds because the RHS grows exponentially in n whereas the LHS grows polynomially. □

Corollary 1.75

$\vartheta(\tau, i+j) = \vartheta(\tau, i)\vartheta(\tau, j)$

Proof. We consider the power series expansions:

$$\sum_{n\geq 0} r(n, i+j)q^n = \left(\sum_{\ell \geq 0} r(\ell, i)q^n\right)\left(\sum_{m\geq 0} r(m, j)q^n\right).$$

We multiply out (justified by absolute convergence) and apply our combinatorial proposition. □

We now investigate transformation properties of ϑ. Note that

$$\vartheta(\tau) := \vartheta(\tau, 1) = \sum_{n\geq 0} r(n, 1)q^n = \sum_{n\in\mathbb{Z}} q^{n^2},$$

since $r(n, 1) = 2$ is n is a square, and 0 otherwise. Note: we saw this function when we found the analytic continuation of ζ; in fact, we found a functional equation using Poisson summation. We bootstrap that result to show the following:

Lemma 1.76

$$\vartheta\left(\begin{pmatrix} 0 & -1 \\ 4 & 0 \end{pmatrix}\tau\right) = \sqrt{-2i\tau}\vartheta(\tau). \tag{1.3}$$

Proof. Recalling the first lecture, if we define $\psi(x) = \sum_{n\geq 1} e^{-n^2\pi x}$, we showed that

$$1 + 2\psi(x) = \frac{1}{\sqrt{x}}(2\psi(1/x) + 1).$$

As $\psi(-2i\tau) = \sum_{n\geq 1} e^{2\pi i n^2 \tau}$, we get that

$$\vartheta(\tau) = 2\psi(-2i\tau) + 1.$$

So by the transformation law with $x = -2i\tau$, we get

$$1 + 2\psi(-2i\tau) = -\frac{1}{\sqrt{-2i\tau}}\left(2\psi(\frac{1}{-2i\tau} + 1)\right).$$

The LHS is clearly $\vartheta(\tau)$, and on can compute that parenthesis on the RHS contain the term $\vartheta(-1/4\tau)$. □

Now, because

$$\begin{pmatrix} 1 & 0 \\ 4 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1/4 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 4 & 0 \end{pmatrix},$$

we have

$$\vartheta\left(\begin{pmatrix} 1 & 0 \\ 4 & 1 \end{pmatrix} \tau\right) = \vartheta\left(\begin{pmatrix} 0 & 1/4 \\ -1 & 0 \end{pmatrix} \tau'\right), \qquad \tau' := \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 4 & 0 \end{pmatrix} \tau = -\frac{1}{4\tau} - 1.$$

We can compute that

$$\begin{aligned} \vartheta\left(\begin{pmatrix} 0 & 1/4 \\ -1 & 0 \end{pmatrix} \tau'\right) &= \sqrt{-2i\tau'}\vartheta(\tau') \\ &= \sqrt{2i(1/4\tau+1)}\vartheta\left(\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 4 & 0 \end{pmatrix} \tau\right) \\ &= \sqrt{2i(1/4\tau+1)}\vartheta\left(\begin{pmatrix} 0 & -1 \\ 4 & 0 \end{pmatrix} \tau\right) \\ &= \sqrt{2i(1/4\tau+1)(-2i\tau)}\vartheta(\tau) \\ &= \sqrt{4\tau+1}\vartheta(t), \end{aligned}$$

where we used the fact that $\vartheta(\tau) = \vartheta(\tau+1)$. What's the point?

We can use this to get the transformation law for $\theta(\tau, n)$ for $n \geq 1$, as follows:

Fact 1.77

We have

$$\vartheta\left(\begin{pmatrix} 1 & 0 \\ 4 & 1 \end{pmatrix} \tau, 4\right) = (4\tau+1)^2\vartheta(\tau, 4).$$

Thus, $\vartheta(\tau, 4)$ is a modular form of weight 2 with respect to the subgroup

$$\Gamma_\vartheta := \left\langle -I, \begin{pmatrix} 1 & 0 \\ 4 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \right\rangle.$$

Proof. By our corollary, $\vartheta(\tau, 4) = \vartheta(\tau)^4$, hence

$$\vartheta\left(\begin{pmatrix} 1 & 0 \\ 4 & 1 \end{pmatrix} \tau, 4\right) = \vartheta\left(\begin{pmatrix} 1 & 0 \\ 4 & 1 \end{pmatrix} \tau, 1\right)^4 = (4\tau+1)^2\vartheta(\tau, 1)^4 = (4\tau+1)^2\vartheta(\tau, 4).$$

And ϑ clearly transforms as it ought to with respect to the other two generators. □

It turns out that this group is one of our congruence subgroups:

Proposition 1.78

In fact, we have $\Gamma_\vartheta = \Gamma_0(4)$.

Proof. Take $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(4)$. We first compute that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a & b' \\ c & nc+d \end{pmatrix}, \qquad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 4n & 1 \end{pmatrix} = \begin{pmatrix} a' & b \\ c+4nd & d \end{pmatrix}$$

The point: each step can make $\min\{|c|, 2|d|\}$ strictly smaller if $c, d \neq 0$. This must stop when c or d equals zero. But then no $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(4)$ satisfies $d = 0$ (since $c \equiv d \equiv 0 \pmod 4$ implies the determinant is zero modulo 4) hence $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is a product of $\begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix}$'s and $\begin{pmatrix} 1 & 0 \\ 4n & 1 \end{pmatrix}$'s with some $\begin{pmatrix} * & * \\ 0 & * \end{pmatrix} \in \Gamma_0(4)$, but this matrix is necessarily some $\begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix}$. It follows that $\Gamma_\vartheta \supseteq \Gamma_0(4)$. The other direction is clear. □

To be fully rigorous in verifying that ϑ is a modular form on $\Gamma_0(4)$, one needs to check the Fourier expansion at each cusp. As there are three cusps, one needs to check the other two. This is super tedious, but not very difficult.

1.11 The weight two Eisenstein series

Now, we'll define weight two Eisenstein series. We showed (using the valence formula) that there is no weight 2 modular form on the whole modular surface. However, we define the next best thing here:

Definition 1.79

The *weight two Eisenstein series* is

$$G_2(z) := \sum_{c \in \mathbb{Z}} \sum_{d \in \mathbb{Z}'_c} \frac{1}{(cz+d)^2},$$

where $\mathbb{Z}'_c = \mathbb{Z} - \{0\}$ if $c = 0$, and $\mathbb{Z}'_c = \mathbb{Z}$ otherwise.

This series actually conditionally converges. But since we don't have absolute convergence, something weird happens. One can show, with some effort:

Proposition 1.80

$G_2(\gamma z)(cz+d)^{-s} = G_2(z) - \frac{2\pi i c}{cz+d}$.

Although $G_2(z)$ is not a modular form on the full modular group, but we can fix it:

Proposition 1.81

The function $G_{2,N}(z) := G_2(z) - NG_2(Nz)$ is a modular form in $M_2(\Gamma_0(N))$.

Proof. Follows directly from previous proposition. □

This yields:

Proposition 1.82

$G_{2,2}, G_{2,4} \in M_2(\Gamma_0(4))$

One can compute $\dim M_2(\Gamma_0(4))$ using the same method as before, but this technique becomes unweildy with contour integration on these congruence subgroups, so the better way to go is to use Riemann Roch (see Diamond-Sherman 3.9.)

Corollary 1.83

$\vartheta(\tau, 4) = -\frac{1}{\pi^2} G_{2,4}(\tau)$.

Proof. This can be seen by comparing Fourier expansions (we only need to compare the first three Fourier coefficients to see that they're equal, since this is 2-dimensional space.) □

By comparing Fourier coefficients, we get:

Corollary 1.84

$8 \sum_{d|n, 4\nmid d} d = r(n, 2)$ for $n \geq 1$.

This is one way to prove the **four-square theorem**, which says we can represent every integer as a sum of four squares. *This is one very nice application of modular forms to a problem in elementary number theory.* There is a guiding principle that modular forms encode a lot of arithmetic information, despite the fact that their origin is very analytic and geometric. The motivation for continuing Eisenstein series to weight 2 was to try to explore this.

It turns out G_2 has a Fourier expansion very similar to $G_{k \geq 4}$.

Proposition 1.85

We have

$$G_2(\tau) = 2\zeta(2) - 8\pi^2 \sum_{n \geq 1} \sigma_1(n) q^n.$$

Furthermore,

$$E_2(\tau) := \frac{G_2(\tau)}{2\zeta(2)} = 1 - 24 \sum_{n \geq 1} \sigma_1(n) q^n.$$

And these converge absolutely.

This implies

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

Definition 1.86

The *Dedekind eta function* is

$$\eta(\tau) := e(\tau/24)\prod_{n\geq 1}(1-q^n).$$

Proposition 1.87

$S(\tau) := \sum_{n\geq 1}\log(1-q^n)$ converges absolutely and uniformly on compact subsets of $\mathbb{H}$, hence $\eta(\tau)$ is holomorphic on $\mathbb{H}$.

Proposition 1.88

We have

$$\eta(-1/\tau) = \sqrt{-i\tau}\eta(\tau).$$

Proof. We compute

$$\begin{aligned}
\frac{\eta'}{\eta}(\tau) &= \frac{d}{d\tau}\log\eta(\tau)\\
&= \frac{d}{d\tau}\left(e(\tau/24)+\sum_{n\geq 1}(1-q^n)\right)\\
&= \frac{\pi i}{12} - 2\pi i\sum_{d\geq 1}\frac{dq^d}{1-q^d}\\
&= \frac{\pi i}{12} - 2\pi i\sum_{d\geq 1} d\sum_{m\geq 1} q^{dm}\\
&= \frac{\pi i}{12} - 2\pi i\sum_{m\geq 1}\sum_{d\geq 1} dq^{dm}\\
&= \frac{\pi i}{12} - 2\pi i\sum_{n\geq 1}\sigma_1(n)q^n\\
&= \frac{\pi i}{12}E_2(\tau),
\end{aligned}$$

which implies that

$$\frac{d}{d\tau}(\log(\eta(-1/\tau))) = \frac{\pi i}{12}\tau^{-2}E_2(-1/\tau).$$

This in turn implies that

$$\frac{d}{dt}(\log(\sqrt{-i\tau}\eta(\tau))) = \frac{1}{2\tau} + \frac{\pi i}{12}E_2(\tau) = \frac{\pi i}{12}(E_2(\tau) + \frac{12}{2\pi i\tau}),$$

so we can infer that

$$\eta(-1/\tau) = c\sqrt{-i\tau}\eta(\tau)$$

for some $c\in\mathbb{C}$. Set $z=i$, we get that $c=1$ and the proposition follows. □

Corollary 1.89

Clearly, $\eta^{24}(\tau)$ is invariant under $\tau \mapsto \tau + 1$, and $\tau^{24}(-1/\tau) = \tau^{12}\eta^{24}(\tau)$. Therefore $\eta^{24} \in S_{12}(\Gamma(1))$. Hence $\Delta = \eta^{24}$, since this space is 1 dimensional (just compare Fourier coefficients.) This proves the famous product formula

$$\Delta = q\prod_{n\geq 1}(1-q^n)^{24} = \sum_{n\geq 1}\tau(n)q^n.$$

In the 19th century, people were obsessed with understanding elliptic integrals, which is why all these things were defined. Coincidentally, they found these mysterious expressions.

So what are $\theta(\tau)$ and $\eta(\tau)$? These are modular forms of weight $1/2$, which are called *half-integral weight modular forms.* These are very mysterious, and understanding these has been a huge industry since the 1980s. Iwaniec had a breakthrough of estimating Fourier coefficients, which led to a breakthough on Linnik's conjecture on equidistribution on the unit sphere of solutions to quadratic equations. Since then, Duke has had many beautiful results in this area. *It turns out that every time you investigate half-integral weight modular forms, you get a random associated beautiful result in geometry/number theory.*

(Lecture 6: September 29, 2020)

1.12 The Hecke operators

We proved that $M = \mathbb{C}[G_4, G_6]$, which on one hand, might be evidence that the space of modular forms isn't so interesting. However, the addition of Hecke operators acting on this space makes everything much more interesting.

We will first consider the general situation. Suppose we have a group G acting on a space X. (We can think of $\mathrm{SL}_2(\mathbb{R}) \curvearrowright \mathbb{H}$; the aim here is to define Hecke operators acting on $L^2(\Gamma\backslash\mathbb{H})$.) Then a discrete subgroup $\Gamma \leq G$ acts discontinuously on X.

Definition 1.90

The *commensurator* subgroup Γ of G is defined to be

$$\mathrm{COM}(\Gamma) := \left\{g \in G : \Gamma \cap g^{-1}\Gamma g \text{ has finite index both in } \Gamma \text{ and } g^{-1}\Gamma g\right\}.$$

(Note that $\mathrm{COM}(\Gamma) \supseteq \Gamma$.)

For $g \in \mathrm{COM}(\Gamma)$, let $\Gamma_g := \Gamma \cap g^{-1}\Gamma g$. As $\Gamma_g\backslash\Gamma$ is a finite collection of cosets, always write Γ as a disjoint union

$$\Gamma = \bigsqcup_{j\in F_g}\Gamma_g\delta_j,$$

where F_g is some set of indices, $\#F_g = [\Gamma : \Gamma_g]$, and δ_j are the right coset representatives for right cosets of

Γ_g in Γ. Correspondingly, we define the operator

$$T_g : L^2(\Gamma\backslash X) \to L^2(\Gamma\backslash X) : f(-) \mapsto \sum_{j\in F_g} f(g\delta_j -).$$

We need to check that T_g is well-defined, when acting on $L^2(\Gamma\backslash X)$. By hypothesis, this is a finite sum, so we only need to show that if $f : X \to X$ is Γ-invariant, then $T_g f$ is as well.

Proposition 1.91

If $f \in L^2(\Gamma\backslash X)$, then for any $g \in \mathrm{COM}(\Gamma)$, we have $T_g f \in L^2(\Gamma\backslash X)$ as well.

Proof. By definition,

$$(T_g f)(\gamma x) = \sum_{j\in F_g} f(g\delta_j \gamma x).$$

Since $\gamma \in \Gamma$, we have $\delta_j\gamma = \delta'_j\delta_{\pi(j)}$ for some permutation π on F_g, and some $\delta_j \in \Gamma_g$ (to see this, act by γ on the right of the coset decomposition.) This implies

$$\begin{aligned}
\sum_{j\in F_g} f(g\delta_j\gamma x) &= \sum_{j\in F_g} f(g\delta'_j\delta_{\pi(j)}x) \\
&= \sum_{j\in F_g} f(g(g^{-1}\mu_j g)\delta_{\pi(j)}x) \qquad \text{where } \mu_j \in \Gamma, \text{ since } \delta_j \in g^{-1}\Gamma g \\
&= \sum_{j\in F_g} f(\mu_j g\delta_{\pi(j)}x) \\
&= \sum_{j\in F_g} f(g\delta_{\pi(j)}(x)) \\
&= (T_g f)(x),
\end{aligned}$$

hence $T_g f$ is invariant under the Γ-action.

It only remains to show that $T_g f$ is square-integrable on the fundamental domain. As $T_g f$ is a finite sum of square integrable functions, this is clear using the triangle inequality. □

Remark: one can show that T_g commutes with every invariant differential operator. So we can consider joint eigenfunctions of T_g and all invariant differential operators, which we'll see are parameterized by Maass forms, if we're working on $\mathrm{SL}_2(\mathbb{R})$.

Now, let us restrict to the case $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$ and $G = \mathrm{PSL}_2(\mathbb{R})$. **This isn't exactly the case of modular forms on $\mathbb{H}$, but we'll see in the next proposition precisely how it is equivalent.** In this case, set $g := \begin{pmatrix} n & 0 \\ 0 & 1 \end{pmatrix}$, then one can show $T_n := T_g$ is given by

$$T_n : \mathrm{Fun}(\mathrm{PSL}_2(\mathbb{Z})\backslash \mathrm{PSL}_2(\mathbb{R})) \to \mathrm{Fun}(\mathrm{PSL}_2(\mathbb{Z})\backslash \mathrm{PSL}_2(\mathbb{R})) : f(-) \mapsto \sum_{\substack{ad=n \\ b \pmod d}} f\left(\begin{pmatrix} a & b \\ 0 & d \end{pmatrix} -\right).$$

How to verify this? The idea is to show that the sum parameterizes coset representatives for

$$\Gamma_g \backslash \Gamma = \left(\mathrm{PSL}_2(\mathbb{Z}) \cap \begin{pmatrix} 1 & -n \\ 0 & 1 \end{pmatrix} \mathrm{PSL}_2(\mathbb{Z}) \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix} \right) \backslash \mathrm{PSL}_2(\mathbb{Z}).$$

One can show directly that these are in fact the only Hecke operators that we can define on this G, i.e. every Hecke operator will be equivalent to one of these T_n.

Proposition 1.92

Let $\Gamma \leq \mathrm{PSL}_2(\mathbb{R})$ be a congruence subgroup, and consider the Iwasawa decomposition on $\mathrm{SL}_2(\mathbb{R})$:

$$\mathrm{SL}_2(\mathbb{R}) \ni g = \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{y} & 0 \\ 0 & 1/\sqrt{y} \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Then, the following are equivalent:

1. $f(z) \in M_k(\Gamma)$, i.e., f is a modular form of weight k for Γ.
2. $F(g) := y^{k/2} f(x+iy) e^{-ki\theta}$ is a Γ-invariant function on $\mathrm{PSL}_2(\mathbb{R})$.

Proof. Do the computation in Iwasawa coordinates. □

This correspondence carries over to our discussion of Hecke operators as well.

Proposition 1.93

Via the above correspondence, the below maps are equivalent:

1. $F \mapsto T_n F$,
2. $f \mapsto n^{k-1} \sum_{b \pmod{d}} \sum_{ad=n} d^{-k} f\left(\frac{az+b}{d}\right)$.

Example: if $n = 4$, then

$$\begin{aligned} T_4 F &= \sum_{b \pmod{d}} \sum_{ad=4} F((ax+b)/d) \\ &= F((4z+0)/1) + F((2z+0)/2) + F((2z+1)/2) \\ &\quad + F((z+0)/4) + F((z+1)/4) + \cdots + F((z+3)/3). \end{aligned}$$

In this case, we're acting on a function living on $\mathrm{SL}_2(\mathbb{R})$, than the case of a function living on $\mathbb{H}$.

Let us return to the action of the Hecke operators on classical modular forms. We present the formal definition now:

Definition 1.94

To define the action of the Hecke operator on the space of *modular forms* $T_n : M_k \to M_k \ldots$

1. For $f \in M_k$, define a function on $\mathrm{PSL}_2(\mathbb{R})$ by setting $F(x, y, \theta) = e^{-ki\theta} y^{k/2} f(z)$.
2. Then $T_n F$ is another $\mathrm{SL}_2(\mathbb{Z})$-invariant function; one can show that, in fact, $(T_n F)(x, y, \theta) = e^{-ki\theta} y^{k/2} g(z)$ for some function g depending only on z.
3. As $T_n F$ is invariant under $\mathrm{SL}_2(\mathbb{Z})$, $g(z)$ is a modular form of weight k.

Now we collect some fundamental facts about these Hecke operators acting the space of modular forms:

Proposition 1.95

The Hecke operators $T_n : M_k \to M_k$ satisfy:

1. $T_n T_m = T_{nm}$ whenever $(n, m) = 1$.
2. $T_n T_m = T_m T_n$ for all m, n.
3. $T_{p^n} T_p = T_{p^{n+1}} + p^{k-1} T_{p^{n-1}}$
4. If $f \in M_k$, then in fact $T_n f \in M_k$ as well; if $f \in S_k$, then $T_n f \in S_k$.

Proof. The first three points follow from some relatively tedious direct computations. For the fourth, if F is invariant under Γ, then $T_n F$ is also invariant under Γ. So $T_n f$ transforms like a weight k function on $\mathbb{H}$; so to check this last point, we only need to check that the resulting function is holomorphic. And for this, it suffices to verify that the Fourier expansion is nice, which we do in the below lemma (which also shows that f is a cusp form implies $T_n f$ is as well.) □

Lemma 1.96

If $f = \sum_{n \geq 0} a_n q^n \in M_k$, then

$$T_n f = \sum_{m \geq 0} b_m q^m, \qquad \text{where } b_m := \sum_{d \mid (n,m)} d^{k-1} a_{mn/d^2}.$$

In particular, $f \in S_k$ implies $T_n f \in S_k$.

Proof. We compute that

$$\begin{aligned} T_n f(z) &= n^{k-1} \sum_{\substack{ad=n \\ b \pmod d}} d^{-k} f\left(\frac{az+b}{d}\right) \\ &= n^{k-1} \sum_{\substack{ad=n \\ b \pmod d}} \sum_{\ell \geq 0} d^{-k} a_\ell e\left(\frac{az+b}{d}\ell\right) \\ &= n^{k-1} \sum_{ad=n} \sum_{\ell \geq 0} d^{1-k} a_\ell e^{2\pi i a z \ell/d} \frac{1}{d} \sum_{b=0}^{d-1} (e^{2\pi i \ell/d})^b \\ &= \sum_{ad=n} \sum_{\ell' \geq 0} a^{k-1} a_{d\ell'} q^{a\ell'}. \end{aligned}$$

Writing $d = n/a$, this is

$$\sum_{ad=n} \sum_{\ell \geq 0} a^{k-1} a_{d\ell} q^{a\ell} = \sum_{\substack{a \mid n \\ \ell > 0}} a^{k-1} a_{n\ell/a} q^{a\ell}.$$

What is the coefficient of q^m? Setting $m = a\ell$, we can rewrite this as

$$\sum_{\substack{a \mid n \\ \ell \geq 0}} a^{k-1} a_{n\ell/a} q^{a\ell} = \sum_{\substack{a \mid n \\ m/a \geq 0}} a^{k-1} a_{nm/a^2} q^m = \sum_{\substack{a \mid n \\ a \mid m}} a^{k-1} a_{nm/a^2} q^m = \sum_{a \mid (m,n)} a^{k-1} a_{nm/a^2} q^m,$$

which is exactly what was to be shown.

In particular, we have

$$b_0 = \sum_{d \mid n} d^{k-1} a_0,$$

so f is cuspidal implies $T_n f$ is as well. □

In particular:

Lemma 1.97

If $f(z) = \sum_{n \geq 0} a_n q^n \in M_k$ is an eigenfunction of T_n with Hecke eigenvalue λ_n, then

$$a_n = \lambda_n a_1.$$

Proof. Write $T_n f = \sum_{m \geq 0} b_m q^m$. On one hand, we have

$$b_1 = \sum_{d \mid (n,1)} d^{k-1} a_{n/d^2} = a_n.$$

On the other hand, $T_n f = \lambda_n f$ implies $b_m = \lambda_n a_m$. Setting $m = 1$ yields the lemma. □

Lemma 1.98

If $f(z) = \sum_{n \geq 1} a_n q^n \in S_k$ is a cusp form that is also a joint eigenfunction of all the Hecke operators, then

$$f(z) = a_1 \sum_{n \geq 1} \lambda_n q^n.$$

Proof. If $T_n f = \lambda_n f$ for all n, then $a_n = \lambda_n a_1$ for all n. □

What's so cool about being a joint eigenfunction of all the Hecke operators? We have $T_n T_m f = \lambda_n \lambda_m f$, and $T_{nm} f = \lambda_{nm} f$, so $(n, m) = 1$ implies $\lambda_n \lambda_m = \lambda_{mn}$. This is the multiplicativity of Hecke eigenvalues:

Lemma 1.99

Assume $f \in M_k$ is a joint eigenfunction of all the Hecke operators $\{T_n\}$, with associated Hecke eigenvalues $\{\lambda_n\}$. Then,

$$\lambda_n \lambda_m = \lambda_{mn} = \lambda_m \lambda_n$$

for all $(m, n) = 1$.

Note: $\{T_n\} : M_k \to M_k$ is a commuting family of linear operators acting on a finite dimensional vector space. *From linear algebra, it follows that the Hecke operators are simultaneously diagonalizable on the space of holomorphic modular forms.* Therefore, it makes sense to talk about joint eigenfunctions of all T_n, and M_k is in fact spanned by joint eigenfunctions of all T_n. Any such function is called a holomorphic Hecke eigenform.

Definition 1.100

$f \in M_k$ is a *holomorphic Hecke eigenform* if it is a joint eigenfunction of all T_n.

So our lemma abose says that *for any holomorphic Hecke eigenform, we can write down the Fourier expansion in terms of the eigenvalues.*

Note that 3 above, applied to f, gives

$$\lambda_{p^n} \lambda_p f = \lambda_{p^{n+1}} f + p^{k-1} \lambda_{p^{k-1}} f,$$

so $f \neq 0$ implies we get a recursive formula for Hecke eigenvalues:

Lemma 1.101

Let $f(z) \in M_k$ be a holomorphic Hecke eigenfunction. Then the Hecke eigenvalues corresponding to f satisfy the recursive formula

$$\lambda_{p^n} \lambda_p = \lambda_{p^{n+1}} + p^{k-1} \lambda_{p^{n-1}}.$$

Some remarks:

1. A take-away from this: *the Hecke eigenvalues are determined by their values at the primes.*

2. An important point: Hecke eigenvalues are real. This follows from self-adjointness of all the Hecke operators, which in turn follows from the original definition using the $g\delta_j$. The adjoint is just given by taking the inverse of $g\delta_j$. I.e., the point is to find T_g' such that $\langle T_g f_1, f_1 \rangle = \langle f_1, T_g' f_2 \rangle$. So the adjoint action is given by shifting f_2 by the inverses of $g\delta_j$ and summing over j. A straightforward but important step in this argument is to show that $g(g^{-1}\delta_j^{-1}g)$ also gives coset representatives.

What we explained today is in fact a solution to two of the Ramanujan conjectures!!

Corollary 1.102

1. $\tau(n)$ is multiplicative on relatively prime arguments.
2. $\tau(n)$ satisify the recursion $\tau(p)\tau(p^n) = \tau(p^{n+1}) + p^{11}\tau(p^{n-1})$.

Proof. Recall that S_{12} is one-dimensional. Since $\{T_n\}$ is a commuting family of linear operators acting on this vector space, we can conclude that the generator Δ is a joint eigenfunction of all T^n, and therefore Δ is a holomorphic Hecke cusp form. The coefficients are normalized to $\tau(1) = a(1) = 1$, so the Hecke eigenvalues are precisely $\tau(n)$. That is, each Hecke operator satisfies

$$T_n : S_{12} \to S_{12} : \Delta \mapsto \tau(n)\Delta.$$

Thus, $\tau(n)$ satisfies every relation satisfied by Hecke eigenvalues. □

Previously, we saw that modular forms are in fact polynomials in the Eisenstein series G_4 and G_6. Now, we're interested in the case where such polynomials are joint eigenfunctions of all the Hecke operators.

Definition 1.103

We define the *normalized Hecke operators* to be

$$\mathcal{T}_n := n^{-(k-1)/2} T_n.$$

The basic results about composing these maps carry over to the normalized setting:

1. $\mathcal{T}_n \mathcal{T}_m = \mathcal{T}_{nm}$ when $(n, m) = 1$
2. $\mathcal{T}_{p^n} \mathcal{T}_p = \mathcal{T}_{p^{n+1}} + \mathcal{T}_{p^{n-1}}$.
3. More generally, we have
$$\mathcal{T}_n \mathcal{T}_m = \sum_{d|(n,m)} \mathcal{T}_{nm/d^2},$$
and $\lambda_n = n^{(k-1)/2}\tilde{\lambda}_n$, where $\tilde{\lambda}_n$ is the *normalized Hecke eigenvalue.*

Ramanujan conjectured was that the τ function satisfies the following bound:

$$|\tau(p)| \leq 2p^{11/2}.$$

Using the above normalization, Ramanujan's conjecture becomes $|\tilde{\tau}(p)| \leq 2$. And the generalized Ramanujan conjecture is the following:

Conjecture 1.104: Ramanujan conjecture

The normalized Hecke eigenvalues $\tilde{\lambda}_p$ satisfy the bound

$$|\tilde{\lambda}_p| \leq 2.$$

In fact, **we know this conjecture holds for normalized Hecke eigenvalues corresponding to holomorphic Hecke eigenforms**, by the work of Deligne which resolved the Riemann Hypothesis for function fields.

Proposition 1.105

Let $\{\tilde{\lambda}_n\}$ be the collection of normalized Hecke eigenvalues corresponding to a holomorphic Hecke eigenform f. By the Ramanujan conjecture, we can write $\tilde{\lambda}_p = 2\cos\theta_p$ for some $\theta_p \in [0, \pi)$. Then,

$$\tilde{\lambda}_{p^n} = \frac{\sin((n+1)\theta_p)}{\sin\theta_p}.$$

Proof. The normalized Hecke eigenvalues satisfy the recurrence $\tilde{\lambda}_{p^{k+1}} + \tilde{\lambda}_{p^{k-1}} = \tilde{\lambda}_{p^k}\tilde{\lambda}_p$. Denoting $a_k := \tilde{\lambda}_{p^k}$, we can rewrite this recurrence as $a_{k+1} + a_{k-1} = a_k a_1$, with $a_0 = 1$ (since f is normalized to have $\tilde{\lambda}_1 = 1$) and $a_1 = e^{i\theta} + e^{-i\theta}$ (since this is exactly $\tilde{\lambda}_p = 2\cos\theta_p$). The characteristic polynomial is

$$x^2 - (e^{i\theta} + e^{-i\theta})x + 1 = 0,$$

which gives solutions $a_k = \alpha e^{ik\theta} + \beta e^{-ik\theta}$. □

Definition 1.106

Let $\{\tilde{\lambda}_n\}$ be the collection of normalized Hecke eigenvalues corresponding to a holomorphic Hecke eigenform f, and write $\tilde{\lambda}_p = 2\cos\theta_p$. Then the *Satake parameter of f at p* is defined to be $e^{i\theta_p}$.

Remark: the Sato–Tate conjecture is precisely a statement about the distribution of Satake parameters, which is why this looks like the setup for the Sato–Tate conjecture.

1.13 The L-function corresponding to a holomorphic Hecke eigenform

Proposition 1.107

Let $f \in S_k$ be a holomorphic Hecke eigenform, normalized so the first Fourier coefficient is λ_1. Let

$$L(f,s) := \sum_{n\geq 1} \frac{\tilde{\lambda}_n}{n^s}, \qquad \Lambda(f,s) := \frac{\Gamma(s + \frac{k-1}{2})}{(2\pi)^s} L(f,s).$$

Then:

1. $L(f,s)$ converges absolutely for $\Re(s) > 3/2$.
2. $\Lambda(f,s)$ has analytic continuation to all of $\mathbb{C}$.
3. $\Lambda(f,s)$ satisfies the functional equation $\Lambda(f,s) = i^k \Lambda(f, 1-s)$.

Proof. For (1), as $f(z) = \sum_{n\geq 1} \lambda_n q^n \in S_k$, the n'th Fourier coefficient of f satisfies the generic cusp form bound $\lambda_n \ll n^{k/2}$ (we prove this in a future lecture). Because $\lambda_n = n^{\frac{k-1}{2}}\tilde{\lambda}_n$, this implies $\tilde{\lambda}_n \ll n^{1/2}$. Therefore $L(f,s)$ indeed converges absolutely for $\Re(s) > 3/2$.

Towards showing (2), we'll compute the Mellin transform of f along the vertical half line $\{iy : y > 0\}$. We have

$$\int_0^\infty f(iy)y^s \frac{dy}{s} = \int_0^\infty \sum_{n\geq 1} \lambda_n e^{-2\pi n y} y^s \frac{dy}{y} = \sum_{n\geq 1} \frac{\lambda_n}{(2\pi n)^s}\Gamma(s) = \frac{\Gamma(s)}{(2\pi)^s}\sum_{n\geq 1}\frac{\tilde{\lambda}_n}{n^{s-\frac{k-1}{2}}}.$$

(Remark: the Gamma function is just the Mellin transform of e^{-x}. So of course applying the Mellin transform to Fourier expansion of a modular form specialized to $z = iy$ will yield a Gamma factor.) But this implies that

$$\frac{\Lambda(f, s-\frac{k-1}{2})}{(2\pi)^{\frac{k-1}{2}}} = \int_0^\infty f(iy)y^{s-1}dy.$$

The integral on the RHS converges absolutely for $s \in \mathbb{C}$. Why is this? The point is that f is a cusp form, so $f(iy) \to 0$ exponentially fast as $y \to \infty$. Rigorously, if $|\lambda_n| \ll n^c$, then for any sufficiently small $\epsilon > 0$, we have

$$\sum_{n\geq 1} |\lambda_n| e^{-2\pi n y} \ll \sum_{n\geq 1} n^c (e^{-2\pi y})^n \ll \sum_{n\geq 1} (e^{-2\pi y} + \epsilon)^n \ll \frac{e^{-2\pi y} + \epsilon}{1 - (e^{-2\pi y} + \epsilon)} \ll e^{-2\pi y} + \epsilon.$$

Choosing $\epsilon = e^{-2\pi y}/M$, where $M = M(y)$ is large enough so that $(M+1)e^{-2\pi y}/M < 1$, this implies

$$\int_0^\infty \sum_{n\geq 1} \lambda_n e^{-2\pi n y} y^s \frac{dy}{y} \ll \int_0^\infty e^{-2\pi y} y^{s-1} dy,$$

which converges for all s. This argument also illustrates why f needs to be a cusp form for this construction to work. It follows immediately that $\Lambda(f,s)$ has analytic continuation to all of $\mathbb{C}$.

Next we show (3). We know $f(-1/z) = z^k f(z)$. Restricting to the imaginary axis $z = iy$, this implies $f(i/y) = (iy)^k f(iy)$. The idea is to insert this functional equation into the above Mellin transform and apply a change of variable, which will yield a functional equation directly analogous to how we obtained the transformation law for ζ during the first lecture. By our above computation, we have

$$\frac{\Gamma(s)}{(2\pi)^s}\sum_{n\geq 1}\frac{\tilde{\lambda}_n}{n^{s-\frac{k-1}{2}}} = \int_0^\infty f(iy)y^s\frac{dy}{y} = \int_0^\infty f(i/y)(iy)^{-k}y^{s-1}dy.$$

Applying the variable transformation $x = 1/y$, so $dx = -y^{-2}dy = -x^2 dy$, we have

$$\begin{aligned}\int_0^\infty f(i/y)(iy)^{-k}y^{s-1}dy &= \int_0^\infty f(ix)(i/x)^{-k}(1/x)^{s-1}\frac{dx}{-x^2} \\ &= i^k\int_0^\infty f(ix)x^{k-s-1}dx \\ &= i^k\frac{\Gamma(k-s)}{(2\pi)^{k-s}}\sum_{n\geq 1}\frac{\tilde{\lambda}_n}{n^{(k-s)-\frac{k-1}{2}}}.\end{aligned}$$

These are related to the completed L-function via the identities

$$\frac{\Lambda(f, s-\frac{k-1}{2})}{(2\pi)^{\frac{k-1}{2}}} = \frac{\Gamma(s)}{(2\pi)^s}\sum_{n\geq 1}\frac{\tilde{\lambda}_n}{n^{s-\frac{k-1}{2}}}, \qquad i^k\frac{\Lambda(f, \frac{k+1}{2}-s)}{(2\pi)^{\frac{k-1}{2}}} = i^k\frac{\Gamma(k-s)}{(2\pi)^{k-s}}\sum_{n\geq 1}\frac{\tilde{\lambda}_n}{n^{(k-s)-\frac{k-1}{2}}}.$$

It follows that

$$\Lambda(f, s - \frac{k-1}{2}) = i^k \Lambda(f, \frac{k+1}{2} - s).$$

We make the variable transformation $s \mapsto s - (k-1)/2$, which yields $\Lambda(f,s) = \Lambda(f, 1-s)$, as needed. This completes the proof. □

Finally, we'll discuss the Euler product of this automorphic L-function. We compute

$$\begin{aligned} L(f,s) &= \sum_{n \geq 1} \frac{\tilde{\lambda}_n}{n^s} \\ &= \prod_p \left(1 + \frac{\tilde{\lambda}_p}{p^s} + \frac{\tilde{\lambda}_{p^2}}{p^{2s}} + \frac{\tilde{\lambda}_{p^3}}{p^{3s}} + \cdots\right) \\ &= \prod_p \left(1 - \frac{\tilde{\lambda}_p}{p^s} + \frac{1}{p^{2s}}\right)^{-1} \\ &= \prod_p \left(1 - \frac{\alpha_1(f_p)}{p^s}\right)^{-1} \left(1 - \frac{\alpha_2(f_p)}{p^s}\right)^{-1}. \end{aligned}$$

where we used unique factorization, and then multiplicity of $\tilde{\lambda}_n$ on relatively prime arguments, and then the recursive relation among the prime powers (note: this is one reason we consider the normalized Hecke eigenvalues) and then the definition of the Satake parameters. **This is the Euler product of the automorphic L-function corresponding to the Hecke eigenform f.** The $\alpha_1(f_p), \alpha_2(f_p)$ are the Satake parameters; there are 2 of them because f is an automorphic form on GL_2.

In general, if π is an automorphic representation on $\mathrm{GL}_m/\mathbb{Q}$, then we'll see an Euler product of the following form (with m Satake parameters; we say m is the degree of the L-function)

$$L(\pi, s) = \prod_p \prod_{j=1}^{m} \left(1 - \frac{\alpha_j(\pi_p)}{p^s}\right)^{-1}.$$

And the **Generalized Ramanujan conjecture** (GRC) says that $|\alpha_p(\pi_p)| = 1$ if π is unramified at p (note that everything is unramified at p in GL_2.) **We don't even know if this is true for Maass forms!!** The current best bound for Maass forms was proven by Kim and Sarnak (2000), that $|\alpha_k(\pi_p)| \leq p^{7/64}$.

(Lecture 7: October 1, 2020)

1.14 Digression: motivating Hecke operators

Recall the zeta function,

$$\zeta(s) := \sum_{n \geq 1} \frac{1}{n^s} = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1}.$$

It's a deep fact that the Dirichlet series is equal to the Euler product. Recall the Legendre symbol

$$\left(\frac{n}{p}\right) = \begin{cases} 1 & n \text{ is a quadratic residue} \\ 0 & p \mid n \\ -1 & n \text{ is not a quadratic residue} \end{cases}$$

We can think of this as a character $\chi_p : (\mathbb{Z}/p\mathbb{Z})^\times \to \{|z| = 1\}$. In general, if you take χ a character $(\mathbb{Z}/d\mathbb{Z})^* \to \{|z| = 1\}$, then the Dirichlet L-function associated to χ also has an Euler product:

$$L(s, \chi) := \sum_{n \geq 1} \frac{\chi(n)}{n^s} = \prod_p \left(1 - \frac{\chi(p)}{p^s}\right)^{-1}.$$

Another thing in common between ζ and $L(s, \chi)$ is the existence of a functional equation: there is $\zeta(s) \leftrightarrow \zeta(1-s)$ and $L(s, \chi) \leftrightarrow L(1-s, \overline{\chi})$.

People wanted to know if L-functions obeying these nice properties (functional equation, Euler product) can be generalized. People found:

1. Hasse-Weil zeta functions (corresponding to an algebraic variety).

2. Artin L-functions (corresponding to Galois representations)

But people wanted to know: *what is the most general L-function one can write down?* Towards this, Langlands is the one who coined the term *automorphic L-function.* These L-functions are cooked up using a representation π of a given algebraic group G. Such an L-function is written $L(s, \pi)$. These are complicated objects. **But people believe that all L-functions that have Euler products and functional equations must be one of the automorphic L-functions that Langlands defined.** The simplest automorphic L-function other than $\zeta(s), L(s, \chi)$ are the L-functions attached to modular forms: $L(s, f)$, where f is a holomorphic Hecke modular form. So that is one reason that it's natural to study Hecke operators at this point.

Let $E : y^2 = x^3 + ax + b$ be an elliptic curve. We want to count

$$a_p := p + 1 - \#\left\{(x, y) \in (\mathbb{Z}/p\mathbb{Z})^2 : y^2 \equiv x^3 + ax + b \pmod{p}\right\}.$$

Similarly, we can compute a_{p^2} to be some normalizing factor minus the count modulo p^2. So if we define

$$L(s, E) := \prod_p \left(1 - \frac{a_p/\sqrt{p}}{p^s} + \frac{1}{p^{2s}}\right)^{-1},$$

then one can show that $L(s, E)$ has a functional equation relating s and $1 - s$. So it's natural to conjecture that there exists some holomorphic Hecke eigenform f (for $\Gamma_0(N)$) such that $L(s, f) = L(s, E)$. This is true for some cases by Andrew Wiles, and this led to the solution of Fermat's last theorem. This is one of the most dramatic applications of L-functions attached to Hecke eigenforms. *So this is one reason that we study Hecke theory.*

1.15 The Petersson trace formula

As we can see from the Euler product, the $L(s, f)$ is completely determined by the Hecke eigenvalues at primes p, the

$$\tilde{\lambda}_p.$$

So these Hecke eigenvalues are interesting to study. For example, people ask questions like "if we vary the modular forms, how do these Hecke eigenvalues vary?" Or "fix the modular form, and consider the Hecke eigenvalues for varying p." These questions are related to the Sato–Tate conjecture. To discuss these questions, we need to discuss the trace formula. *Roughly speaking, the trace formula will tell us what the average of the n'th Hecke eigenvalue $\lambda_f(n)$ is, as f runs over $M_k(\Gamma)$.* It is a formula that computes a quantity of the shape

$$\sum_{f \in M_k(\Gamma)} \lambda_f(n).$$

This is a natural first statistical quantity to compute. We need to first define the Petersson inner product.

Proposition 1.108

If $f, g \in M_k(\Gamma)$, then $y^k f(z)\overline{g}(z)$ is Γ-invariant.

Proof. Note that $\Im(\gamma z)^k = y^k/|cz+d|^{2k}$, so because $f(\gamma z) = (cz+d)^k f(z)$ and $g(\gamma z) = (cz+d)^k g(z)$, then

$$\Im(\gamma z)^k f(\gamma z)\overline{g(\gamma z)} = y^k f(z) g(z),$$

as needed. □

We endow an inner product on the finite dimensional vector space $M_k(\Gamma)$.

Definition 1.109

For $f, g \in M_k(\Gamma)$, we define

$$\langle f, g \rangle = \int_{\Gamma \backslash \mathbb{H}} f(z)\overline{g(z)} y^k \frac{dxdy}{y^2}.$$

According to the third homework, $dxdy/y^2$ is Γ-invariant, so this is a well-defined measure. So by the previous propositon, the whole quantity is well-defined.

Proposition 1.110

$M_k(\Gamma)$ (resp. $S_k(\Gamma)$) equipped with the Petersson inner product is a Hilbert space.

Proof. It is a complete, separable vector space (since it's finite dimensional) equipped with a non-degenerate inner product (since $\langle f, f \rangle > 0$.) □

Now we need to introduce Poincaré series. To simplify notation, fix $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$, and

$$\Gamma_\infty = \left\{ \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix} : n \in \mathbb{Z} \right\} \subseteq \Gamma.$$

Fix $\phi : \mathbb{H} \to \mathbb{C}$ a holomorphic function such that $p(z) = p(z+1)$, so p is not necessarily modular. Consider: if f is a modular form, then

$$f(z) = f(\gamma z)(cz+d)^{-k} =: f|_\gamma(z).$$

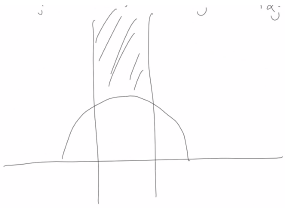
Thus, f is a modular form with respect to Γ iff $f = f|_\gamma$ for all $\gamma \in \Gamma$. And one can check that $f|_{\gamma_1}|_{\gamma_2} = f|_{\gamma_1\gamma_2}$. Because p is not a modular form, $p \neq p|\gamma$ in general, and one way of turning p into a modular form is to take its average over the slash operator. Namely, if

$$P = \sum_{\gamma\in\Gamma} p|_\gamma$$

is well-defined, then P is a modular form. Unfortunately, this sum diverges, since it contains the sub-sum $\sum_{n\in\mathbb{Z}} p|\begin{pmatrix}1 & n\\ 0 & 1\end{pmatrix}$. So we remedy this by actually defining

$$P := \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} p|_\gamma.$$

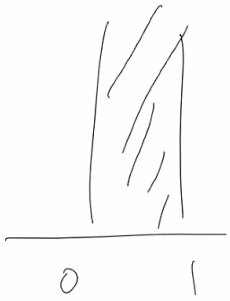
So if we define $\Gamma = \cup_j \Gamma_\infty \alpha_j$, then this means we're taking $P = \sum_j P|_{\alpha_j}$. Illustrating this quotient: we have the fundamental domain



T shifts this to the right one:

and S sends the triangle:

We know $\Gamma\mathcal{F} = \mathbb{H}$ since $\mathcal{F}$ is a fundamental domain. But $\alpha_j\mathcal{F}$ collects fundamental domains on the half plane, but only one among the shifts. Meaning, we can choose α_j so that we only have the vertical strip, since we're picking only one from the translations:



So picture the quotient $\Gamma\backslash\mathbb{H}$ as the transformations moving fundamental domains into that above rectangular strip.

The series $P = \sum_{\gamma\in\Gamma} p|_\gamma$ will only converge if p is nice enough. Concretely,

$$P(z) = \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} (cz+d)^{-k} p(\gamma z).$$

This is clearly invariant under $|_\gamma$, so this is a modular form only given absolute convergence for any particular p that we consider. The most important Poincaré series we consider is the following, taking $p(z) := e(mz)$:

Definition 1.111

The m'th Poincaré series of weight k is

$$P_m(z) = \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} (cz+d)^{-k} e(m\gamma z).$$

One can show that $P_0 = E_k$. The way to see this is by parameterizing the quotient via

$$\Gamma_\infty\backslash\Gamma = \Gamma_\infty \begin{pmatrix} * & * \\ c & d \end{pmatrix}$$

with $(c,d) = 1$. And summing over $(c,d) = 1$ gives the normalized Eisenstein series.

Proposition 1.112

If $f = \sum_{n\geq 0} a_n q^n \in M_k$, then

$$\langle f, P_m\rangle = \frac{\Gamma(k-1)}{(4\pi m)^{k-1}} a_m.$$

Proof. We compute

$$\begin{aligned}\int_{\gamma\backslash\mathbb{H}} y^k f(z) \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} \overline{(cz+d)^{-k} e(m\gamma z)} d\mu(z) &= \int_{\gamma\backslash\mathbb{H}} \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} y^k \overline{(cz+d)^{-k}} (f(\gamma z)(cz+d)^{-k}) \overline{e(m\gamma z)} d\mu(z) \\ &= \int_{\gamma\backslash\mathbb{H}} \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} \frac{y^k}{|cz+d|^{2k}} f(\gamma z) \overline{e(m\gamma z)} d\mu(z) \\ &= \int_{\gamma\backslash\mathbb{H}} \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} \Im(\gamma z) f(\gamma z) \overline{e(m\gamma z)} d\mu(z) \\ &= \int_{\Gamma_\infty\backslash\mathbb{H}} \Im(z)^k f(z) \overline{e(mz)} d\mu(z)\end{aligned}$$

This technique is called "unfolding." We can now evaluate this integral directly:

$$\begin{aligned}\int_{\Gamma_\infty\backslash\mathbb{H}} \Im(z)^k f(z) \overline{e(mz)} d\mu(z) &= \int_0^\infty \int_0^1 y^k f(z) e^{-2\pi i m z - 2\pi m y} \frac{dx\,dy}{y^2} \\ &= \int_0^\infty a_m e^{2\pi i m y - 2\pi m y} y^{k-2} dx.\end{aligned}$$

since the only term in the Fourier expansion of f which survives the integration is $n = m$. After a change of variable, this is just a gamma function, which finishes the proof. □

Corollary 1.113

$\{P_m : m \geq 0\}$ spans M_k.

Proof. The span of the P_m is a closed subspace of M_k, because M_k is finite-dimensional. If the span is a proper subspace of M_k, then there exists an orthogonal compliment of this span, so there exists $f \in M_k$ so that $\langle P_m, f\rangle = 0$ for all $m \geq 0$. This implies $a_n = 0$ for all m, so $f = 0$. □

(Lecture 8: October 6, 2020)

Theorem 1.114

We have

$$T_n P_m = \sum_{d|(m,n)} \left(\frac{n}{d}\right)^{k-1} P_{mn/d^2}$$

Proof. We first give a different realization of Hecke operators. Define

$$R(n) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_{2\times 2}(\mathbb{Z}) : ad - bc = n \right\}.$$

Then $R(1)\backslash R(n)$ is parameterized by

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

Letting $j_\gamma(z) = cz + d$, where $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, and define

$$f|_\gamma(z) = (\det \gamma)^{k/2} j_\gamma(z)^{-k} f(\gamma z),$$

then we can prove that

$$T_n f = n^{\frac{k}{2}-1} \sum_{\gamma \in R(1)\backslash R(n)} f|_\gamma.$$

This implies that

$$\begin{aligned}(T_n P_m)(z) &= T_n \left(\sum_{g \in \Gamma_\infty \backslash R(1)} j_g(z)^{-k} e(mgz) \right) \\ &= n^{k-1} \sum_{g \in \Gamma_\infty \backslash R(n)} j_g(z)^{-k} e(mgz)\end{aligned}$$

where we used $j_{\gamma'}(z) j_\gamma(\gamma' z) = j_{\gamma\gamma'}(z)$.

Now let H and B be any sets of right coset representatives of $\Gamma_\infty \backslash R(1)$, and $R(1)\backslash R(N)$. Simple but powerful observation: HG is a set of right coset representatives of $\Gamma_\infty \backslash R(n)$, and the same can be said for GH. If we use this fact, then we can write down

$$\begin{aligned}(T_n P_m)(z) &= n^{k-1} \sum_{\rho \in G} \sum_{\tau \in H} j_{\rho\tau}(z)^{-k} e(m\rho\tau z) \\ &= n^{k-1} \sum_{\rho \in G, \tau \in H} j_\tau(z)^{-k} j_\rho(\tau z)^{-k} e(m\rho\tau z) \\ &= n^{k-1} \sum_{ad=n} d^{-k} \sum_{b \pmod d} \sum_{\tau \in H} j_\tau(z^{-k}) e(m \frac{a\tau z + b}{d}) \\ &= n^{k-1} \sum_{ad=n, d|m} d^{1-k} \sum_{\tau \in H} j_\tau(z)^{-k} e(\frac{am}{d} \tau z) \\ &= \sum_{d|(m,n)} \left(\frac{n}{d}\right)^{k-1} P_{mn/d^2}(z).\end{aligned}$$

We used our parameterization above for $R(1)\backslash R(n)$ to get $j_\rho(\tau z)^{-k} = d^k$ (since ρ always has lower left coordinate zero). Then we sum over b; that vanishes unless $d \mid m$, since $e(mb/d) = 1$ if $d \mid m$. □

Hence, no Poincare series other than P_0 is the eigenfunction of a Hecke operator. However, the Hecke action on a Poincaré series can be written explicitly in terms of the Poincaré series of lower order.

Taking $m = 0$ we get the Eisenstein series so:

Corollary 1.115

$T_n E_k = \sum_{d|n} \left(\frac{n}{d}\right)^m E_k = \sigma_{k-1}(n) E_k$, which means E_k is a joint eigenfunction of all Hecke operators.

This means that **the Eisenstein series itself is a holomorphic Hecke eigenform.**

Corollary 1.116

The action $T_n \curvearrowright S_k$ is self adjoint with respect to the Petersson inner product.

Proof. It suffices to show self-adjointness on the Poincaré series, since they span. We compute

$$m^{k-1} T_n P_m = n^{k-1} T_m P_n,$$

by symmetry of the formula that we just proved. Now let $f = \sum_{n \geq 1} a_n e(nz)$ be a cusp form, and write

$$T_n f = \sum_{m \geq 1} a_m(n) e(mz).$$

Then, $a_m(n) = a_n(m) = \sum_{d|(n,m)} d^{k-1} a_{nm/d^2}$ (call this 1), which implies that

$$m^{k-1} \langle T_n f, P_m \rangle = n^{k-1} \langle T_m f, P_n \rangle,$$

which we call 2. From these symmetry formulas, we can argue that

$$\begin{aligned} \ell^{k-1} \langle T_n P_\ell, P_m \rangle &=_1 n^{k-1} \langle T_\ell P_n, P_m \rangle \\ &=_2 \left(\frac{n}{m}\right)^{k-1} \ell^{k-1} \langle T_m P_n, P_\ell \rangle \\ &=_1 \ell^{k-1} \langle T_n P_m, P_\ell \rangle. \end{aligned}$$

Thus far, we've showed that

$$\langle T_n P_\ell, P_m \rangle = \langle T_n P_m, P_\ell \rangle.$$

The corollary will follow when we show that the Fourier coefficients of Poincaré are real, because if this is the case, then we can switch the order of the inner product (which we recall conjugates the second coordinate) and we'll get that $\langle T_n P_m, P_\ell \rangle = \langle P_\ell, T_n P_m \rangle$.

We'll do this below. □

Towards the Fourier expansion of Poincaré series:

Lemma 1.117

The double coset decomposition of $\Gamma = \mathrm{SL}_2(\mathbb{Z})$:

$$\Gamma = \Gamma_\infty \cup \bigcup_{c>0} \bigcup_{d \pmod c} \Gamma_\infty \begin{pmatrix} * & * \\ c & d \end{pmatrix} \Gamma_\infty.$$

In other words, $\Gamma_\infty \backslash \Gamma / \Gamma_\infty$ can be parameterized by $\begin{pmatrix} * & * \\ c & d \end{pmatrix}$ with $c > 0, 0 < d < c, (d, c) = 1$.

Proof. The way to see this is first to notice that $\Gamma_\infty \backslash \Gamma$ is parameterized by $\begin{pmatrix} * & * \\ c & d \end{pmatrix}$ with (c, d) coprime. Then quotienting on the right by Γ_∞ gets us the rest of the way there. □

Using this, we write

$$P(z) = p(z) + \sum_{1 \neq \gamma \in \Gamma_\infty \backslash \Gamma / \Gamma_\infty} I_\gamma(z)$$

where

$$\begin{aligned} I_\gamma(z) &= \sum_{\tau \in \Gamma_\infty} j_{\gamma\tau}(z)^{-k} p(\gamma\tau z) \\ &= \sum_{n \in \mathbb{Z}} (c(z+n)+d)^{-k} p\left(\frac{a}{c} - \frac{1}{c(c(z+n)+d)}\right) \\ &= \sum_{n \in \mathbb{Z}} \int_{-\infty}^{\infty} (c(z+v)+d)^{-k} p\left(\frac{a}{c} - \frac{1}{c(c(z+v)+d)}\right) e(-nv) dv. \end{aligned}$$

by Poisson summation. Now let $p(z) = e(mz)$, and make the variable transformation $m \mapsto m - z - d/c$, so we continue

$$I_\gamma(z) = \sum_{n \in \mathbb{Z}} e(nz + \frac{ma+nd}{c}) J_c(m,n)$$

where $J_c(m,n) = \int_{-\infty+iy}^{\infty+iy} (cv)^{-k} e(-\frac{m}{cv} - nv) dv$. For $n \leq 0$, we can send $y \to \infty$ and get $e(-nv) \to 0$, which tells us that $J_c(m,n) = 0$. And for $n > 0$ and $m > 0$, this is (by definition of Bessel functions)

$$\frac{2\pi}{i^k c} \left(\frac{n}{m}\right)^{\frac{k-1}{2}} J_{k-1}\left(\frac{4\pi\sqrt{mn}}{c}\right).$$

Therefore,

Theorem 1.118

The Fourier expansion of the m'th Poincaré series of weight k is

$$P_m(z) = e(mz) + \frac{2\pi}{i^k m^{(k-1)/2}} \sum_{n \geq 1} e(nz) n^{(k-1)/2} \sum_{c>0} \frac{S(m,n,c)}{c} J_{k-1}\left(\frac{4\pi\sqrt{mn}}{c}\right)$$

where

$$S(m,n,c) := \sum_{d \pmod c} e(\frac{md^* + nd}{c}),$$

is called the *Kloosterman sum.*

It's one of the most important exponential sums one would like to understand in analytic number theory.

Corollary 1.119

$P_m(z), m \geq 1$ is cuspidal.

Corollary 1.120

$\{P_m(z) : m \geq 1\}$ spans S_k.

Corollary 1.121

The Fourier coefficients are real.

Proof. Kloosterman sums are real, and Bessel functions are real. □

This completes the proof that the Hecke operator is self-adjoint with respect to Petersson inner product. Note: this can also be proved using the definition directly.

Theorem 1.122

S_k is spanned by joint eigenfunctions of $\{T_n\}_{n\geq 1}$. These joint eigenfunctions are called *holomorphic Hecke cuspforms.*

Proof. It is a fact from linear algebra that self-adjoint commuting linear operators have a simultaneous orthonormal eigenbasis. □

Next, we'll get one of the most powerful techniques people use to study modular forms.

Let B_k be the basis of S_k consisting only of Hecke cusp forms. For $f \in B_k$, let $f(z) = \sum a_f(n)e(nz)$, where we assume that f is Petersson normalized via $\langle f, f\rangle = 1$. Then

$$P_m(z) = \sum_{f\in B_k} \langle P_m, f\rangle f(z) = \frac{\Gamma(k-1)}{(4\pi m)^{k-1}} \sum_{f\in B_k} \overline{a_f(m)} f(z)$$

(The first equality is true in general; we have an orthonormal basis of a finite dimensional vector space, so any vector in the space can be decomposed into a sum over its projections on to each basis element.) Now, we take the inner product:

$$\begin{aligned}\langle P_m, P_n\rangle &= \frac{(\Gamma(k-1))^2}{(4\pi m)^{k-1}(4\pi n)^{k-1}} \sum_{f\in B_k} a_f(n)\overline{a_f(m)} \\ &= \frac{\Gamma(k-1)}{(4\pi\sqrt{mn})^{k-1}} \left(\delta_{m,n} + \frac{2\pi}{i^k} \sum_{c>0} \frac{S(m,n,c)}{c} J_{k-1}\left(\frac{4\pi\sqrt{mn}}{c}\right)\right)\end{aligned}$$

Theorem 1.123: (Petersson trace formula, version 1)

For $m, n \geq 1$, we have

$$\frac{\Gamma(k-1)}{(4\pi\sqrt{mn})^{k-1}} \sum_{f\in B_k} a_f(n)\overline{a_f(m)} = \delta_{m,n} + \frac{2\pi}{i^k} \sum_{c>0} \frac{S(m,n,c)}{c} J_{k-1}\left(\frac{4\pi\sqrt{mn}}{c}\right)$$

The LHS is the "spectral side" and the RHS is the "arithmetic side."

The arithmetic side is referred to as such because it contains arithmetic sums. We'll now present a different version of the Petersson trace formula. Because f is a Hecke eigenform, we know $a_f(n) = a_f(1)\lambda_f(n)$, where $\lambda_f(n)$ is the n-th Hecke eigenvalue. Now, let us normalize f by setting $a_f(1) = 1$, i.e., we write

$$f(z) = \sum_{n\geq 1} \lambda_f(n) n^{\frac{k-1}{2}} e(nz).$$

I.e., f is *Hecke normalized.* (We're abusing notation here; now $\lambda_f(n)$ is denoting the normalized Hecke eigenvalue, whereas before it denoted the much larger Hecke eigenvalue.) In this case, the spectral side of the Petersson trace formula becomes

$$\frac{\Gamma(k-1)}{(4\pi\sqrt{mn})^{k-1}}\sum_{f\in B_k} a_f(n)\overline{a_f(m)} = \frac{\Gamma(k-1)}{(4\pi)^{k-1}}\sum_{f\in H_k}\frac{\lambda_f(n)\lambda_f(m)}{\langle f,f\rangle}$$

Here H_k denotes a basis of Hecke normalized cusp forms. The extra inner product in the denominator pops out because these are no longer L^2-normalized.

1.16 Rankin-Selberg convolution

Now, we're going to talk about *what will be the typical size of the Petersson norm of a Hecke normalized modular form.* In order to do this, we'll need to introduce real analytic Eisenstein series and the Rankin-Selberg convolution.

Definition 1.124

The real analytic Eisenstein series is

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

This converges absolutely with $\Re s > 1$, but we can analytically continue it to complex plane with poles potentially at $s = 0, 1$. Now, for $f, g \in H_k$ (Hecke cusp forms of weight k, which are Hecke normalized) we look at

$$\begin{aligned}
\int_{\Gamma\backslash\mathbb{H}} y^k f(z)\overline{g(z)}E(z,s)\frac{dx}{dy}y^2 &= \int_{\Gamma\backslash\mathbb{H}} y^k f(z)\overline{g(z)}\sum_{\gamma\in\Gamma_\infty\backslash\Gamma}\Im(\gamma z)^s\frac{dxdy}{y^2}\\
&= \int_{\Gamma_\infty\backslash\mathbb{H}} y^k f(z)\overline{g(z)}\Im(z)^s\frac{dxdy}{y^2}\\
&= \int_0^\infty\int_0^1 f(z)\overline{g(z)}y^{s+k}\frac{dxdy}{y^2}\\
&= \sum_{n\geq 1} n^{k-1}\lambda_f(n)\lambda_g(n)\int_0^\infty e^{-4\pi ny}y^{s+k-2}dy\\
&= (4\pi)^{-(s+k-1)}\sum_{n\geq 1}\frac{\lambda_f(n)\lambda_g(n)}{n^s}\Gamma(s+k-1)\\
&= \frac{\Gamma(s+k-1)}{(4\pi)^{s+k-1}}L(s,f\otimes g).
\end{aligned}$$

Justifying the fourth line: when we conjugate the expansion of g and integrate is against the expansion of f with respect to the variable x, the exponentials are orthogonal to 1 unless they both came from q^n, in which case the exponentials are exactly one.

So it's natural to define

Definition 1.125

The *Rankin-Selberg convolution* corresponding to f and g is

$$L(s, f \otimes g) := \sum_{n \geq 1} \frac{\lambda_f(n)\lambda_g(n)}{n^s}.$$

Collecting facts about the Rankin-Selberg convolution:

1. When $f = g$, the LHS has a simple pole at $s = 1$ (coming from the real analytic Eisenstein series) with residue $\langle f, f \rangle$. And the RHS has a simple pole at $s = 1$ with residue

$$\frac{\Gamma(k)}{(4\pi)^k \zeta(2)} L(1, \mathrm{sym}^2 f)$$

(Note: the value of this is finite.)

Thus, the spectral side of Petersson trace formula becomes

$$\frac{\Gamma(k-1)}{(4\pi)^{k-1}} \sum_{f \in H_k} \frac{\lambda_f(n)\lambda_f(m)}{\langle f, f \rangle} = \frac{4\pi\zeta(2)}{k-1} \sum_{f \in H_k} \frac{\lambda_f(n)\lambda_f(m)}{L(1, \mathrm{sym}^2, f)}.$$

So we may think of this as a weighted average of $\lambda_f(n)\lambda_f(m)$, where the weight is given by a special value of the symmetric square L-function of f. We can think of this as an average because $\#H_k \approx k/12$, and we're dividing this by $k - 1$, which is linear in k.

But what is the typical size of $L(1, \mathrm{sym}^2, f)$?

Theorem 1.126

For any $\epsilon > 0$,

$$k^{-\epsilon} \ll_\epsilon L(1, \mathrm{sym}^2, f) \ll_\epsilon k^\epsilon.$$

So this varies very moderately as the weight $k \to \infty$. The upper bound is easy, we can prove it ourselves if we try to follow the definition of the L-series. The lower bound is quite serious, it's actually JJ's favorite theorem of Hoffstein-Lockhart. They did this by studying the Siegel zeros of various L-functions. (Just like you can prove this for Dirichlet L-functions, which gives you a similar lower bound $L(1, \chi_d) \gg d^{-\epsilon}$.) The lower bound is ineffective.

Returning to Petersson: the trace formula, in this case, tells us the weighted average is 1 when $m = n$, and is something small when $m \neq n$. Actually showing the summation involving Kloosterman sums is smaller than the main contribution is difficult, but there are many techniques people use to show that type of estimate.

(Lecture 9: October 8, 2020)

1.17 Basic estimates on modular forms

We can estimate the Fourier coefficients of cusp forms as follows:

Proposition 1.127

Suppose we have a cusp form $f(z) = \sum_{n\geq 1} a(n)e(nz) \in S_k$.

1. $a(n) = O_f(n^C)$.
2. $F(z) := y^{k/2}|f(z)|$ is a bounded function in $z \in \mathbb{H}$.
3. $\sum_{n=1}^{N} |a_n|^2 \ll_f e^{2\pi N y} y^{-k}$.
4. $a(n) \ll n^{k/2}$.

Proof. For the first point, since $M = \mathbb{C}[G_4, G_6]$, and G_4 and G_6 have polynomially growing coefficients, any polynomial in G_4 and G_6 only grows polynomially fast.

For the second, we know $y^k|f(z)|^2$ is Γ-invariant, as

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

As $F(z) \geq 0$, this implies $F(z)$ is Γ-invariant as well. Notice that $e(nz) = e(nx)e^{-2\pi ny}$, so as $y \to \infty$, $|f(z)| \to 0$ by considering the Fourier expansion. So $f(z)$ is bounded in the main fundamental domain. Then by Γ-invariance, $F(z)$ is bounded everywhere.

Now if we let $c := ||F||_{L^\infty(F)}$ be the supremum in one fundamental domain, then for any $y > 0$ we have

$$\int_0^1 |f(z)|^2 dx = \int_0^1 (\sum_{n\geq 1} a(n)e(n(x+iy)))(\sum_{n\geq 1} \overline{a(n)e(n(x+iy))}) = \sum_{n\geq 1} |a(n)|^2 e^{-4\pi ny}.$$

Because $y^{k/2}|f| \leq c$, we see that the above integral is bounded by $c^2 y^{-k}$, and therefore

$$\sum_{n=1}^{N} |a_n|^2 \ll_f e^{2\pi N y} y^{-k}.$$

For the fourth, setting $y = 1/N$ in the above estimate, this implies

$$\sum_{n=1}^{N} |a(n)|^2 \ll_f N^k.$$

Hence $|a(n)|^2 \ll_f n^k$, so $a(n) = O_f(n^{k/2})$. Recall the Ramanujan conjecture, which predicts that

$$a(n) = O_f(n^{(k-1)/2}).$$

for Hecke modular forms on the full modular surface, Deligne proved that we have the strongest possible bound, the Ramanujan bound. But for very general modular forms (e.g. half integral weight) we don't yet have this Ramanujan bound. □

Next homework will be a single problem, and he'll give us two weeks to prove it (and he's giving us all the necessary ingredients to prove it.) It's to prove this:

Theorem 1.128

If f is a Hecke cusp form on $\mathrm{SL}_2(\mathbb{Z})$ which is Petersson normalized, and $F(z) = y^{k/2} f(z)$, then

$$k^{\frac{1}{4}-\epsilon} \ll_\epsilon \sup_{z \in \mathbb{H}} |F(z)| \ll_\epsilon k^{\frac{1}{4}+\epsilon}.$$

This tells us that the supremum of a weighted modular form is approximately $k^{1/4}$. This is actually the main theorem of a paper which was published about 10 years ago. But this is a very good exercise to learn the basic techniques of analytic number theory.

2 Equidistribution in number theory

2.1 Diophantine approximation

Eventually we'll talk about equidistribution theorems related to modular forms. Before that, we'll get an introduction to equidistribution in general. The model question is the following:

Fix $\alpha \in \mathbb{R}$. How are the fractional parts $\{\{\alpha n\} : n \in \mathbb{N}\}$ distributed?

For example:

1. if $\alpha = 2/7$, then $\{\{\alpha n\} : n \in \mathbb{N}\} = \{i/7 : i = 0, \dots, 6\}$. So this will give

$$\frac{2}{7} \to \frac{4}{7} \to \frac{6}{7} \to \frac{1}{7} \to \frac{3}{7} \to \frac{5}{7} \to \frac{0}{7} \to \cdots.$$

   So as you increase n, you hit each element in the set quite "evenly".

2. If $\alpha = \sqrt{2}$, then

$$\{\alpha\} = 0.4142\dots$$
$$\{2\alpha\} = 0.8284\dots$$
$$\{3\alpha\} = 0.2426\dots$$
$$\{4\alpha\} = 0.6568\dots$$

   and these will spread out densely.

Note: density does NOT equal equidistribution. In this example, equidistribution means that the proportion of time the sequence spends in each interval (a, b) is $\int_a^b d\mu = b - a$.

The questions we can ask about distribution of numbers in this context:

1. Is $\{n\alpha\}$ dense?

2. Is $\{n\alpha\}$ uniformly distributed (i.e., equidistributed with respect to the standard measure)?

The answers in this case are yes (for both). The first person to do this was Kronecker.

Theorem 2.1: (Kronecker)

Let $\alpha \in \mathbb{R} - \mathbb{Q}$. Then $\{\{n\alpha\}\}$ is dense in $[0, 1)$.

Theorem 2.2: (Dirichlet)

If $\alpha \in \mathbb{R}$ and $N \in \mathbb{N}$, then there exist $p, q \in \mathbb{Z}$ with $0 < q \leq N$ such that

$$|q\alpha - p| < \frac{1}{N}.$$

Proof. Set $\alpha_0 = 0, \alpha_1 = \{\alpha\}, \alpha_i := \{2\alpha\}, \ldots, \alpha_N := \{N\alpha\}$. By the pigeonhole principle, there exist i, j such that $|\alpha_i - \alpha_j| < 1/N$. This implies

$$|i\alpha - \lfloor i\alpha \rfloor - (j\alpha - \lfloor j\alpha \rfloor)| < \frac{1}{N}.$$

Therefore, set $q = i - j$ and $p = \lfloor i\alpha \rfloor - \lfloor j\alpha \rfloor$. □

Corollary 2.3

For $\alpha \in \mathbb{R} \setminus \mathbb{Q}$, there exist infinitely many p, q such that

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^2}.$$

Proof. Fix any $N_1 \in \mathbb{N}$. By Dirichlet's theorem, there exist $p_1, q_1 \in \mathbb{Z}$ with $0 < q_1 \leq N_1$ so that

$$\left|r - \frac{p_1}{q_1}\right| < \frac{1}{Nq_1} < \frac{1}{q_1^2}.$$

Inductively, we have $\left|r - \frac{p_n}{q_n}\right| \neq 0$ because r is irrational. So we may choose N_{n+1} large enough so that

$$\left|r - \frac{p_1}{q_1}\right|, \ldots, \left|r - \frac{p_n}{q_n}\right| > \frac{1}{N_{n+1}}.$$

By Dirichlet's theorem, there exist $p_{n+1}, q_{n+1} \in \mathbb{Z}$ with $0 < q_{n+1} \leq N_{n+1}$ such that

$$\left|r - \frac{p_{n+1}}{q_{n+1}}\right| < \frac{1}{q_{n+1}N_{n+1}} < \frac{1}{q_{n+1}^2}.$$

In particular, $\left|r - \frac{p_{n+1}}{q_{n+1}}\right| < \frac{1}{N_{n+1}}$. As p_{n+1}/q_{n+1} is closer to r than any of the lower order p_i/q_i, it follows that $p_{n+1}/q_{n+1} \notin \{p_i/q_i : i = 1, \ldots, n\}$, as needed. □

This is the strongest theorem of this type:

Theorem 2.4: (Hurziwtz)

If $\alpha \in \mathbb{R} \setminus \mathbb{Q}$, then there exist infinitely many p, q such that

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{\sqrt{5}q^2}.$$

Why is that theorem the strongest of its type?

Proposition 2.5

If $r = (1-\sqrt{5})/2$ and $A > \sqrt{5}$, then

$$\left|r - \frac{p}{q}\right| < \frac{1}{Aq^2} \tag{2.1}$$

has only finitely many solutions p/q.

Proof. Suppose there exists $A > \sqrt{5}$ such that (2.1) holds for infinitely many p/q. Thus, there exist infinitely many p, q such that

$$r = \frac{p}{q} + \frac{\delta}{q^2},$$

where $\delta = \delta(p,q)$ satisfies $|\delta| < 1/\sqrt{5}$. But

$$\begin{aligned} r = \frac{p}{q} + \frac{\delta}{q^2} &\iff \frac{\delta}{q} = qr - p \\ &\iff \frac{\delta}{q} + \frac{q\sqrt{5}}{2} = \frac{q}{2} - p \\ &\implies \frac{\delta^2}{q^2} + \delta\sqrt{5} = p^2 - pq - q^2. \end{aligned}$$

As $|\delta| < \sqrt{5}$, for q sufficiently large, the LHS is strictly less than 1. In particular, since there exist infinitely many p/q satisfying (2.1), the RHS necessarily vanishes for some p, q. Thus, for this choice of p, q, we have

$$4p^2 - 4pq - 4q^2 = 0 \implies 4p^2 - 4pq + q^2 = 5q^2 \implies (2p-q)^2 = 5q^2.$$

This is impossible, as $v_5((2p-q)^2) \in 2\mathbb{Z}$, whereas $v_5(5q^2) \in 2\mathbb{Z} + 1$. □

Theorem 2.6: (Liouville)

Suppose $\alpha \in \mathbb{R}$ is algebraic of degree $n > 1$. Then there exists some constant $A > 0$ such that, for all $p, q > 0$, we have

$$\left|\alpha - \frac{p}{q}\right| \geq \frac{A}{q^n}.$$

Proof. Let $f \in \mathbb{Z}[x]$ be the minimal polynomial of α, so f is irreducible, so for all $p/q \in \mathbb{Q}$, we have $q^n f(p/q) \in \mathbb{Z} \setminus \{0\}$ (if this were zero, it would be reducible.) So by the mean value theorem, there exists

$x_0 \in [\alpha, p/q)$ such that

$$f(p/q) - f(\alpha)p/q - \alpha - f'(x_0).$$

Hence,

$$\left|\frac{p}{q} - \alpha\right| - \frac{|q^n f(p/q)}{f'(x_0)q^n} \geq \frac{1}{q^n} \frac{1}{\sup_{|x-\alpha|<\delta} |f'|}.$$

If we take this constant to be A, then we get the result. □

In particular, if α is algebraic of degree 2, then

$$\left|\alpha - \frac{p}{q}\right| \geq \frac{A}{q^2}$$

which means *algebraic numbers are the worst in terms of diophantine approximation.*

Historical remarks:

1. (Thue) the exponent n in Liouville's theorem can be replaced by $\frac{n}{2} + 1 + \epsilon$
2. (Siegel) the exponent n in Liouville's theorem can be replaced by $2\sqrt{n}$
3. (Dyson) the exponent n in Liouville's theorem can be replaced by $\sqrt{2n}$
4. (Roth) the exponent n in Liouville's theorem can be replaced by $2 + \epsilon$. This tell us that we can find some constant $A := A(\alpha)$ such that $|\alpha - p/q| \geq A/q^{2+\epsilon}$.

We'll use this to give a proof of Kronecker's theorem.

Proof of Kronecker's theorem. We want to show that, for any $x \in [0, 1)$, there exists α_{n_j} such that $\alpha_{n_j} \to x$. Given $\epsilon > 0$, let $q > 0$ be chosen so that

$$\left|\alpha - \frac{p}{q}\right| < \frac{1}{q^2}$$

with $1/q < \epsilon$. Take an integer j so that

$$|j(\alpha q - p) - x| < \frac{1}{q},$$

by Dirichlet's theorem. Then $|(jq\alpha - jp) - x| < \epsilon$. If we choose ϵ to be small enough so that an ϵ-neighborhood of x is contained in $(0, 1)$, then $jq\alpha - jp = \{jq\alpha\} = \alpha_{jq}$ suffices. □

2.2 Uniform distribution

Definition 2.7

A sequence $\{a_n\} \subseteq [0, 1)$ is *uniformly distributed* if, for every $(b, c) \subseteq [0, 1)$,

$$\lim_{N\to\infty} \frac{\#\{n \leq N : a_n \in (b, c)\}}{N} = c - b.$$

Let's play a game!! True and false questions:

1. For irrational α, $\{n\alpha\}$ is uniformly distributed: True

2. For irrational α, $\{n^2\alpha\}$ is uniformly distributed: True

3. $\{\log n\}$ is uniformly distributed: False! Once can actually show this accumulates near zero sometimes... or more precisely, it fluctuates a lot.

4. $\{n!e\}$ is not uniformly distributed. In fact, its not even dense!

   Proof. $e = \sum_{n\geq 1} \frac{1}{n!}$, so

   $$n!e \in \mathbb{Z} + \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \cdots + < \frac{1}{n+1}\left(1 + \frac{1}{1!} + \cdots +\right) = \frac{e}{n+1}.$$

   □

5. What about $\{\log p_n\}$? Along the same lines as $\{\log n\}$, the answer is no.

6. $\{\sqrt{n}\}$ is uniformly distributed.

7. $\{\log n!\}$ is unformly distributed, but $\{\log\log n!\}$ is not uniformly distributed.

Next time, we'll discuss Weyl's criterion, which lets us check whether a sequence is uniformly distributed by associating to an exponential sum. **Later, we'll discuss equidistribution of Hecke eigenvalues, and (effective) vertical Sato–Tate.** The point of this mini-unit is to warm us up for that; this will bridge randomness and number theory. For example, consider the Möbius function $\mu(n)$. We believe that $\mu(n)$ is random. Consider a random sequence x_n in $\{\pm 1\}$. Then

$$\lim_{N\to\infty} \frac{1}{\sqrt{N}} \sum_{n=1}^{N} x_n$$

is in fact the normal distribution!! So the estimate $\sum_{n=1}^{N} x_n \ll N^{1/2+\epsilon}$ happens most of the time. So formally speaking,

$$\limsup_{N\to\infty} \text{Prob}\left(\sum_{n=1}^{N} x_n \ll N^{1/2+\epsilon}\right) = 1.$$

If $\mu(n)$ were truly random, then we must also have

$$\sum_{n=1}^{N} \mu(n) \ll N^{1/2+\epsilon}$$

for large N. **Merten's conjecture** is that this estimate holds; this is in fact equivalent to RH. *So this is one example of why understanding randomness in number theory is important; being able to prove true randomness in some objects gives us hard and fast true things.*

(Lecture 10: October 13, 2020)

Now, a brute-force proof of uniform distribution of the fractional parts:

Theorem 2.8

For irrational α, $\{n\alpha\}$ is uniformly distributed.

Proof. Define $||\cdot||$ to be the distance to the nearest integer. Given $\epsilon > 0$, pick $M > 0$ sufficiently large so that $1/M < \epsilon$, and choose $1 \leq m \leq M$ such that $\delta := ||m\alpha|| < 1/M$. The existence of such an m is guaranteed by a lemma that we proved last time. We want to show that

$$\{\{\alpha n\} : 1 \leq n \leq N, n \equiv i \pmod{m}\}$$

is well-distributed, in a sense to be made explicit. Why that particular set? As $m\alpha$ is ϵ-close to an integer, we have

$$|\{m\alpha + x\} - \{x\}| < \frac{1}{M}.$$

So if we consider the sequence $\{x\}, \{x + m\alpha\}, \{x + 2m\alpha\}, \ldots$, any element of this sequence will be at most $1/M$ far away from its neighbors.

To formalize all this, let us reexpress this sequence as

$$\begin{aligned}\{\{\alpha n\} : 1 \leq n \leq N, n \equiv i \pmod{m}\} &= \{\{j(m\alpha) + i\alpha\} : 1 \leq j \leq J_i\} \\ &= \{\{\delta j + \gamma\} : 1 \leq j \leq J_i\}\end{aligned}$$

where $J_i = \lfloor N/m \rfloor = \frac{N}{m} + O(1)$, and where

$$\gamma := \begin{cases} i\alpha & \delta = \{m\alpha\} \\ i\alpha - \delta(J_i + 1) & 1 - \delta = \{m\alpha\}. \end{cases}$$

Now, for $0 \leq \gamma \leq 1$, and $K = [\delta J_i + \gamma]$,

$$\begin{aligned}\#\{j \leq J_i : \{\delta j + \gamma\} \in [b, c)]\} &= \sum_{k=0}^{K} \#\{j \leq J_i : \delta_j + \gamma \in [k + b, k + c)\} \\ &= (K + O(1)) \left(\frac{c - b}{\delta} + O(1)\right) \\ &= (c - b)J_i + O\left(\frac{c - b}{\delta} + \delta J_i + 1\right)\end{aligned}$$

by the definition of K. Next we combine all of these for $i = 1, \ldots, m - 1$. This implies that

$$\begin{aligned}\#\{n \leq N : \{\alpha n\} \in (b, c)\} &= (c - b)J_i m + O\left(\frac{(c - b)m}{\delta} + \delta J_i m + m\right) \\ &= (c - b)N + O\left(\frac{m}{\delta} + \delta N\right).\end{aligned}$$

Next, we choose $N \approx m/\delta^2$ (so that these error terms balance each other.) Then we get

$$(c - b)N + O\left(\frac{m}{\delta} + \delta N\right) = ((c - b) + O(\epsilon))\, N.$$

This implies that

$$\lim_{N\to\infty} \frac{\#\{n \le N : \{\alpha n\} \in (b,c)\}}{N} = c - b.$$

□

That was a brute force of the theorem. *This type of proof is nice when we don't have a nice technique to handle the distribution of sequences: just break it up into subsequences where each subsequence is manageable (in this case, "arithmetic progressions") and then combine them to prove the theorem.* This type of thing is quite standard in the theory of character sums and the circle method. But we'll soon present a much simpler proof of this theorem.

Before that, another thing about uniform distribution.

Proposition 2.9

1. Any sequence has a rearrangement that is not uniformly distributed.
2. Any dense sequence has a rearrangement that is uniformly distributed.

Proof. (1) is trivial: if it's not already uniformly distributed, divide it into two halves, say those in $[0, 1/2]$ are the subsequence b_i and those in $(1/2, 1]$ are c_i. Then $b_1, b_2, c_1, b_3, b_4, c_2 \ldots$ is clearly not uniformly distributed since it spends more time in the left half of the interval.

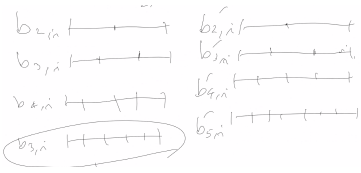
For (2), let c_n be the arbitrary sequence. if we define $a_{\binom{m}{2+1}} = b_{m,i} = i/m$, then this is uniformly distributed. This marches along in 4 steps, then in 8, then in 16, etc. Now set $m_{m,i} = c_n$, such that

$$\frac{i-1}{m} \le c_n \le \frac{i}{m}.$$

Then

$$d_{\binom{m}{2}} = b_{m_i}.$$

Basically, what we're trying to do is mimic the sequence i/m.



But since c_n can be arbitrary, the new $b'_{m,i}$ mimics the $b_{m,i}$. □

We mention that just to emphasize the role of ordering when talking about the ordering of numbers. Now, a very important theorem in the world of equidistribution:

Theorem 2.10: (Weyy's criterion)

The following are equivalent:

1. $\{a_n\}$ is uniformly distributed modulo 1.
2. For all continuous f, $\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} f(\{a_n\}) = \int_0^1 f dx$.
3. $\lim \frac{1}{N}\sum_{n=1}^{N} e(ma_n) = \delta_{m,0}$

Proof. (1 $\implies$ 2): 1 means 2 is true for step functions.

(2 $\implies$ 3): this is obvious (if it's true for every function, in particular it must be for exponential functons)

(3 $\implies$ 2): you can approximate every sufficiently continuous function using Fourier expansion.

(2 $\implies$ 1) : you can approximate a step function by continuous functions. □

Example for (2):

$$\sum_{n=0}^{N} e(m\alpha n) = \frac{e((N+1)m\alpha) - 1}{e(m\alpha) - 1}$$

is valid as long as $\alpha \notin \mathbb{Q}$ and $m \neq 0$. As we increase N, the absolute value of this expression is uniformly bounded. Hence

$$\frac{1}{N}\sum_{n=0}^{N} e(m\alpha n) \to 0$$

if $\alpha \notin \mathbb{Q}$ and $m \neq 0$.

This is why people care about exponential sums; if you can exhibit cancellations in exponential sums, then you can say things about the distribution of numbers. For example, to prove the existence of arbitrarily long arithmetic progressions in the primes, Green-Tao showed that nilpotent flow is disjoint from the Mobius function, which is very related to some cancellation in exponential sums.

2.3 A more general notion of equidistribution

Definition 2.11

Let (X, μ) be a measure space. We say $\{x_n\} \subseteq X$ is *equidistributed* with respect to μ if any one of the following holds:

1. $\frac{1}{N}\sum_{n=1}^{N} \delta_{x_n} \to \mu$ in the sense of weak convergence.
2. $\frac{1}{N}\sum_{n=1}^{N} f(x_n) \to \int_X f d\mu$ for all $f \in C(X)$.
3. For any open ball A, $\frac{1}{N}\#\{x_n \in A : n < N\} \to \mu(A)/\mu(X)$.

The equivalence of these properties can be proven using techniques similar to Weyl's criterion. Just like Weyl's criterion, *it is sufficient to check condition (2) for a basis* $\{f_n\}$ for the set of continuous functions.

Why do we mention this? The exponential function might not necessarily be the best choice of basis. For example, if the points are equidistributed with respect to the non-uniform measure, then there might be a better basis to work with. (Like the Chebyshev polynomials in the Sato–Tate distribution.) An important trick towards this that he wishes were taught earlier: **orthogonal polynomials.**

Let μ be a measure on $\mathbb{R}$ such that $\int f(x)d\mu(x) < \infty$ for all polynomials f. Then

$$\langle f, g\rangle := \int fgd\mu$$

defines an inner product (as it's finite for any choice of f and g by hypothesis.)

Definition 2.12

The sequence $\{p_n\}_{n\geq 0}$ of orthogonal polynomials is defined by

$$\deg p_n = n, \qquad \langle P_m, P_n\rangle = 0 \text{ when } m \neq n.$$

These are the *orthogonal polynomials corresponding to* μ.

P_0 will necessarily be a constant, then inductively, $P_0, \ldots, P_{n-1}$ determines P_n by the inner product. (All the above orthogonal polynomials are defined this way!!)

Example:

1. The *Jacobi polynomials* are the orthogonal polynomials deterimed by the measure $(1-x)^\alpha(1+x)^\beta\chi_{[-1,1]}(x)dx$, and these polynomials are denoted $P_n^{(\alpha,\beta)}(x)$.

2. The *Hermite polynomials* are the orthogonal polynomials deterimed by the measure $e^{-x^2}dx$, and they're denoted by $H_n(x)$.

3. The *Gegenbauer polynomials* are the orthogonal polynomials deterimed by the measure $(-x^2)^\alpha\chi_{[-1,1]}(x)$. In particular, these are a special case of the Jacobi polynomials. They're denoted $C_n^{(\alpha)}(x)$.

4. The *Legendre polynomials* are the orthogonal polynomials deterimed by the measure $\chi_{[-1,1]}(x)dx$. These are denoted $P_n(x)$.

5. *Chebyshev polynomials of the first kind* are denoted $T_n(x)$. They correspond to the measure $dx/\sqrt{1-x^2}$.

6. *Chebyshev polynomials of the second kind* and denoted $U_n(x)$. They correspond to the measure $\sqrt{1-x^2}dx$.

   These appear naturally when you're trying to understand the Laplacian on the sphere; they're referred to as *spherical harmonics* in general.

   Note that

   $$U_n(\cos\theta) = \frac{\sin((n+1)\theta)}{\sin\theta},$$

   so if we write $\lambda_f(p) = 2\cos\theta_p$, then $\lambda_f(p^k) = U_k(\lambda_f(p)/2)$. In particular, the recurrence relation of Hecke eigenvalues can be expressed in a very simple way in terms of these Chebyshev polynomials.

The last four classes of orthogonal polynomials all come about when you try to solve the eigenfunction equation on a sphere. They're the eigenfunctions in some coordinates.

Let us end the lecture by briefly discussing **special functions**. It's good to know the asymptotic behavior and expansions of all of these that come up. What do we do when we encounter special functions in our research? There are two good options:

1. Look it up on the NIST functions site; this is an excellent database.
2. There is a book called *Table of integrals, series, and products* by Gradshteyn–Ryzhik. "It's the bible" says Jeff and JJ.

(Lecture 11: October 15, 2020)

2.4 Weighted vertical Sato–Tate

Today we'll discuss a Weighted distribution of Hecke eigenvalues as $k \to \infty$. Recall that by the Ramanujan conjecture, the normalized Hecke eigenvalues correponding to a holomorphic Hecke eigenform satisfy the bound

$$\lambda_f(p) \in [-2, 2].$$

We want to see how these numbers are distributed within this interval, as we vary k. And since we just discussed Petersson, it's natural (and easier) to discuss the weighted average of these numbers.

Recall one version of Petersson, where we weight by the symmetric square L-function: if B_k is a basis of S_k consisting of Hecke normalized Hecke cusp forms, then

$$\frac{4\pi\zeta(2)}{k-1} \sum_{f \in B_k} \frac{\lambda_f(n)\lambda_f(m)}{L(1, \mathrm{sym}^2, f)} = \delta_{m,n} + \frac{2\pi}{i^k} \sum_{c>0} \frac{S(m,n,c)}{c} J_{k-1}\left(\frac{4\pi\sqrt{mn}}{c}\right).$$

It is natural to define the following measure:

$$\nu_{p,k} := \frac{4\pi\zeta(2)}{k-1} \sum_{f \in B_k} \frac{1}{L(1, \mathrm{sym}^2, f)} \delta_{\lambda_f(p)}.$$

This is a weighted average of the Dirac delta masses supported on each Hecke eigenvalue $\lambda_f(p)$, where the weight comes from $L(1, \mathrm{sym}^2, f)$. But we know this is a very mild weight since $k^{-\epsilon} \ll L(1, \mathrm{sym}^2, f) \ll k^{\epsilon}$. Then we have

$$\int_{-2}^{2} U_n(x/2) d\nu_{p,k} = \frac{4\pi\zeta(2)}{k-1} \sum_{f \in B_k} \frac{\lambda_f(p^n)}{L(1, \mathrm{sym}^2, f)}$$

because we can compute this integral on the Dirac masses to be

$$\int_{-2}^{2} U_n(x/2) d\delta_{\lambda_f(p)} = U_n(\lambda_f(p)/2) = \lambda_f(p^n).$$

The Petersson trace formula with $n := p^n$ and $m := 1$ implies that

$$\frac{4\pi\zeta(2)}{k-1} \sum_{f \in B_k} \frac{\lambda_f(p^n)}{L(1, \mathrm{sym}^2, f)} = \delta_{n,0} + \frac{2\pi}{i^k} \sum_{c>0} \frac{S(p^n, 1, c)}{c} J_{k-1}\left(\frac{4\pi p^{n/2}}{c}\right) \tag{2.2}$$

(Note that $\lambda_f(1) = 1$, as $a_n = \lambda_f(n)a_1$ for all n.) Denote by E_k the series on the RHS. We'll argue that it is an error term, i.e., $E_k \to 0$ as $k \to \infty$.

To do this, we'll need to introduce a bound for the Kloosterman sum

$$S(n,m,c) = \sideset{}{^*}\sum_{x \pmod c} \left(\frac{nx + mx^*}{c}\right).$$

This is trivially bounded by $|S(n,m,c)| \leq \phi(c)$, as each summand has size at most one. But this is too crude of an estimate for our applications; as $\phi(c) = c\prod_{p|c}(1-p^{-1})$, If c has few prime factors, this is bounded away from zero; in general, the product is $\gg 1/\log c$, so in fact $\phi(c) \gg c/\log c$, which is essentially linear growth. So, we want to exploit the cancellation that happens inside this exponential sum. The best bound is obtained by Weil:

Theorem 2.13: (Weil's bound)

We have

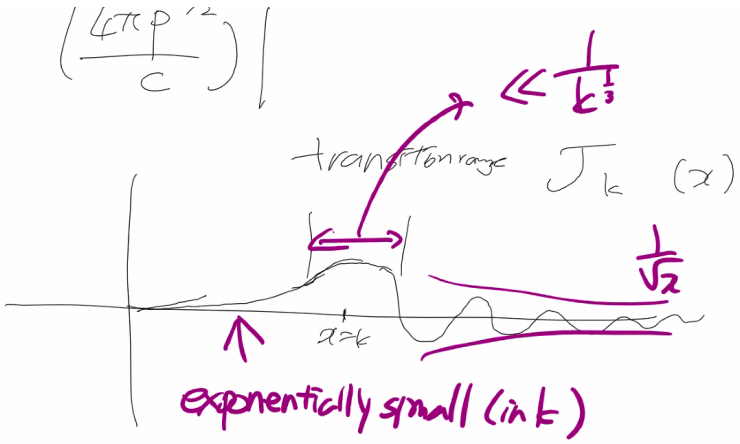
$$|S(m,n,c)| \leq \sigma_0(c)\sqrt{\gcd(m,n,c)}\sqrt{c}.$$

This implies that $|S(p^n,1,c)| \leq \sigma(c)\sqrt{c}$. *Note: oftentimes, Weil's bound is enough to prove the theorem one is considering, but many times you need to do your own better error in the particular situation that you're considering. But Weil's bound is sharp in this general setting.*

Continuing from above, we can use Weil's bound to estimate

$$|E_k| \leq 2\pi \sum_{c>0} \frac{\sigma(c)}{\sqrt{c}} J_{k-1}\left(\frac{4\pi p^{n/2}}{c}\right).$$

Next, we need some bound on the Bessel functions. For large k, J_k is exponentially small for $x < k$; then it enters the transition range (where it becomes large); and then it decays and oscillates.



This type of behavior is common; there is exponential growth, followed by some transition range, followed by some oscillatory behavior, where the transition range is defined by some Airy function. The reason for this is from harmonic analysis; namely, these special functions have integral representations which have

stationaly phase with singularity of order 3 (this explains why we're seeing 1/3 everywhere.) The explicit bounds that we'll use:

Proposition 2.14

1. $J_{k-1}(x) \ll 1/k^{1/3}$ for x close to $k-1$.
2. If $\nu \geq 0$ and $0 < x \leq 1$, then $1 \leq \frac{J_\nu(\nu x)}{x^\nu J_\nu(\nu)} \leq e^{\nu(1-x)}$.
3. In the transition range: $J_\nu(\nu + a\nu^{1/3}) = \frac{2^{1/3}}{\nu^{1/3}} A_i(-2^{1/3}a) + O(1/\nu)$, where A_i is the *Airy function*, for each $a \in I$, where I is a fixed compact interval.

Proof. We can find these on NIST. □

We may assume k is sufficiently large so that $4\pi p^{n/2}/c$ is less than $k-1$, which lets us apply the second part of the proposition. Namely, if $k-1 > \alpha/c$, where $\alpha = 4\pi p^{n/2}$, and so $\nu = k-1$ and $x = \alpha/c(k-1)$, then we have $J_\nu(\nu x) \leq e^{\nu(1-x)} x^\nu J_\nu(\nu)$, so

$$\begin{aligned} J_{k-1}\left(\frac{4\pi p^{n/2}}{c}\right) &\leq e^{(k-1)\left(1-\frac{\alpha/c}{k-1}\right)} \left(\frac{\alpha}{c(k-1)}\right)^{k-1} J_{k-1}(k-1) \\ &\ll e^k \left(\frac{\alpha}{c(k-1)}\right)^{k-1} \\ &\ll \frac{e^{-k}}{c^{k-1}} \end{aligned}$$

as $J_{k-1}(k-1) \ll k^{1/3}$; and as $(k-1)^{-(k-1)}$ decays much faster than e^k grows, so we can get the very crude upper bound of e^{-k} for their product. Finally, this tells us that

$$|E_k| \ll \sum_{c>0} \frac{\sigma(c)}{\sqrt{c}} \frac{e^{-k}}{c^{k-1}} \ll e^{-k}$$

since this series converges to a constant bounded by something independent of k.

In summary, we showed that

$$\frac{4\pi\zeta(2)}{k-1} \sum_{f \in B_k} \frac{\lambda_f(p^n)}{L(1, \mathrm{sym}^2, f)} = \delta_{n,0} + O(e^{-k}),$$

as $k \to \infty$. But by our computation above, this implies that

$$\int U_n(x/2) d\nu_{p,k} = \delta_{n,0} + O(e^{-k}),$$

so we deduce that

$$\lim_{k\to\infty} \int U_n(x/2) d\nu_{p,k} = \delta_{n,0}. \tag{2.3}$$

Next, recall that U_n are the orthogonal polynomials with respect to $\sqrt{1-x^2}dx$; by a change of variables, this implies that

$$\int_{-2}^{2} U_n(x/2)\sqrt{4-x^2}dx = \delta_{n,0}.$$

which implies that $d\nu_{p,k} \to \sqrt{4-x^2}dx$. In particular, we have

$$\lim_{k\to\infty} \frac{4\pi\zeta(2)}{k-1} \sum_{f\in B_k} \frac{\lambda_f(p^n)}{L(1,\mathrm{sym}^2,f)} = \int_{-2}^{2} U_n(x/2)\sqrt{4-x^2}dx.$$

In summary, we've shown that

$$\lim_{k\to\infty} \int_{-2}^{2} U_n(x/2)d\nu_{p,k} = \int_{-2}^{2} U_n(x/2)\sqrt{4-x^2}dx.$$

In particular, every $g \in C([-2,2])$ can be written in a basis of $U_n(x/2)$, so by linearity we get

$$\lim_{k\to\infty} \int_{-2}^{2} g(x/2)d\nu_{p,k} = \int_{-2}^{2} g(x/2)\sqrt{4-x^2}dx.$$

Expanding the definition of the measure on the LHS, we get the following:

Theorem 2.15: (Weighted vertical Sato–Tate)

The following equidistribution statement holds:

$$\lim_{k\to\infty} \frac{2\pi\zeta(2)}{k-1} \sum_{f\in B_k} \frac{g(\lambda_f(p)/2)}{L(1,\mathrm{sym}^2,f)} = \int_{-2}^{2} g(x/2)\sqrt{4-x^2}dx.$$

The Sato–Tate conjecture, which is now a theorem, says that, for a fixed $f \in S_k(\mathrm{SL}_2(\mathbb{Z}))$, the normalized Hecke eigenvalues $\lambda_f(p)$ become equidistributed with respect the semicircular measure $\sqrt{4-x^2}dx$. We proved *weighted vertical Sato–Tate*, which is so-called because in that version we fixed p and varied f. It is indeed possible to order to obtain an *unweighted vertical Sato–Tate*; this is our next task. To do this, we'll first need to discuss Selberg's trace formula.

2.5 Eichler-Selberg trace formula

The Petersson trace formula didn't really compute the trace of anything. In contrast, the Eichler-Selberg trace formula explicitly computes the trace of the Hecke operator acting on M_k. Although it's impossible in general to write down each individual Hecke eigenvalues, this result gives us an explicit formula for the trace. The full proof of Zagier appears in one of the books by Serge Lang.

We want to compute

$$\operatorname{Tr} T_m = \sum_{f\in B_k} \lambda_f(m),$$

where B_k is a basis of S_k consisting of Hecke cusp forms. As $T_m : S_k \to S_k$ is a linear operator, for some $H(z,z')$ which is bi-invariant under $\mathrm{SL}_2(\mathbb{Z})$, we can define a linear operator

$$T_H : f \mapsto \iint_{\Gamma\backslash\mathbb{H}} H(z,z')f(z)\frac{dxdy}{y^2}.$$

There ought to be some integral kernel of the Hecke operator; namely, there should be some kernel H so that $T_H = T_m$. Suppose we have found some H. Then the trace of such an operator T_H is (roughly) computed

by the formula which integrates the integral kernel over the diagonal:

$$\operatorname{Tr} T_H = \iint_{\Gamma\backslash\mathbb{H}} H(z,z)\frac{dxdy}{y^2}.$$

This is the whole idea. What Zagier did was conjure a specific H that gives T_n. The final result:

Theorem 2.16: (Eichler-Selberg trace formula)

If $k \geq 4$ and $m \geq 1$, then the trace of the Hecke oparator T_m on S_k is given by

$$\operatorname{Tr} T_m = -\frac{1}{2}\sum_{t\in\mathbb{Z}} p_k(t,m)H(4m-t^2) - \frac{1}{2}\sum_{dd'=m}\min(d,d')^{k-1}.$$

Some points:

1. In this trace formula, T_m is the un-normalized Hecke operator.
2. Here, H is the Hurwitz class number, so $H(n) = 0$ if $n < 0$, $H(0) = -1/12$, and for $n > 0$, $H(n)$ is the number of $\mathrm{SL}_2(\mathbb{Z})$-equivalence classes of positive-definite binary quadratic forms $ax^2 + bxy + cy^2$ with discriminant $b^2 - 4ac = -n$. This class number comes with some weight: we count forms equivalent to a multiple of
$$\begin{cases} x^2 + y^2 & \text{with multiplicity } 1/2 \\ x^2 + xy + y^2 & \text{with multiplicity } 1/3. \end{cases}$$
3. The above $p_k(t,N)$ are defined by
$$\frac{1}{(1-tx+Nx^2)} = \sum_{k\geq 1} x^{k-2}p_k(t,N), \qquad \text{where} \qquad p_k(t,N) := \frac{\rho^{k-1}-\overline{\rho}^{k-1}}{\rho-\overline{\rho}};$$
here, $\rho + \overline{\rho} = t$ and $\rho\overline{\rho} = N$. This expression is valid only when $t^2 < 4N$, which is acceptable, as the Hurwitz class number vanishes on negative arguments. In particular, notice that $H(4m-t^2)$ is zero for all but finitely many t, so the sum over $t \in \mathbb{Z}$ is finitely supported.

(Lecture 12: October 20, 2020)

In the Eichler-Selberg trace formula, the main term corresponds to the $4m - t^2 = 0$ summand. When $t^2 = 4m$, writing $\rho + \overline{\rho} = t$ and $\rho\overline{\rho} = t^2/4$, then we have $\rho = \overline{\rho} = t/2$. In this particular case, we compute

$$P_k(t,N) = \rho^{k-1} + \rho^{k-2}\overline{\rho} + \cdots + \overline{\rho}^{k-2} = (k-1)(t/2)^{k-2} = (k-1)m^{\frac{k-2}{2}},$$

so we can rewrite the trace formula as

$$\operatorname{Tr} T_m = \frac{k-1}{12}m^{\frac{k-2}{2}}\delta_{\sqrt{m}} - \frac{1}{2}\sum_{t^2<4m} P_k(t,m)H(4m-t^2) - \frac{1}{2}\sum_{dd'=m}\min(d'd')^{k-1},$$

where we define

$$\delta_{\sqrt{m}} = \begin{cases} 1 & \sqrt{m} \in \mathbb{Z} \\ 0 & \text{else.} \end{cases}$$

If we fix m and send $k \to \infty$, then we can estimate:

1. When m is fixed, we have the class number bound $H(4m-t^2) \ll_m 1$, as this doesn't depend on k at all.

2. As $\rho\overline{\rho} = m$ and $\rho + \overline{\rho} = t$, we estimate

$$\left|\frac{\rho^{k-1} - \overline{\rho}^{k-1}}{\rho - \overline{\rho}}\right| = \left|\frac{\rho^{k-1} - \overline{\rho}^{k-1}}{\sqrt{4m-t^2}}\right| \leq 2|\rho|^{k-1} = 2m^{\frac{k-1}{2}} = O_m(m^{\frac{k-1}{2}}).$$

3. As $\min(d, d') \leq \sqrt{m}$, we have

$$\sum_{dd'=m} \min(d,d')^{k-1} \leq \sigma_0(m) m^{\frac{k-1}{2}} = O_m(m^{\frac{k-1}{2}})$$

So if we divide by $m^{\frac{k-1}{2}}$, then we get that the trace of the normalized Hecke operator is

$$\mathrm{Tr}(\tilde{T}_m) = \frac{k-1}{12}\frac{1}{\sqrt{m}}\delta_{\sqrt{m}} + O_m(1), \qquad \text{as } k \to \infty.$$

As the space of modular forms of weight k has dimension about $k/12$, we can conclude:

Theorem 2.17

As $k \to \infty$, the average normalized Hecke eigenvalues of $\tilde{T}_m$ acting on S_k is

$$\frac{1}{|B_k|}\sum_{f \in B_k} \lambda_f(m) = \frac{1}{\sqrt{m}}\delta_{\sqrt{m}} + O_m(1/k).$$

Remark: we should expect to only get behavior like this when m is a square, because $\lambda_f(p^2) = \lambda_f(p)^2 - 1$ (this is a special case of the recurrence satisfied by Hecke eigenvalues, $\lambda(n)\lambda(m)\sum_{d|(n,m)}\lambda_f(nm/d^2)$) so on a square, these sums are positive and pile up, and otherwise they'll scatter and cancel out.

Next, consider the weighted measure

$$\mu_{p,k} = \frac{1}{|B_k|}\sum_{f \in B_k} \delta_{\lambda_f(p)}.$$

(We looked at a weighted version of this measure last time.) Then, one can show that

$$\lim_{k\to\infty}\int_{-2}^{2} U_n(x/2)d\mu_{p,k} = \begin{cases} p^{-n/2} & n \in 2\mathbb{Z} \\ 0 & n \in 2\mathbb{Z}+1. \end{cases}$$

Sketch of proof: use the formula above; you pick up the main term behavior if and only if m is a square, which happens if and only if n is even.

Proposition 2.18

The measure

$$d\mu_p(x) := \frac{p+1}{\pi} \frac{\sqrt{1-x^2/4}}{(\sqrt{p}+1/\sqrt{p})^2 - x^2} dx$$

defined on $[-2,2]$ is the unique measure that satisfies

$$\int_{-2}^{2} U_n(x/2) d\mu_p(x) = \begin{cases} p^{-n/2} & \text{if } n \text{ is even} \\ 0 & \text{if } n \text{ is odd} \end{cases} \tag{2.4}$$

for all non-negative integers n.

Proof. In three steps.

Step 1: We will argue that

$$\sum_{m \geq 0} \frac{U_{2m}(x/2)}{p^m} = \frac{p+1}{(\sqrt{p}+1/\sqrt{p})^2 - x^2}. \tag{2.5}$$

First, we recall the generating function

$$\sum_{n \geq 0} U_n(x) t^n = \frac{1}{1 - 2tx + t^2}.$$

For $x := x/2$ and $t := 1/\sqrt{p}$, this implies

$$\sum_{n \geq 0} \frac{U_n(x/2)}{p^{n/2}} = \frac{\sqrt{p}}{(\sqrt{p}+1/\sqrt{p}) - x},$$

which implies that

$$\left(\sum_{n \geq 0} \frac{U_n(x/2)}{p^{n/2}} \right) \left(\sum_{m \geq 0} \frac{U_m(-x/2)}{p^{m/2}} \right) = \frac{p}{(\sqrt{p}+1/\sqrt{p})^2 - x^2}.$$

Therefore, it suffices to show that

$$(p+1) \left(\sum_{n \geq 0} \frac{U_n(x/2)}{p^{n/2}} \right) \left(\sum_{m \geq 0} \frac{U_m(-x/2)}{p^{m/2}} \right) = p \sum_{m \geq 0} \frac{U_{2m}(x/2)}{p^m}. \tag{2.6}$$

We compute

$$\left(\sum_{n \geq 0} \frac{U_n(x/2)}{p^{n/2}} \right) \left(\sum_{m \geq 0} \frac{U_m(-x/2)}{p^{m/2}} \right) = \sum_{\ell \geq 0} \frac{\sum_{k=0}^{\ell} U_k(x/2) U_{\ell-k}(-x/2)}{p^{\ell/2}}.$$

We will find a closed form expression for the coefficient of $p^{-\ell/2}$:

$$\begin{aligned}\sum_{k=0}^{\ell} U_k(x/2)U_{\ell-k}(-x/2) &= \sum_{k=0}^{\ell}(-1)^{\ell-k}U_k(x/2)U_{\ell-k}(x/2)\\ &= \sum_{k=0}^{\lfloor \ell/2\rfloor}(-1)^{\ell-k}U_k(x/2)U_{\ell-k}(x/2)\\ &\quad + \sum_{k=\lfloor \ell/2\rfloor+1}^{\ell}(-1)^{\ell-k}U_k(x/2)U_{\ell-k}(x/2)\\ &= \sum_{k=0}^{\lfloor \ell/2\rfloor}(-1)^{\ell-k}\sum_{n=0}^{k}U_{\ell-2k+2n}(x/2)\\ &\quad + \sum_{k=\lfloor \ell/2\rfloor+1}^{\ell}(-1)^{\ell-k}\sum_{n=0}^{\ell-k}U_{2k-\ell+2n}(x/2),\end{aligned}$$

where we used the product formula $U_m(x)U_n(x) = \sum_{k=0}^{n} U_{m-n+2k}(x)$, valid for $m \geq n$. Expanding this, and cancelling greatly, we see that only the first term in every other inner sum contributes; concretely, this sum is precisely

$$\begin{aligned}N_\ell(x/2) := (-1)^{\ell-\lfloor \ell/2\rfloor}&\sum_{k=0}^{\left\lfloor \frac{\lfloor \ell/2\rfloor}{2}\right\rfloor} U_{\ell-2\lfloor \ell/2\rfloor+4k}(x/2)\\ + (-1)^{\ell-\lfloor \ell/2\rfloor-1}&\sum_{k=0}^{\left\lfloor \frac{\ell-\lfloor \ell/2\rfloor-1}{2}\right\rfloor} U_{2\lfloor \ell/2\rfloor-\ell+2+4k}(x/2).\end{aligned}$$

With this new notation, by (2.6) it suffices to show that

$$\sum_{\ell\geq 0}\frac{N_\ell(x/2)}{p^{\frac{\ell}{2}-1}} + \sum_{\ell\geq 0}\frac{N_\ell(x/2)}{p^{\ell/2}} = \sum_{m\geq 0}\frac{U_{2m}(x/2)}{p^{m-1}}.$$

Upon reindexing, this is clearly equivalent to

$$N_\ell(x/2) + N_{\ell-2}(x/2) = \begin{cases} 0 & \text{if } \ell \text{ is odd}\\ U_\ell(x/2) & \text{if } \ell \text{ is even.}\end{cases}$$

And this can be verified directly from the definition of N_ℓ.

Step 2: We will verify that the orthogonality relation (2.4) holds. Using the basic orthogonality relation

$$\int_{-1}^{1} U_n(x)U_m(x)\frac{\sqrt{1-x^2}}{2/\pi}dx = \delta_{m,n}$$

we compute

$$\begin{aligned}\int_{-2}^{2} U_n(x/2)d\mu_p(x) &= \int_{-2}^{2} U_n(x/2)\left(\sum_{m\geq 0}\frac{U_{2m}(x/2)}{p^m}\right)\frac{\sqrt{1-x^2/4}}{\pi}dx \\ &= \sum_{m\geq 0}\frac{1}{p^m}\int_{-2}^{2} U_n(x/2)U_{2m}(x/2)\frac{\sqrt{1-x^2/4}}{\pi}dx \\ &= \sum_{m\geq 0}\frac{1}{p^m}\delta_{n,2m} \\ &= p^{-n/2}.\end{aligned}$$

Step 3: We will argue that $d\mu_p$ is the unique measure satisfying the orthogonality relation (2.4). As the U_n are orthogonal polynomials, knowing the value $\int_{-2}^{2} U_n(x/2)d\mu_p$ for all n is equivalent to knowing the moments $m_n := \int_{-2}^{2} x^n d\mu(u)$ for all n. But by Carleman's condition, if every m_n is finite and $\sum_{n\geq 1} m_{2n}^{-1/2n} = +\infty$, then $d\mu_p$ is the only measure on $[-2,2]$ with (m_n) its sequence of moments. We can estimate that

$$|m_n| \leq \int_{-2}^{2} |x^n| d\mu_p \leq 2^n \int_{-2}^{2} d\mu_p = 2^n,$$

so the moments of $d\mu_p$ are finite. As the even moments are positive, we can estimate that

$$\frac{1}{m_{2n}^{1/2n}} \geq \frac{1}{(2^{2n})^{1/2n}} = \frac{1}{2},$$

hence $\sum_{n\geq 1} m_{2n}^{-1/2n} = +\infty$, as needed. □

So in other words,

$$d\mu_{p,k} \to d\mu_p \qquad \text{as } k \to \infty,$$

in the weak sense. This is a theorem of Serre (1997) published in JAMS. This is referred to as the *vertical Sato–Tate theorem*. This has been generalized to various contexts, as we can write down Hecke theory for basically all algebraic groups. But the ultimate version of generalization that one can write down is due to Shin-Templier. They studied every case where one can write down the trace formula; by assuming some functoriality conjecture, they computed the limiting distribtuion of Hecke eigenvalues for basically every algebraic group G. They classified the types of distributions you get, depending on G.

2.6 Effective equidistruibution

So far, the equidistribution results that we stated ($\{\alpha n\}$, vertical Sato–Tate) didn't specify a rate of convergence. But often we want to know "how fast" things equidistribute. Say $\{a_n\}$ is uniformly distributed mod 1, and let

$$\mu_N = \frac{1}{N}\sum_{n=1}^{N}\delta_{a_n}.$$

Then $d\mu_N(x) \to dx$ on $[0,1]$ as $N \to \infty$. There are two common ways of quantifying the rate of convergence:

1. Take a function on the unit circle, say $f \in C^\infty(S^1)$, and bound
$$\left| \int f d\mu_N - \int f dx \right|$$
in terms of some Sobolev norm $||f||_{W^{p,k}}$, and N.

2. The "discrepancy" is defined as
$$D(d\mu_N, dx) := \sup_{[\alpha,\beta]\subseteq[0,1]} |\mu_N(\alpha,\beta) - (\beta-\alpha)|.$$
If we can quantify how this converges to zero at $N \to \infty$, then this is one way of quantifying the rate of convergence.

These two notions are certainly not the same. For example, take $\{a_n\} = \{n\alpha\}$, where $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. We'll compute both discrepancies.

1. If f is a smooth function, then it has a fast-converging Fourier expansion $f = \sum b_m e(mx)$, so
$$\int f d\mu_N - \int f dx = \frac{1}{N} \sum_{m\neq 0} a_m \frac{e((N+1)m\alpha) - 1}{e(m\alpha) - 1}.$$
   (a) Take $\delta > 2$. Then we showed that there exists $M_{\delta,\alpha} > 0$ such that, for any $m > M_{\delta,\alpha}$, and for any $n \in \mathbb{Z}$, we have
   $$\left|\alpha - \frac{n}{m}\right| > \frac{1}{m^\delta}.$$
   (As there are only finitely many solutions if the inequality is flipped.) This is equivalent to $|m\alpha - n| > m^{1-\delta}$, hence
   $$|e(m\alpha) - 1| \gg_{\delta,\alpha} |m|^{1-\delta}.$$
   So by our above equality, we get
   $$\int f d\mu_N - \int f dx \ll_{\delta,\alpha} \frac{1}{N} \sum_{m\neq 0} |m|^{\delta-1} |b_m|.$$
   (b) Now we take care of $|b_m|$. By integration by parts,
   $$(2\pi m i)^k b_m = \int_0^1 e(-mx) f^{(k)}(x) dx.$$
   Namely, we just take the k'th derivative of f's Fourier expansion and integrate. This implies that
   $$|b_m| \leq \frac{1}{(2\pi|m|)^k} ||f||_{W^{k,\infty}(S^1)}.$$
   This implies that the size of the m'th Fourier coefficient of f is related to the differentiability of f.
   Now, let $k = 3$ and $\delta = 2.1$. Then we get
   $$\left| \int f d\mu_N - \int f d\mu \right| \ll_\alpha \frac{1}{M} ||f||_{W^{3,\infty}(S^1)}.$$
   So this is how you obtain the rate of decay in the first context.

2. Now we'll discuss how to bound the discrepancy, using the Erdős-Turan inequality. This inequality says that, for any $n > 1$, we have

$$D(d\mu_N, dx) \ll \frac{1}{n} + \sum_{k=1}^{n} \frac{\hat{\mu}_N(k)}{k},$$

where $\hat{\mu}_N(k) = \int e(kx) d\mu_N(x)$ is the k'th Fourier coefficient of the associated measure. Applying this to our case, we get that this is

$$\hat{\mu}_N(k) = \frac{1}{N} \frac{e((N+1)k\alpha) - 1}{e(k\alpha) - 1} \ll_{\delta,\alpha} \frac{|k|^{\delta-1}}{N},$$

so the discrepancy can be estimated as

$$D \ll_{\delta,\alpha} \frac{1}{n} + \sum_{k=1}^{n} \frac{k^{\delta-1}}{N} \ll \frac{1}{n} + \frac{n^\delta}{N}.$$

Now, choose n such that $1/n$ and n^δ/N both go to zero, and we want to choose it so that they're balanced (i.e. they go to zero at the same rate.) So we want $1/n = n^\delta/N$, so we choose $N = n^{1+\delta}$, so choose $n = N^{1/3}$, then we get that

$$D \ll_{\delta,\alpha} N^{-1/3+\epsilon}.$$

In fact, one can prove that the discrepancy is

$$D = O(\log N/N),$$

which is much better.

2.7 Density type results

Kronecker's theorem says $\{\alpha n\}$ becomes dense in $[0, 1]$. Density type results are those which quantify results such as Kronecker's theorem. These are even more complicated to deal with than equidistribution results. For example, one must answer questions like: *what is the minimum of $\epsilon_N > 0$ such that $\{a_n\}_{n=1}^N$ intersects any ball of radius ϵ_N?* (E.g. $\{a_n\}_{n=1}^N \subseteq [0, 1]$ implies $\epsilon_N > 1/N$ by the pigeonhole principle.)

Theorem 2.19

For $\{\{\alpha n\}\}$, we have

$$\limsup_{n\to\infty} N\epsilon_N > 1 + \frac{2\sqrt{5}}{5}.$$

Furthermore, equality is obtained exactly when $\alpha = (a\phi + d)/(c\phi + d)$, where ϕ is the Golden ratio, and $\gamma \in \mathrm{GL}_2(\mathbb{Z})$.

The standard proof of this uses continued fractions. This result says that the fractional part $\{\alpha n\}$ has large holes, since the theorem is only "obviously" true for 1 on the RHS. The smaller the lim sup for a particular α, the better it's distributed than other α's, since it'll have fewer holes. **This theorem says "the golden ratio is the best at filling out** $[0, 1]$**."** Typically, finding a sharp estimate for ϵ_N is extremely difficult.

2.8 Effective vertical Sato–Tate conjecture

Let us start with a variant of the Erdős-Turan inequality. This inequality only works for sequences that become uniformly distributed, so if we're working with a sequence that equidistributes with respect to another measure, then we need a variant.

A fact that we'll accept without proof: the sequence $\{x_n\} \subseteq [0,1]$ is equidistributed with respect to some continuous measure μ if and only if the following is true:

1. The "Weyl limit"
$$c_m := \lim_{N\to\infty} \frac{1}{N} \sum_{n\le N} e(mx_n)$$
exists for all $m \in \mathbb{Z}$, and in fact converges to $\int e^{mx} d\mu$. (This is clearly a necessary condition, but is not sufficient. So we must impose the following...)

2. $\lim_{N\to\infty} \frac{1}{N} \sum_{|m|<N} |c_m|^2 = 0$. (This imposes that the resulting measure is indeed continuous.)

If we assume further that
$$\sum |c_m| < \infty,$$
then we can define μ explicitly by $d\mu = f(x)dx$, where
$$f(x) = \sum c_m e(-mx).$$
The point is that finiteness of $\sum |c_m|$ guarantees absolute convergence of the RHS. Without this condition, we only have measure-wise equality; but with this finiteness condition, we're guaranteed pointwise equality.

Murty and Sinha proved the following variant of the Erdős-Turan inequality:

Theorem 2.20: (Murty-Sinha, 2009)

$$D(d\mu_N, d\mu) < \frac{||f||_{L^\infty}}{M+1} + 2 \sum_{1\le|m|\le M} \frac{1}{m} |\hat{\mu}_N(m) - c_m|.$$

This recovers Erdős-Turan when we set $d\mu_N = dx$; so in this case, the right-hand sum simplifies.

Now that we have a general Erdős-Turan with a general target measure, we can prove an effective Sato–Tate. One can try to work this out using
$$d\mu = \mu_p(x) = \frac{p+1}{\pi} \frac{\sqrt{1-x^2/4}}{(\sqrt{p}+1/\sqrt{p})^2 - x^2} \chi_{[-2,2]}(x) dx.$$
from before; but the problem is that computing the Weyl limit is not natural. It essentially comes out to computing the Fourier transform of this measure, but it's hard to integrate the above against $e(mx)$. Much more natural is to integrate this against $U_n(\cos(mx))$ with respect to x.

We first note that, under the change of variable $\theta \mapsto 2\cos\theta = x$, any measure $d\nu(x)$ on $[-2,2]$ corresponds to a unique "even" measure $d\tilde{\nu}(\theta)$. (Why is this? θ varies from -2π to 2π; but $\cos\theta$ is even; so if we have an even measure in terms of θ, that'll correspond to an even measure on $[-2,2]$ under this change of variable. For example, dx is sent to $-2\sin\theta d\theta$.) Next, observe that

$$D(d\mu_N, d\mu) = D(d\tilde{\mu}_N, d\tilde{\mu}).$$

(Why is this? The LHS quantity is the supremum of $|\int_a^b d\mu_N - \int_a^b d\mu|$, but under the change of variable given by the above correspondence, this is difference is precisely $2|\int_{a'}^{b'} d\tilde{\mu}_N - \int_{a'}^{b'} d\tilde{\mu}|$.) We'll need the identity

$$U_m(\cos\theta) - U_{m-2}(\cos\theta) = \frac{\sin(m+1)\theta - \sin(m-1)\theta}{\sin\theta} = 2\cos(m\theta).$$

This tells us that the Weyl limit for $\tilde{\mu}_{p,k}$ is

$$\widehat{\tilde{\mu}_{p,k}} = \int \cos(m\theta) d\tilde{\mu}_{pk}(x)$$

because $\tilde{\mu}_{p,k}$ is even. But this is

$$\begin{aligned}\int \cos(m\theta) d\tilde{\mu}_{pk}(x) &= \frac{1}{2}\int U_m(\cos\theta) - U_{m-2}(\cos\theta) d\tilde{\mu}_{p,k}(\theta) \\ &= \frac{1}{2}\int_{-2}^{2} U_m(x/2) - U_{m-2}(x/2) d\mu_{p,k}(x) \\ &= \frac{1}{2}\frac{1}{|B_k|}\sum_{f\in B_k} \lambda_f(p^m) - \lambda_f(p^{m-2}).\end{aligned}$$

So, we have an explicit Weyl limit by doing a change of variable. If we estimate the Eichler-Selberg trace formula, keeping track of p-dependency in the error estimate, we get the following quick application:

$$\left|\hat{\tilde{\mu}}_{p,k}(m) - c_m\right| \ll \frac{1}{k} p^{3m/2}\log p^m.$$

Applying the generalized Erdős-Turan inequality for $\tilde{\mu}_{p,k}$, we can estimate the discrepancy

$$D(d\mu_{p,k}, d\mu_p) \ll \frac{1}{M} + \frac{1}{k} p^{3M/2}(\log p^M)(\log M).$$

It's clear that fixing M and taking k large gives vertical Sato–Tate. But we want a precise error estimate. If we take $M = \lfloor c\log k/\log p\rfloor$, then we get

$$p^{3M/2} \ll p^{\frac{3}{2}c\frac{\log k}{\log p}} = k^{\frac{3}{2}c}.$$

So by taking c sufficiently small, we have that $p^{3M/2}/k$ goes to zero almost linearly in k. This means that, with this choice of M,

$$\frac{1}{k} p^{3M/2}(\log p^M)(\log M) \ll k^{-1+\delta}$$

for some small $\delta > 0$. As for the first part, we have $M^{-1} \ll \log p/\log k$, and since p is fixed, this goes to zero as $1/\log k$. This proves vertical Sato–Tate a la Murty and Sinha:

Theorem 2.21: (Murty-Sinha)

We have

$$D(\mu_{p,k}, \mu_p) \ll \frac{\log p}{\log k}.$$

Some philosophical remarks:

1. Because of the term $p^{3M/2}$ in the original discrepancy estimate, one has to take M at least logarithmically small in k, since we need $p^{3M/2}$ to grow slower than k. But this means the whole expression can't decay any faster than $1/\log k$, from the original Murty-Sinha estimate. That is, by the same term, improving the depending constant c in the final statement requires a power saving estimate in the error term of the trace formula, which is extremely difficult. For example, we saw $c < 2/3$. So to make the final statement effective, and to get a good constant, we need the $3/2$ in the $p^{3M/2}$ to be smaller, since that gives the barrier. This is where the "effective" comes from.

   In particular, if we want some bound like $1/k^\delta$, then we'd need some exponentially good savings for k in the Eichler-Selberg trace formula. Which is absurd; one can prove this is not even true.

2. *Thus, if we're going to prove a discrepancy using Erdős-Turan, then there is no way to improve this past* $1/\log k$. So we probably need some tools that are better than the Erdős-Turan inequality to improve the Murty–Sinha bound. But this is not known, at the moment. But philosophically, we don't think not having the right tool is the real problem. Rather, we think that the definition of the discrepancy $D(\cdot, \cdot)$ as the supremums of integrals, being naturally not smooth, is the main reason why we can't prove anything stronger than $1/\log k$. And this may be seen by the Fourier transform.

3. Experimental computations suggest that the discrepancy is better than the Murty–Sinha bound. From a theorem of Gamburd-Jacobson-Sarnak:

   Theorem 2.22

   $D(\mu_{p,k}, \mu_p) = \Omega_p(\frac{1}{\sqrt{k}(\log k)^2})$ as $k \to \infty$. Furthermore, numerical experiments predict that

   $$D(\mu_{p,k}, \mu_p) = O(k^{-\frac{1}{2}+\epsilon}).$$

   This follows from the square root cancellation from the sums of Hecke eigenvalues (numerical experiments show us that normalized Hecke eigenvalues behave like random numbers.)

2.9 Digression: motivating arithmetic quantum chaos

All of the theorems we've proven so far fall into the field of **arithmetic quantum chaos**. Suppose $\{x_n\}$ is a sequence which is i.i.d., with $E(x_n) = 0$ (so it's centered at zero) and its variance is $V(x_n) = 1$. The

central limit theorem says

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{N} x_n \to N(0,1),$$

the standard normal distribution. What does this imply? For large N, we have that

$$P\left(|\frac{1}{\sqrt{N}}\sum_{n=1}^{N} x_n| < A\right) \sim \frac{1}{\sqrt{\pi}}\int_{-A}^{A} e^{-x^2}dx.$$

So, for any $\epsilon > 0$,

$$P\left(|\frac{1}{\sqrt{N}}\sum_{n=1}^{N} x_n| < N^{\epsilon}\right) \to 1.$$

In other words, it is "almost surely" true that

$$\left|\sum_{n=1}^{N} x_n\right| < N^{\frac{1}{2}+\epsilon}.$$

For example, we think that $\{\mu(n)\}$ is random, i.e. that it takes values in $1, 0$, and 1 with probability $3/\pi^2, 1-6/\pi^2$, and $3/\pi^2$, respectively. (Why these numbers? They're the density of square-free numbers! Specifically, $6/\pi^2$ is the probability that a random number has a square factor. This is a relatively "naive" way of guessing the distribution.)

If this prediction is true, it implies RH, but only if we can quantify the rate of convergence and show that there is square-root cancellation. Namely, RH is equivalent to the claim that

$$|\sum_{n=1}^{N} \mu(n)| < N^{\frac{1}{2}+\epsilon}$$

for all large N. We at least do know that $\mu(n)$ is equidistributed with respect to the above measure. And in fact, that is equivalent to the prime number theorem (with no error bound.)

For example, for Dirichlet L-functions, the corresponding quantities are $\{\chi_d(n)\mu(n)\}$. We know that

$$\sum_{n<N} \chi_d(n)\mu(n) = o(N),$$

which is equivalent to Dirichlet's theorem on primes in arithmetic progressions. In this case,

$$\sum_{n<N} \chi_d(n)\mu(n) = O_\epsilon(N^{\frac{1}{2}+\epsilon}) \iff \text{GRH}.$$

The estimate on the LHS says *Dirichlet characters and the Mobius function aren't correlated.*

So why is the field called arithmetic quantum chaos? Quantum chaos is all about the relationship between Hamiltonian systems and the corresponding quantized system, when the underlying Hamiltonian system has chaotic behavior. But if you're looking at the Hamiltonian which corresponds to the free particle, then the corresponding quantized system is going to be described by the Laplacian, as the Schrodinger equation is not going to have any potential on it. Therefore, understanding the corresponding quantized system can be done by looking at the eigenspaces of the Laplacian. And on $\mathbb{H}$, geodesic flow is chaotic, so Hamiltonian

dynamics is chaotic (this is a famous "hyperbolicity theorem.") Therefore, if you want to understand the corresponding quantized system on $\mathbb{H}$, then you must understand the spectrum of the Laplacian, namely, modular forms and Maass forms.

(Lecture 14: October 27, 2020)

To recast our earlier discussion, $\mu(n)$ is uncorrelated with the constant sequence $1, 1, 1, \ldots$. And the prime number theorem in arithmetic progressions tells us $\mu(n)$ is uncorrelated with $\chi_d(1), \chi_d(2), \ldots$ But the latter sequence is determined by the first d elements of the sequence, by periodicity. So the sequence $\chi_d(1), \chi_d(2), \ldots$ is very far from being random; in fact, it's a deterministic sequence.

The Möbius disjointness conjecture says: for any deterministic sequence $\{a_n\}$, we must have

$$\sum_{n \leq N} a_n \mu(n) = o(N).$$

To wave our hands at what it means for a sequence $\{a_n\}$ to be deterministic... consider $a_1, \ldots, a_N$, and within this truncated sequence, look at the n-length adjacent subsequences. There are $N - n$ of these, and we count the number of distinct segments. For example, the constant sequence gives 1, uniformly in n and N; looking at the Dirichlet character, you'll get at most d distinct segments, unformly in N. So we can try to estimate how the number of distinct segments grows in terms of n and N; and this gives a quantity called the "entropy" of the sequence. For a sequence to be "deterministic" means the entropy is zero. The Möbius disjointness conjecture implies Chowla's conjecture.

Chowla's conjecture: If $r_1, \ldots, r_m \in \{1, 2\}$ with not all $r_i = 2$, $a_1, \ldots, a_m$ are distinct, then

$$\frac{1}{N} \sum_{n \leq N} \mu^{r_1}(n + a_1) \cdots \mu^{r_m}(n + a_m) \to 0.$$

This conjecture is so hard that not even a single case has been proven to date. For example, we don't even know if the Mobius function is uncorrelated with itself, i.e., we don't know if

$$\frac{1}{X} \sum_{n \leq X} \mu(n)\mu(n + 1) \to 0.$$

We bring this up since it was a very hot topic recently in analytic number theory. Recent theorems in this direction:

1. If we define $\lambda(n) = (-1)^{\#\text{prime factors of } n}$, then Matomäki-Radziwill proved that

$$\liminf \frac{1}{X} \sum_{n \leq X} \lambda(n)\lambda(n + 1) > -\frac{1}{6}.$$

   If there were no cancellation, this would be either -1 or 1. But what they showed is that there does exist some positive proportion of cancellation in this summation. This was published in the Annals, and people care a lot about this result because it represents progress towards Chowla's conjecture.

2. "The logarithmic Chowla's conjecture" says

$$\frac{1}{\log X}\sum \frac{\mu^{r_1}(n+a_1)\cdots\mu^{r_m}(n+a_m)}{n} = o(1).$$

This was proved in 2018, and was also published in the Annals.

The takeaway: recent work on Chowla has included proving it with extra weight. The reason these are so important:

3. (Green-Tao) Mobius disjointness conjecture is true for nil-sequences (this is a deterministic sequence which is determined by a nilpotent flow.) This allowed them to show that for all N, there exist N distinct primes $p_1, \ldots, p_N$ and d such that $p_{i+1} - p_i = d$.

The upshot: you can get some very incredible results by looking at randomness in μ, as well as progress towards Chowla's conjecture.

2.10 Overview of Hamiltonian dynamics

In this class we've been studying modular forms. There is in fact some randomness going on in this area as well; so there should be some conjectured analog to Chowla that explains the expected behavior of modular forms. In order to discuss this, we need to understand some physics.

The Hamiltonian, which represents total energy, is

$$H = T + V,$$

where $T = P^2/2m$ is kinetic energy, P is momentum, m is the mass, and V is the potential. Any motion is going try to minimize the total energy. The case relevant to modular forms is when there is no potential at all, $V = 0$. So the dynamics is described by the total energy T; in this case, H describes the energy of free particle. So movement is only along geodesics, since geodesics minimize length. And if you think of a particle with fixed mass, your total energy H is a function of x and $P = m\dot{x}$. In other words, the total energy H is a function on a phase space $\{(x,p)\}$, where you record the position and the velocity at the same time. If you bring this to a manifold M, then the setup of pairs $\{(x,p)\}$ is identified with TM, the tangent bundle over M. *So the Hamiltonian dynamics described by a potential-free Hamiltonian H is actually equivalent to geodesic flow on M.* The *geodesic flow* on M is defined by

$$\Phi : TM \to TM : (x,p) \mapsto (\mathrm{Exp}_x(p), q).$$

Here, (x,p) specifies a point x on M and its direction p; $\mathrm{Exp}_x(p)$ is the point obtained by moving a point x according to the tangent vector p along a geodesic; and the resulting vector is the parallel transport q, which is obtained by moving p along this geodesic. This is called geodesic flow because any trajectory of Φ is given by a geodesic. Note that this geodesic flow is equivalent to Hamiltonian dynamics for a free particle, as the latter follows geodesics when it's free to move in this space.

The way that physicists quantize this, thereby obtain quantum dynamics, is by mapping the momentum P to a differential operator; i.e. the corresponding "quantum system" is obtained by naively mapping

$$P \mapsto \frac{1}{i}\partial_x.$$

In this case, the total energy becomes the negative Laplacian,

$$H = -\Delta,$$

since $P^2 = -(\partial_x)^2$. The dynamics is described by the evolution equation

$$i\partial_t \phi(t,x) = \Delta\phi(t,x),$$

which is *the potential-free version of the Schrodinger equation.* One strategy of solving this equation is by separation of variables; if we do this, we obtain

$$\phi(t,x) = \sum a_n e^{-it\lambda_n}\phi_n(x),$$

where $-\Delta\phi_n(x) = \lambda_n\phi_n(x)$. In words, we write the solution $\phi(t,x)$ as an infinite series, where each summand is the product of something that oscillates in t and an eigenfunction of the Laplacian. This means that *the eigenstates $\phi_n(x)$ of $-\Delta$ describe the potential-free quantum system.* Even though (on the surface) this quantum system has nothing to do with the classical Hamiltonian system, physicists believe there is a close relation between these two systems.

The subject *quantum chaos* concerns the relationship between classical Hamiltonian dynamics and quantum dynamics (namely, the behavior of these eigenstates ϕ_n.) The general philsophy of this correspondence is based on Berry's 1997 paper, which is referred to as "Berry's Random Wave Model." It essentially says that *if the classical dynamics is chaotic, then the corresponding quantum dynamics must be random.* Meaning, you shouldn't see the eigenfunctions align, or anything like that. In particular, he stated: ϕ_n **should behave like a random wave.**

Now we explain the relevance to modular forms. $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ has hyperbolic structure, meaning that the sectional curvature is -1 everywhere; in contrast, a sphere has curvature 1 everywhere and the Euclidian plane 0. Therefore, $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ is *dispersing*, meaning that if you shoot two geodesics into directions that are slightly different, then the distance between those particles will grow like an exponential function; in constrast, in the Euclidian plane, the distance only grows like a linear function, and in positively curves surfaces such as S^2, they will always meet each other eventually. We know that in this case, the geodesic flow is chaotic. So according to the Random Wave Model, eigenfunctions of the Laplace Beltrami operator, which in this case is $-\Delta = -y^2(\partial_x^2 + \partial_y^2)$, should behave like a random wave. We know that these eigenfunctions are Maass forms.

There are many conclusions we may draw from the Random Wave Model, i.e., the naïve belief that the Maass forms ϕ_n behave like a random wave.

1. *The L^∞ conjecture* says
$$\|\phi_n\|_{L^\infty} = O_\epsilon(\lambda_n^\epsilon).$$
This is important because it implies the Lindelöf hypothesis for ζ. A non-trivial estimate for L^∞ norm for Maass forms was proven in the early 90's.

2. It's conjectured that $\lim_{n\to\infty} ||\phi_n||_{L^p}$, for p even, is the p-th moment of the standard normal distribution, $N(0,1)$. We have a good understanding of the fourth moment, from the past few years. A group of people recently computed this fourth moment, assuming the Ramanujan conjecture for Maass forms.

3. If you're given an eigenstate, then the corresponding density function is given by $|\phi_n|^2 dV$. If things were random, then this measure should not be concentrated everywhere; it ought to "see" every part of the surface $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ equally. Therefore, we expect $|\phi_n|^2 dV$ to converge to the volume form dV as $n \to \infty$. This is the *arithmetic quantum unique ergodicity theorem* of Lindenstrauss and Soundararajan (and this won Lindenstrauss the Fields medal to Lindenstrauss.)

Often, these implications are too strong to be claimed. So, we want to understand where the truth lies in. That's why we try to give a better estimate for L^∞ norms of modular and Maass forms.

Arithmetic quantum chaos is mainly about properties of Maass forms, as these are the relevant eigenfunctions. But there is a more general belief: we believe that modular forms f, as the weight $k \to \infty$, should exhibit the same asymptotic behavior as Maass forms ϕ_n, as the Laplacian eigenvalue $\lambda_n \to \infty$. Physicists gave us the philosophy of Maass forms ought to behave like a random wave; the number theorists ran with this, and now believe that modular forms should as well. A consequence of this is the **mass equidistribution of modular forms.** This analog of the arithmetic quantum unique ergodicity theorem of Sound and Lindenstrauss says is that
$$|f|^2 y^k dV \to dV.$$
And this was proven by Sound-Holowinsky around 2010. **The point: increasing the weight of your form to ∞ is exactly the analog of increasing your Laplace eigenvalue to ∞.** The Petersson trace formula and the Eichler-Selberg trace formula are the most important tools for studying this phenomenon. If this is true, then for example, then Hecke eigenvalues must behave like a random number. And this was proven by Serre, which is referred to as vertical Sato–Tate.