MODULAR FORMS (DRAFT, CTNT 2016)

KEITH CONRAD

1. Introduction

A modular form is a holomorphic function on the upper half-plane

$$\mathfrak{h} = \{x + iy : x \in \mathbf{R}, y > 0\} = \{\tau \in \mathbf{C} : \operatorname{Im} \tau > 0\}$$

that transforms in a certain way under a discrete matrix group and has a nice behavior at infinity. To explain this more precisely (see Definition 1.2 below) we introduce a few 2×2 real matrix groups.

Definition 1.1. Set

$$\begin{aligned} \mathrm{GL}_2(\mathbf{R}) &= \{A \in \mathrm{M}_2(\mathbf{R}) : \det A \neq 0\}, \\ \mathrm{GL}_2^+(\mathbf{R}) &= \{A \in \mathrm{M}_2(\mathbf{R}) : \det A > 0\}, \\ \mathrm{SL}_2(\mathbf{R}) &= \{A \in \mathrm{M}_2(\mathbf{R}) : \det A = 1\}. \end{aligned}$$

These are all groups under matrix multiplication, with identity $I_2 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. The notations GL and SL stand for "general linear" and "special linear," where the word "special" is shorthand for "determinant 1." Clearly $\mathrm{GL}_2(\mathbf{R}) \supset \mathrm{GL}_2^+(\mathbf{R}) \supset \mathrm{SL}_2(\mathbf{R})$.

We will be interested in discrete subgroups of $\mathrm{GL}_2(\mathbf{R})$, especially the integer-matrix analogue of $\mathrm{SL}_2(\mathbf{R})$, which is[1]

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

If you pick three integers in a 2×2 matrix and solve for the fourth to have $ad - bc = 1$, usually it won't be an integer so you don't get a matrix in $\mathrm{SL}_2(\mathbf{Z})$. To create a matrix in $\mathrm{SL}_2(\mathbf{Z})$ "randomly," pick any pair of relatively prime integers for the first column and solve for the second column using Euclid's algorithm. For example, to find a matrix $\left(\begin{smallmatrix} 18 & x \\ 25 & y \end{smallmatrix}\right)$ in $\mathrm{SL}_2(\mathbf{Z})$ is the same as solving $18y - 25x = 1$ in integers x and y.

Definition 1.2. Let $k \in \mathbf{Z}$. A *modular form of weight* k for $\mathrm{SL}_2(\mathbf{Z})$ is a function $f \colon \mathfrak{h} \to \mathbf{C}$ such that

(1) f is holomorphic on $\mathfrak{h}$,
(2) $f\left(\dfrac{a\tau + b}{c\tau + d}\right) = (c\tau + d)^k f(\tau)$ for all $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbf{Z})$ and all $\tau \in \mathfrak{h}$,
(3) the values $f(\tau)$ are bounded as $\operatorname{Im} \tau \to \infty$.

Often "$\operatorname{Im} \tau \to \infty$" is written as $\tau \to i\infty$ and we think of $i\infty$ as a point infinitely high up in $\mathfrak{h}$, analogous to ∞ and $-\infty$ lying infinitely far to the right or left of $\mathbf{R}$.

[1] The group $\mathrm{GL}_2(\mathbf{Z})$ is *not* the 2×2 integer matrices with nonzero determinant, since that is not a group: the inverse of such a matrix need not have integer entries. Instead, $\mathrm{GL}_2(\mathbf{Z}) = \{A \in \mathrm{M}_2(\mathbf{Z}) : \det A = \pm 1\}$.

Remark 1.3. The three defining properties of a modular form are independent of each other: there are functions $\mathfrak{h} \to \mathbf{C}$ satisfying any two of the three properties but not satisfying the third (for some choice of k).

The zero function on $\mathfrak{h}$ is a modular form of every weight. We will eventually see that the only modular form of negative weight, odd weight, or weight 2 for $\mathrm{SL}_2(\mathbf{Z})$ is the function 0, the only modular forms of weight 0 for $\mathrm{SL}_2(\mathbf{Z})$ are constant functions, and for every even $k \geq 4$ we'll use a construction called Eisenstein series in Section 4 to give a nonzero example of a modular form of weight k for $\mathrm{SL}_2(\mathbf{Z})$.

The second property in the definition of a modular form is called the *modularity condition.* Let's make it explicit in three examples.

Example 1.4. For the matrix $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbf{Z})$, the modularity condition means $f(\tau+1) = f(\tau)$ for all $\tau \in \mathfrak{h}$. The weight k plays no role here.

Example 1.5. For the matrix $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbf{Z})$, the modularity condition means $f(-1/\tau) = \tau^k f(\tau)$ for all $\tau \in \mathfrak{h}$. Here we see k appears prominently.

Example 1.6. For the matrix $\left(\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbf{Z})$, the modularity condition means $f(\tau) = (-1)^k f(\tau)$ for all $\tau \in \mathfrak{h}$, so if k is odd then f is identically zero: the only modular form of any odd weight for $\mathrm{SL}_2(\mathbf{Z})$ is the zero function.[2]

It is no surprise that modular forms might have (and do have!) applications in complex analysis, since by definition they are certain holomorphic functions. They are also connected to many other areas of math, such as combinatorics, number theory, geometry (both hyperbolic geometry and algebraic geometry), representation theory, and mathematical physics. Here are some reasons for these other connections.

(1) Modular forms can be expanded into power series in the complex variable $q = e^{2\pi i\tau}$ (this is called a q-expansion), and many q-series in combinatorics turn out to be modular forms or closely related to modular forms.
(2) The theta-function of a positive-definite quadratic form in number theory is a modular form and the L-function of an elliptic curve over $\mathbf{Q}$ (a generalization of the Riemann zeta-function) is also the L-function of a modular form. The link between elliptic curves and modular forms is how Wiles proved Fermat's Last Theorem: a counterexample to Fermat's Last Theorem leads to a contradiction of what we know about modular forms.
(3) The upper half-plane $\mathfrak{h}$ is a model for hyperbolic geometry, and constructions on $\mathfrak{h}$ that are relevant to modular forms (*e.g.*, fundamental domains and the Petersson inner product) have an appealing interpretation using the language of hyperbolic geometry.
(4) Modular forms provide embeddings of certain algebraic varieties into projective space.
(5) A modular form can be turned into a representation of an adelic matrix group.
(6) Generating functions in string theory and conformal field theory can be described in terms of modular forms.

[2]Modular forms can be defined for finite-index subgroups of $\mathrm{SL}_2(\mathbf{Z})$, and when the subgroup does not contain $-I_2$ there might be nonzero modular forms of odd weight for that subgroup.

2. Why the modularity condition?

Why would anyone think the equation

$$f\left(\frac{a\tau+b}{c\tau+d}\right) = (c\tau+d)^k f(\tau)$$

in the definition of a modular form is interesting? It arose from 19th century developments in complex analysis and geometry, which we will discuss in this section.

While the group $\mathrm{GL}_2(\mathbf{R})$ acts on $\mathbf{R}^2$ by linear transformations (any 2×2 matrix A sends each vector $\mathbf{v}$ in $\mathbf{R}^2$ to the vector $A\mathbf{v}$ in $\mathbf{R}^2$, and $I_2\mathbf{v} = \mathbf{v}$ and $A(B\mathbf{v}) = (AB)\mathbf{v}$ for all A and B in $\mathrm{GL}_2(\mathbf{R})$), the group $\mathrm{GL}_2^+(\mathbf{R})$ acts on $\mathfrak{h}$ by linear fractional transformations: for $\tau \in \mathfrak{h}$, define

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\tau := \frac{a\tau+b}{c\tau+d}. \tag{2.1}$$

The reason (2.1) lies in $\mathfrak{h}$ follows from the imaginary part formula

$$\mathrm{Im}\left(\frac{a\tau+b}{c\tau+d}\right) = \frac{(ad-bc)\,\mathrm{Im}\,\tau}{|c\tau+d|^2}, \tag{2.2}$$

for $\tau \in \mathbf{C} - \{-d/c\}$ and real a, b, c, d. By this formula, which the reader can check as an exercise, if $\tau \in \mathfrak{h}$ and $ad - bc > 0$ then $(a\tau + b)/(c\tau + d) \in \mathfrak{h}$. To show (2.1) defines a (left) group action of $\mathrm{GL}_2^+(\mathbf{R})$ on $\mathfrak{h}$, check that $I_2\tau = \tau$ and $A(B\tau) = (AB)\tau$ for all A and B in $\mathrm{GL}_2^+(\mathbf{R})$.

For $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{GL}_2^+(\mathbf{R})$ and $x \in \mathbf{R}^\times$, the matrix $\left(\begin{smallmatrix} xa & xb \\ xc & xd \end{smallmatrix}\right)$ is in $\mathrm{GL}_2^+(\mathbf{R})$ (its determinant is $x^2(ad - bc)$) and it acts on $\mathfrak{h}$ in the same way as $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ does since $(xa\tau + xb)/(xc\tau + xd) = (a\tau + b)/(c\tau + d)$. This is different from $\mathrm{GL}_2(\mathbf{R})$ acting as linear transformations on $\mathbf{R}^2$, where different matrices have different effects somewhere (in fact on either $\binom{1}{0}$ or $\binom{0}{1}$). Using $x = 1/\sqrt{ad-bc}$ shows every matrix in $\mathrm{GL}_2^+(\mathbf{R})$ acts on $\mathfrak{h}$ in the same way as a matrix in $\mathrm{SL}_2(\mathbf{R})$.

One of the reasons for interest in linear fractional transformations of $\mathfrak{h}$ by matrices in $\mathrm{SL}_2(\mathbf{R})$ is the classification of compact surfaces. Aside from the Riemann sphere $\widehat{\mathbf{C}} = \mathbf{C} \cup \{\infty\}$ and a torus $\mathbf{C}/L$ for any lattice L in $\mathbf{C}$, every other compact *orientable* surface can be realized as a quotient space $\Gamma\backslash\mathfrak{h} = \{\Gamma\tau : \tau \in \mathfrak{h}\}$ where $\mathfrak{h}$ is acted on from the left by some discrete subgroup Γ of $\mathrm{SL}_2(\mathbf{R})$ using linear fractional transformations. This should be thought of as a two-dimensional analogue of the construction of a circle as a quotient space $\mathbf{R}/\mathbf{Z}$, where $\mathbf{Z}$ acts on $\mathbf{R}$ as discrete additive translations ($x \mapsto x + n$ for $n \in \mathbf{Z}$).[3]

The similarity between a quotient of $\mathbf{C}$ by a lattice and a quotient of $\mathfrak{h}$ by a discrete subgroup of $\mathrm{SL}_2(\mathbf{R})$ becomes more striking when we use the language of geometry: a lattice in $\mathbf{C}$ acts on $\mathbf{C}$ as a discrete group of additive translations that each preserve Euclidean distances on $\mathbf{C}$, while linear fractional transformations of $\mathfrak{h}$ coming from matrices in $\mathrm{SL}_2(\mathbf{R})$ each preserve non-Euclidean distances on $\mathfrak{h}$ when we view $\mathfrak{h}$ as the hyperbolic plane (see Appendix A). From the viewpoint of Euclidean and non-Euclidean geometry, compact orientable surfaces other than $\widehat{\mathbf{C}}$ have similar descriptions: they arise as a model geometric

[3] While $\mathbf{R}$ has a group structure, with $\mathbf{Z}$ a subgroup of $\mathbf{R}$, $\mathfrak{h}$ does not have a group structure and discrete subgroups of $\mathrm{SL}_2(\mathbf{R})$ are generally noncommutative, so we write $\Gamma\backslash\mathfrak{h}$ rather than $\mathfrak{h}/\Gamma$ to emphasize the leftness of the group action. In contrast, there is no real difference between $\mathbf{R}/\mathbf{Z}$ and $\mathbf{Z}\backslash\mathbf{R}$ since the group structure on $\mathbf{R}$ is commutative. The backslash $\backslash$ in $\mathbf{Z}\backslash\mathbf{R}$ is important since writing $\mathbf{Z}/\mathbf{R}$ would be terrible.

space ($\mathbf{C}$ or $\mathfrak{h}$) modulo the action of an appropriate[4] discrete group of distance-preserving transformations of that space.

An important way to study a space is to study nice functions (continuous, smooth, analytic) on the space. For a discrete group Γ in $\mathrm{SL}_2(\mathbf{R})$, creating nice nonconstant complex-valued functions on $\Gamma\backslash\mathfrak{h}$ is the same thing as creating nice functions $f\colon \mathfrak{h} \to \mathbf{C}$ that are Γ-invariant: $f(\gamma\tau) = f(\tau)$ for all $\gamma \in \Gamma$ and $\tau \in \mathfrak{h}$. Two non-invariant functions lead to an invariant function if they fail to be invariant by the same fudge factor: if

$$f\left(\frac{a\tau+b}{c\tau+d}\right) = (c\tau+d)^k f(\tau) \quad\text{and}\quad g\left(\frac{a\tau+b}{c\tau+d}\right) = (c\tau+d)^k g(\tau)$$

for all $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$ and $\tau \in \mathfrak{h}$, and the same "weight" k, then the ratio $f(\tau)/g(\tau)$ is Γ-invariant:

$$\frac{f((a\tau+b)/(c\tau+d))}{g((a\tau+b)/(c\tau+d))} = \frac{(c\tau+d)^k f(\tau)}{(c\tau+d)^k g(\tau)} = \frac{f(\tau)}{g(\tau)}.$$

But why should we use fudge factors of the form $(c\tau+d)^k$?

Suppose for a function $f\colon \mathfrak{h} \to \mathbf{C}$ that $f(\gamma\tau)$ and $f(\tau)$ are always related by a factor determined by $\gamma \in \Gamma$ and $\tau \in \mathfrak{h}$:

$$f(\gamma\tau) = j(\gamma,\tau)f(\tau) \tag{2.3}$$

for some function $j\colon \Gamma \times \mathfrak{h} \to \mathbf{C}$. That (2.1) defines a (left) group action of $\mathrm{SL}_2(\mathbf{R})$ on $\mathfrak{h}$ means in part that $(\gamma_1\gamma_2)\tau = \gamma_1(\gamma_2\tau)$, so $f((\gamma_1\gamma_2)\tau) = f(\gamma_1(\gamma_2\tau))$. This turns (2.3) into

$$j(\gamma_1\gamma_2,\tau)f(\tau) = j(\gamma_1,\gamma_2\tau)f(\gamma_2\tau). \tag{2.4}$$

Since $f(\gamma_2\tau) = j(\gamma_2,\tau)f(\tau)$, (2.4) holds if

$$j(\gamma_1\gamma_2,\tau) = j(\gamma_1,\gamma_2\tau)j(\gamma_2,\tau), \tag{2.5}$$

which looks like the chain rule $(f_1 \circ f_2)'(x) = f_1'(f_2(x))f_2'(x)$. This suggests a natural example of (2.5) using differentiation: when $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ set

$$j(\gamma,\tau) := \left(\frac{a\tau+b}{c\tau+d}\right)' = \frac{a(c\tau+d)-c(a\tau+b)}{(c\tau+d)^2} = \frac{ad-bc}{(c\tau+d)^2},$$

and for $\gamma \in \mathrm{SL}_2(\mathbf{R})$ this says $j(\gamma,\tau) = 1/(c\tau+d)^2$. When $j(\gamma,\tau)$ fits (2.5) so does $j(\gamma,\tau)^m$ for each $m \in \mathbf{Z}$, which motivates the consideration of the modularity condition with factors $1/(c\tau+d)^k$, at least for even k.

Exercises.

1. Prove (2.2).
2. Prove (2.1) defines a (left) group action of $\mathrm{GL}_2^+(\mathbf{R})$ on $\mathfrak{h}$.
3. Prove two matrices in $\mathrm{GL}_2^+(\mathbf{R})$ act in the same way everywhere on $\mathfrak{h}$ if and only if they are scalar multiples of each other.

[4]For some discrete subgroups Γ of $\mathrm{SL}_2(\mathbf{R})$, $\Gamma\backslash\,\mathrm{SL}_2(\mathbf{R})$ is not compact.

3. Simplifying the modularity condition for $\mathrm{SL}_2(\mathbf{Z})$

The only modular forms we have seen are boring: the zero function in any weight and constant functions in weight 0. Before giving interesting example of modular forms will use group theory to simplify the modularity condition in the definition of a modular form. It is an infinite set of equations, one for each matrix in $\mathrm{SL}_2(\mathbf{Z})$, but the following lemma will let us check the modularity condition on a set of generators for $\mathrm{SL}_2(\mathbf{Z})$ to know it holds for all matrices in the group.

Lemma 3.1. *If a function $f\colon \mathfrak{h} \to \mathbf{C}$ satisfies the modularity condition with weight k for two matrices γ_1 and γ_2 in $\mathrm{SL}_2(\mathbf{Z})$ then it satisfies the modularity condition with weight k for $\gamma_1\gamma_2$ and for the inverse γ_1^{-1}.*

Proof. Let $\gamma_1 = \left(\begin{smallmatrix} a_1 & b_1 \\ c_1 & d_1 \end{smallmatrix}\right)$ and $\gamma_2 = \left(\begin{smallmatrix} a_2 & b_2 \\ c_2 & d_2 \end{smallmatrix}\right)$. The modularity condition with weight k for these matrices says $f(\gamma_1\tau) = (c_1\tau + d_1)^k f(\tau)$ and $f(\gamma_2\tau) = (c_2\tau + d_2)^k f(\tau)$ for all $\tau \in \mathfrak{h}$. It follows that for all τ,

$$\begin{aligned} f((\gamma_1\gamma_2)\tau) &= f(\gamma_1(\gamma_2\tau)) \\ &= (c_1\gamma_2\tau + d_1)^k f(\gamma_2\tau) \\ &= (c_1\gamma_2\tau + d_1)^k (c_2\tau + d)^k f(\tau). \end{aligned}$$

Since $\gamma_2\tau = (a_2\tau + b_2)/(c_2\tau + d_2)$, a calculation shows

$$(c_1\gamma_2\tau + d_1)^k (c_2\tau + d)^k = ((c_1a_2 + d_1c_2)\tau + (c_1b_2 + d_1d_2))^k,$$

so

$$f((\gamma_1\gamma_2)\tau) = ((c_1a_2 + d_1c_2)\tau + (c_1b_2 + d_1d_2))^k f(\tau), \tag{3.1}$$

and the bottom matrix entries of

$$\gamma_1\gamma_2 = \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix}\begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix} = \begin{pmatrix} * & * \\ c_1a_2 + d_1c_2 & c_1b_2 + d_1d_2 \end{pmatrix}$$

are exactly the "c" and "d" that appear when we write $f((\gamma_1\gamma_2)\tau)$ as $(c\tau + d)^k f(\tau)$ in (3.1). Thus f satisfies the modularity condition with weight k for $\gamma_1\gamma_2$.

We now want to prove that if $f(\gamma_1\tau) = (c_1\tau + d_1)^k f(\tau)$ for all $\tau \in \mathfrak{h}$ then the same condition holds with γ_1 replaced by γ_1^{-1}, which is $\left(\begin{smallmatrix} d_1 & -b_1 \\ -c_1 & a_1 \end{smallmatrix}\right)$ because γ_1 has determinant 1. Replacing τ with $\gamma_1^{-1}\tau$ in the modularity condition for the matrix γ_1, we get

$$f(\tau) = (c_1(\gamma_1^{-1}\tau) + d_1)^k f(\gamma_1^{-1}\tau)$$

for all τ. Dividing both sides by $(c_1(\gamma_1^{-1}\tau) + d_1)^k$,

$$f(\gamma_1^{-1}\tau) = \frac{1}{(c_1\gamma_1^{-1}\tau + d_1)^k} f(\tau)$$

for all τ. Since $c_1\gamma_1^{-1}\tau + d_1 = (a_1d_1 - b_1c_1)/(-c_1\tau + a_1) = 1/(-c_1\tau + a_1)$,

$$f(\gamma_1^{-1}\tau) = (-c_1\tau + a_1)^k f(\tau)$$

for all τ, which is the modularity condition for γ_1^{-1}. □

Theorem 3.2. *If the set $\{\gamma_1, \ldots, \gamma_m\}$ generates $\mathrm{SL}_2(\mathbf{Z})$ and a function $f\colon \mathfrak{h} \to \mathbf{C}$ satisfies the modularity condition with weight k for each γ_i then f satisfies the modularity condition with weight k for all of $\mathrm{SL}_2(\mathbf{Z})$.*

Proof. By Lemma 3.1, the set of all $\gamma \in \mathrm{SL}_2(\mathbf{Z})$ for which f satisfies the modularity condition with weight k is a subgroup of $\mathrm{SL}_2(\mathbf{Z})$ (clearly the modularity condition holds when $\gamma = I_2$). Therefore if this subset contains a set of generators of $\mathrm{SL}_2(\mathbf{Z})$ it is all of $\mathrm{SL}_2(\mathbf{Z})$. □

Two particular elements in $\mathrm{SL}_2(\mathbf{Z})$ are

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

The matrix S has order 4 (check $S^2 = -I_2$), while the matrix T has infinite order (check $T^n = (\begin{smallmatrix} 1 & n \\ 0 & 1 \end{smallmatrix})$). As linear fractional transformations of $\mathfrak{h}$,

$$S\tau = -\frac{1}{\tau}, \quad T\tau = \tau + 1, \tag{3.2}$$

so as a transformation of $\mathfrak{h}$ the order of S is 2 rather than 4, while T has infinite order on $\mathfrak{h}$.

Theorem 3.3. *The group* $\mathrm{SL}_2(\mathbf{Z})$ *is generated by* S *and* T.

Proof. Let $G = \langle S, T\rangle$ be the subgroup of $\mathrm{SL}_2(\mathbf{Z})$ generated by S and T. We will give two proofs that $G = \mathrm{SL}_2(\mathbf{Z})$, one algebraic and the other geometric.

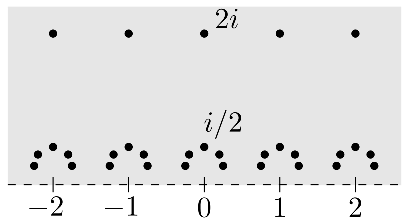
For the algebraic proof, we start by writing down the effect of S and T^n on any matrix by multiplication from the left:

$$S\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} -c & -d \\ a & b \end{pmatrix}, \quad T^n\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a+nc & b+nd \\ c & d \end{pmatrix}. \tag{3.3}$$

Now pick any $\gamma = (\begin{smallmatrix} a & b \\ c & d \end{smallmatrix})$ in $\mathrm{SL}_2(\mathbf{Z})$. Suppose $c \neq 0$. If $|a| \geq |c|$, divide a by c: $a = cq + r$ with $0 \leq r < |c|$. By (3.3), $T^{-q}\gamma$ has upper left entry $a - qc = r$, which is smaller in absolute value than the lower left entry c in $T^{-q}\gamma$. Applying S switches these entries (with a sign change), and we can apply the division algorithm in $\mathbf{Z}$ again if the lower left entry is nonzero in order to find another power of T to multiply by on the left so the lower left entry has smaller absolute value than before. Eventually multiplication of γ on the left by enough copies of S and powers of T gives a matrix in $\mathrm{SL}_2(\mathbf{Z})$ with lower left entry 0. Such a matrix, since it is integral with determinant 1, has the form $(\begin{smallmatrix} \pm 1 & m \\ 0 & \pm 1 \end{smallmatrix})$ for some $m \in \mathbf{Z}$ and common signs on the diagonal. This matrix is either T^m or $-T^{-m}$, so there is some $g \in G$ such that $g\gamma = \pm T^n$ for some $n \in \mathbf{Z}$. Since $T^n \in G$ and $S^2 = -I_2$, we have $\gamma = \pm g^{-1}T^n \in G$, so we are done.

In this algebraic proof, G acted on the set $\mathrm{SL}_2(\mathbf{Z})$ by left multiplication. For the geometric proof, we make G act on $\mathfrak{h}$ by linear fractional transformations. This action does not distinguish between matrices that differ by a sign (γ and $-\gamma$ act on $\mathfrak{h}$ in the same way), but this will not be a problem for the purpose of using this action to prove $G = \mathrm{SL}_2(\mathbf{Z})$ since $-I_2 = S^2 \in G$.

The key geometric idea is that when $\mathrm{SL}_2(\mathbf{Z})$ acts on a point in $\mathfrak{h}$, the orbit appears to accumulate towards the x-axis. This is illustrated by the picture below, which shows points in the $\mathrm{SL}_2(\mathbf{Z})$-orbit of $2i$ (including $S(2i) = -1/(2i) = i/2$). It appears that the imaginary parts of points in the orbit never exceed 2.



With the picture in mind, pick $\gamma \in \mathrm{SL}_2(\mathbf{Z})$ and set $\tau := \gamma(2i)$.

For any $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ in G, so $ad - bc = 1$, (2.2) tells us

$$\mathrm{Im}(g\tau) = \frac{\mathrm{Im}\,\tau}{|c\tau + d|^2}.$$

Write τ as $x + yi$. Then in the denominator

$$|c\tau + d|^2 = (cx + d)^2 + (cy)^2,$$

since $y \neq 0$ there are only finitely many integers c and d with $|c\tau + d|$ less than a given bound. Here τ is not changing but c and d are. Therefore $\mathrm{Im}(g\tau)$ has a *maximum* possible value as g runs over G (with τ fixed), so there is some $g_0 \in G$ such that

$$\mathrm{Im}(g\tau) \leq \mathrm{Im}(g_0\tau)$$

for all $g \in G$.

Since $Sg_0 \in G$, the maximality property defining g_0 implies $\mathrm{Im}((Sg_0)\tau) \leq \mathrm{Im}(g_0\tau)$, so (2.2) with $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = S$ gives us

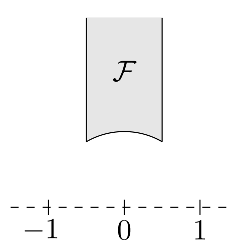
$$\mathrm{Im}(S(g_0\tau)) = \frac{\mathrm{Im}(g_0\tau)}{|g_0\tau|^2} \leq \mathrm{Im}(g_0\tau).$$

Therefore $|g_0\tau|^2 \geq 1$, so $|g_0\tau| \geq 1$. Since $\mathrm{Im}(T^n g_0\tau) = \mathrm{Im}(g_0\tau)$ and $T^n g_0 \in G$, replacing $g_0\tau$ with $T^n g_0\tau$ and running through the argument again shows $|T^n g_0\tau| \geq 1$ for all $n \in \mathbf{Z}$.

Applying T (or T^{-1}) to $g_0\tau$ adjusts its real part by 1 (or -1) without affecting the imaginary part. For some n, $T^n g_0\tau$ has real part between $-1/2$ and $1/2$. Using this power of T, we've shown that $\tau = \gamma(2i)$ has an element of its G-orbit in the set

$$\mathcal{F} = \{\tau \in \mathfrak{h} : |\mathrm{Re}(\tau)| \leq 1/2, |\tau| \geq 1\}. \tag{3.4}$$

See the picture below. Note $\mathrm{Im}\,\tau \geq \sqrt{3}/2 > 1/2$ for all $\tau \in \mathcal{F}$.



We started by picking the number $2i$ in $\mathcal{F}$ and any γ in $\mathrm{SL}_2(\mathbf{Z})$, and we showed there is some $g \in G$ such that the point $g(\gamma(2i)) = (g\gamma)(2i)$ is also in $\mathcal{F}$. By (2.2),

$$g\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbf{Z}) \Longrightarrow \mathrm{Im}((g\gamma)(2i)) = \frac{2}{4c^2 + d^2} \geq \frac{\sqrt{3}}{2},$$

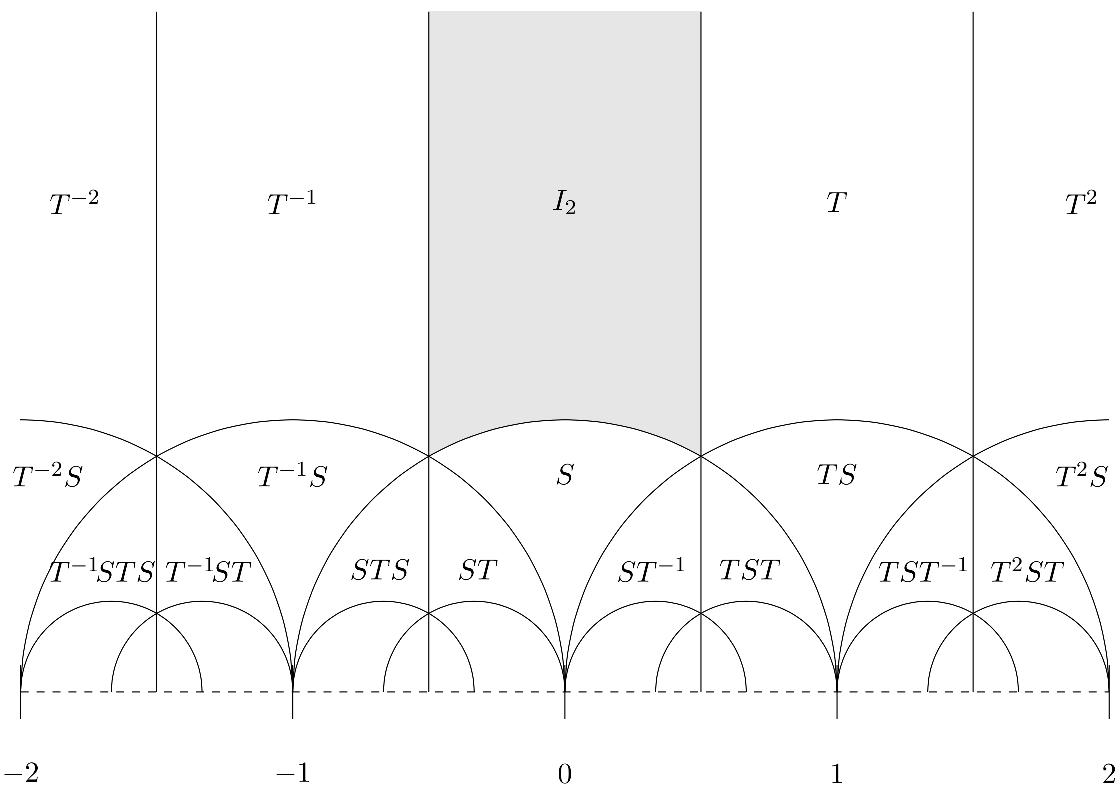
so $c = 0$ (otherwise the imaginary part is at most $2/(4c^2) \leq 1/2 < \sqrt{3}/2$). Then $ad = 1$, so $a = d = \pm 1$ and

$$\begin{pmatrix} a & b \\ 0 & d \end{pmatrix}(2i) = \frac{2ai + b}{d} = 2i \pm b.$$

For this to have real part between $\pm 1/2$ forces $b = 0$, so $g\gamma = \pm I_2$. Thus $\gamma = \pm g^{-1}$. Since $-I_2 = S^2 \in G$, we conclude $\gamma \in G$. □

The region $\mathcal{F}$ we drew above is called a *fundamental domain* for the action of $\mathrm{SL}_2(\mathbf{Z})$ on $\mathfrak{h}$. It is analogous to the role of $[0, 1]$ as a fundamental domain for the translation action of $\mathbf{Z}$ on $\mathbf{R}$: each point in the space ($\mathfrak{h}$ or $\mathbf{R}$) has a point of its orbit (by $\mathrm{SL}_2(\mathbf{Z})$ or $\mathbf{Z}$) in the fundamental domain ($\mathcal{F}$ or $[0, 1]$) and the only points in the fundamental domain that lie in the same orbit are on the boundary.

Below is a dissection of $\mathfrak{h}$ into translates $\gamma(\mathcal{F})$ as γ runs over $\mathrm{SL}_2(\mathbf{Z})$, with $\gamma = I_2$ corresponding to $\mathcal{F}$ itself. Different translates overlap only along boundary curves, and as we get closer to the x-axis $\mathfrak{h}$ is filled by infinitely many more of these translates. The fundamental domain and its translates are called "ideal triangles" since they are each bounded by three sides and have two endpoints in $\mathfrak{h}$ but one "endpoint" not in $\mathfrak{h}$: the third endpoint is either a rational number on the x-axis or (for the regions $T^n(\mathcal{F})$ with $n \in \mathbf{Z}$) is $i\infty$.



The description of $\mathcal{F}$ in (3.4) uses Euclidean geometry (the absolute value measures Euclidean distances in $\mathfrak{h}$) and is somewhat awkward. If we treat $\mathfrak{h}$ as the hyperbolic plane, for which the action of $\mathrm{SL}_2(\mathbf{Z})$ and more generally $\mathrm{SL}_2(\mathbf{R})$ is by isometries for the hyperbolic metric d_H (see Appendix A), then there is a prettier description of $\mathcal{F}$:

$$\mathcal{F} = \{\tau \in \mathfrak{h} : d_H(\tau, 2i) \leq d_H(\tau, \gamma(2i)) \text{ for all } \gamma \in \mathrm{SL}_2(\mathbf{Z})\}.$$

That is, $\mathcal{F}$ is the points of $\mathfrak{h}$ whose distance (as measured by the hyperbolic metric) to $2i$ is minimal compared to the distance to all points in the $\mathrm{SL}_2(\mathbf{Z})$-orbit of $2i$. The boundary of $\mathcal{F}$ is the points equidistant (for the hyperbolic metric) between $2i$ and one of its nearest $\mathrm{SL}_2(\mathbf{Z})$ translates $T(2i) = 2i+1$, $T^{-1}(2i) = 2i-1$, or $S(2i) = i/2$.[5] Part of what makes this geometric description of $\mathcal{F}$, called a *Dirichlet polygon*, attractive is that it also works for discrete groups actings by isometries on Euclidean spaces. For example, when $\mathbf{Z}$ acts on $\mathbf{R}$ by integer translations, for any $a \in \mathbf{R}$ the numbers whose distance to $a+\mathbf{Z} = \{a+n : n \in \mathbf{Z}\}$ is minimal at a is $[a-1/2, a+1/2]$ and this is a fundamental domain for $\mathbf{Z}$ acting on $\mathbf{R}$.

Example 3.4. We will carry out the algebraic proof of Theorem 3.3 to express $A = \left(\begin{smallmatrix} 17 & 29 \\ 7 & 12 \end{smallmatrix}\right)$ in terms of S and T.

Since $17 = 7 \cdot 2 + 3$, we want to subtract $7 \cdot 2$ from 17:

$$T^{-2}A = \begin{pmatrix} 3 & 5 \\ 7 & 12 \end{pmatrix}.$$

Now we want to switch the roles of 3 and 7. Multiply by S:

$$ST^{-2}A = \begin{pmatrix} -7 & -12 \\ 3 & 5 \end{pmatrix}.$$

Dividing -7 by 3, we have $-7 = 3 \cdot (-3) + 2$, so we want to add $3 \cdot 3$ to -7. Multiply by T^3:

$$T^3ST^{-2}A = \begin{pmatrix} 2 & 3 \\ 3 & 5 \end{pmatrix}.$$

Once again, multiply by S to switch the entries of the first column (up to sign):

$$ST^3ST^{-2}A = \begin{pmatrix} -3 & -5 \\ 2 & 3 \end{pmatrix}.$$

Since $-3 = 2(-2) + 1$, we compute

$$T^2ST^3ST^{-2}A = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix}.$$

Mutliply by S:

$$ST^2ST^3ST^{-2}A = \begin{pmatrix} -2 & -3 \\ 1 & 1 \end{pmatrix}.$$

Since $-2 = 1(-2) + 0$, multiply by T^2:

$$T^2ST^2ST^3ST^{-2}A = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}.$$

Multiply by S:

$$ST^2ST^2ST^3ST^{-2}A = \begin{pmatrix} -1 & -1 \\ 0 & -1 \end{pmatrix} = -T = S^2T.$$

Solving for A,

$$\begin{pmatrix} 17 & 29 \\ 7 & 12 \end{pmatrix} = A = T^2S^{-1}T^{-3}S^{-1}T^{-2}S^{-1}T^{-2}S^{-1}(S^2T) = T^2ST^{-3}ST^{-2}ST^{-2}ST \tag{3.5}$$

since $S^{-1} = -S$.

[5]We can replace $2i$ by yi for any $y > 1$ and the same description of $\mathcal{F}$ works.

Remark 3.5. Multiplication by the matrices S and T is closely related to continued fractions for rational numbers, with the caveat that the continued fraction algorithm should use nearest integers from above rather than from below. To illustrate, the matrix $\left(\begin{smallmatrix} 17 & 29 \\ 7 & 12 \end{smallmatrix}\right)$ is in $\mathrm{SL}_2(\mathbf{Z})$, and to obtain an expression for it in terms of S and T, we look at the ratio of the numbers in the first column, 17/7:

$$\frac{17}{7} = 3 - \frac{4}{7} = 3 - \frac{1}{7/4} = 3 - \frac{1}{2 - 1/4}.$$

Using the entries 3, 2, and 4 as exponents for T,

$$T^3ST^2ST^4S = \begin{pmatrix} 17 & -5 \\ 7 & -2 \end{pmatrix},$$

whose first column is what we are after. To get the correct second column, we solve $\left(\begin{smallmatrix} 17 & 29 \\ 7 & 12 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 17 & -5 \\ 7 & -2 \end{smallmatrix}\right)M$ for M, which is $\left(\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right) = T^2$, so

$$\begin{pmatrix} 17 & 29 \\ 7 & 12 \end{pmatrix} = \begin{pmatrix} 17 & -5 \\ 7 & -2 \end{pmatrix} T^2 = T^3ST^2ST^4ST^2.$$

This is a different expression for $\left(\begin{smallmatrix} 17 & 29 \\ 7 & 12 \end{smallmatrix}\right)$ than the one we found in (3.5). The representation of an element of $\mathrm{SL}_2(\mathbf{Z})$ as a product of powers of S and T is not unique.

Here, finally, is the simplified description of the modularity condition in the definition of a modular form for $\mathrm{SL}_2(\mathbf{Z})$.

Corollary 3.6. *For $k \in \mathbf{Z}$, a function $f\colon \mathfrak{h} \to \mathbf{C}$ is a modular form of weight k for* $\mathrm{SL}_2(\mathbf{Z})$ *if and only if*

(1) *f is holomorphic on $\mathfrak{h}$,*
(2) *$f(\tau + 1) = f(\tau)$ and $f\left(-\frac{1}{\tau}\right) = \tau^k f(\tau)$ for all $\tau \in \mathfrak{h}$,*
(3) *the values $f(\tau)$ are bounded as* $\mathrm{Im}\,\tau \to \infty$.

Proof. Use Theorems 3.2 and 3.3 together with (3.2). □

Exercises.

1. Find a matrix in $\mathrm{SL}_2(\mathbf{Z})$ with first column $\binom{39}{14}$.
2. Express the matrix $\left(\begin{smallmatrix} 8 & 7 \\ 9 & 8 \end{smallmatrix}\right)$, which is in $\mathrm{SL}_2(\mathbf{Z})$, as a product of powers of the matrices S and T.
3. If $f\colon \mathfrak{h} \to \mathbf{C}$ is a function satisfying the modularity condition for weight 4, show $f(\omega) = 0$ where $\omega = -1/2 + i\sqrt{3}/2$ is a nontrivial cube root of unity in $\mathbf{C}$, and if instead f satisfies the modularity condition for weight 6 then prove $f(i) = 0$.
4. For $k \in \mathbf{Z}$, a matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ in $\mathrm{GL}_2^+(\mathbf{R})$, and a function $f\colon \mathfrak{h} \to \mathbf{C}$, define the function $f|_k\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)\colon \mathfrak{h} \to \mathbf{C}$ by the formula

$$\left(f|_k \begin{pmatrix} a & b \\ c & d \end{pmatrix}\right)(\tau) = \frac{1}{(c\tau + d)^k} f\left(\frac{a\tau + b}{c\tau + d}\right).$$

   (a) Prove this formula defines a (right) group action of $\mathrm{GL}_2^+(\mathbf{R})$ on functions: $f|_k I_2 = f$ and $(f|_k A)|_k B = f|_k(AB)$ for all A and B in $\mathrm{GL}_2^+(\mathbf{R})$.

(b) If we want to view this action on functions as defined by the group of linear fractional transformations, not by matrices, why should we change the definition of the action by multiplying the formula by $(ad-bc)^{k/2}$? (See Exercise 2.3.)

5. For each $N \geq 1$, the *principal congruence subgroup of level* N is

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

where the matrix congruence is componentwise. This is the kernel of the reduction homomorphism $\mathrm{SL}_2(\mathbf{Z}) \to \mathrm{SL}_2(\mathbf{Z}/N\mathbf{Z})$, so $\Gamma(N)$ is a normal subgroup of $\mathrm{SL}_2(\mathbf{Z})$ with finite index.

Prove $\Gamma(2)$ is generated by the matrices $-I_2$, $\left(\begin{smallmatrix} 1 & 0 \\ 2 & 1 \end{smallmatrix}\right)$, and $\left(\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right)$. (Hint: Instead of the usual division algorithm in the first proof of Theorem 3.3, use a modified division algorithm: $a = bq + r$ where $|r| \leq |b/2|$ and possibly $r < 0$.)

4. Eisenstein Series and q-expansions

The most basic example of a nonconstant modular form for $\mathrm{SL}_2(\mathbf{Z})$ is an Eisenstein series.

Definition 4.1. For even $k \geq 4$, the weight k *Eisenstein series* is

$$G_k(\tau) := \sum_{\substack{(m,n)\in\mathbf{Z}^2 \\ (m,n)\neq(0,0)}} \frac{1}{(m\tau+n)^k}.$$

Our goal is to prove G_k is a modular form of weight k for $\mathrm{SL}_2(\mathbf{Z})$. The definition of $G_k(\tau)$ makes sense for odd $k \geq 3$, but in that case the series vanishes since the terms at (m,n) and $(-m,-n)$ cancel, so it is boring. (We already saw the only modular form of odd weight for $\mathrm{SL}_2(\mathbf{Z})$ is 0.)

First we prove absolute convergence.

Lemma 4.2. *For each $\tau \in H$ there is a $\delta = \delta_\tau \in (0,1)$ such that*

$$|m\tau + n| \geq \delta|mi+n|$$

for all $m, n \in \mathbf{Z}$.

Proof. If $m = 0$ then the desired inequality holds for all n provided we use $\delta \in (0,1)$.

If $m \neq 0$, then $|m\tau+n| \geq \delta|mi+n|$ is equivalent to $|\tau + n/m| \geq \delta|i+n/m|$, which in turn is equivalent to

$$\left|\frac{\tau+n/m}{i+n/m}\right| \geq \delta.$$

Rather than working with rational n/m, let's treat this as a task in real variables: set $f_\tau \colon \mathbf{R} \to \mathbf{R}$ by $f_\tau(x) = |(\tau-x)/(i-x)|$, so $f_\tau(x) > 0$ for all x. This is a continuous function and $f_\tau(x) \to 1$ as $x \to \pm\infty$. Therefore there is a large positive number R (depending on τ) such that $f_\tau(x) \geq 1/2$ for $|x| > R$. For $x \in [-R, R]$, positivity of $f_\tau(x)$ implies by compactness of $[-R,R]$ that there is some $c > 0$ such that $f_\tau(x) \geq c_\delta$ for all $x \in [-R,R]$. Therefore $f_\tau(x) \geq \delta$ for all $x \in \mathbf{R}$ when $\delta = \min(1/2, c)$. □

Theorem 4.3. *The Eisenstein series $G_k(\tau)$ is absolutely convergent: for each $\tau \in \mathfrak{h}$, the series $\sum_{(m,n)\neq(0,0)} 1/|m\tau+n|^k$ converges.*

Proof. Let $\delta = \delta_\tau$ be chosen as in Lemma 4.2. Then

$$\frac{1}{|m\tau+n|^k} \le \frac{1}{\delta^k|mi+n|^k} = \frac{1}{\delta^k\sqrt{m^2+n^2}^k}.$$

The exponent $k/2$ is greater than 1, so absolute convergence of $G_k(\tau)$ follows from absolute convergence of $\sum_{(m,n)\neq(0,0)} 1/\sqrt{m^2+n^2}^k$ for $k > 2$, which is proved in Section B as a special case of convergence of a lattice sum in any number of dimensions. □

Theorem 4.4. *For even $k \ge 4$, the Eisenstein series G_k is a modular form of weight k for $\mathrm{SL}_2(\mathbf{Z})$.*

Proof. By Theorem 4.3, $G_k(\tau)$ makes sense for each τ and the order of summation can be rearranged by absolute convergence. To prove G_k is holomorphic, we want to derive this from each term $1/(m\tau+n)^k$ in the series being holomorphic in τ. We will use a fundamental result of complex analysis about limits of holomorphic functions being holomorphic: if a sequence of holomorphic functions $\{f_n\}$ on a common domain $\Omega \subset \mathbf{C}$ converges uniformly on compact subsets of Ω then the pointwise limit $f(z) = \lim_{n\to\infty} f_n(z)$ is holomorphic on Ω.[6]

To apply this result to G_k, we will use a strengthening of Lemma 4.2: on each half-strip of the form $S_{a,b} = \{x+iy \in \mathfrak{h} : |x| \le a, y \ge b\}$ where $a > 0$ and $b > 0$, a value of δ can be chosen in Lemma 4.2 that works for all τ in $S_{a,b}$. The proof that such δ exists is left to the reader as an exercise (Exercise 4.1). Using this δ in the proof of Theorem 4.3 shows $G_k(\tau)$ converges uniformly on each $S_{a,b}$ by the Weierstrass M-test: the series $\sum_{(m,n)\neq(0,0)} 1/|m\tau+n|^k$ for $\tau \in S_{a,b}$ is bounded above termwise by $\sum_{(m,n)\neq(0,0)} 1/\delta^k\sqrt{m^2+n^2}^k$, which is independent of τ. Every compact subset of $\mathfrak{h}$ is contained in some $S_{a,b}$, so G_k converges uniformly on compact subsets of $\mathfrak{h}$ and thus is holomorphic.

To prove G_k satisfies the modularity condition with weight k, Corollary 3.6 tells us we have to check just two cases: $G_k(\tau+1) \overset{?}{==} G_k(\tau)$ and $G_k(-1/\tau) \overset{?}{=} \tau^k G_k(\tau)$. For the first condition,

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

As (m,n) runs over $\mathbf{Z}^2 - \{(0,0)\}$, so does $(m, m+n)$, so absolute convergence of the Eisenstein series lets us rearrange the terms:

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

For the second condition,

$$G_k(-1/\tau) = \sum_{(m,n)\neq(0,0)} \frac{1}{(-m/\tau+n)^k} = \tau^k \sum_{(m,n)\neq(0,0)} \frac{1}{(n\tau-m)^k}.$$

This last series is $G_k(\tau)$ by rearranging terms, so $G_k(-1/\tau) = \tau^k G_k(\tau)$.

The final property we have to check is behavior of $G_k(\tau)$ as $\tau \to i\infty$. We can assume $\operatorname{Im}\tau \ge 1$, and since $G_k(\tau+1) = G_k(\tau)$ we may also assume $|\operatorname{Re}(\tau)| \le 1/2$ as $\tau \to \infty$

[6]The analogue of this in real analysis is false: the Stone–Weierstrass theorem implies $|x|$ is a uniform limit of polynomials on $(-1,1)$, and polynomials are real-analytic but $|x|$ is not real-analytic on $(-1,1)$ because there's a problem at 0.

This is the half-strip $S_{1,1}$ described earlier in the proof, so there is some $\delta > 0$ such that $|m\tau + n| \geq \delta|mi + n|$ for all $\tau \in S_{1,1}$ and $m, n \in \mathbf{Z}$.

Rearrange the terms of $G_k(\tau)$:

$$G_k(\tau) = \sum_{n\neq 0} \frac{1}{n^k} + \sum_{m\neq 0}\sum_{n\in\mathbf{Z}} \frac{1}{(m\tau+n)^k} = 2\sum_{n\geq 1}\frac{1}{n^k} + 2\sum_{m\geq 1}\sum_{n\in\mathbf{Z}}\frac{1}{(m\tau+n)^k}, \tag{4.1}$$

where we write the sum over nonzero n and outer sum over nonzero m as twice a sum over positive n and positive m using evenness of k. We will show the double series, where every term has τ in it, tends to 0 as $\tau \to \infty$, so $G_k(\tau) \to 2\sum_{n\geq 1} 1/n^k$ as $\tau \to i\infty$.

For any $N \geq 1$,

$$\begin{aligned}\sum_{m\geq 1}\sum_{n\in\mathbf{Z}}\frac{1}{|m\tau+n|^k} &= \sum_{m+|n|\leq N}\frac{1}{|m\tau+n|^k} + \sum_{m+|n|>N}\frac{1}{|m\tau+n|^k}\\ &\leq \sum_{m+|n|\leq N}\frac{1}{|m\tau+n|^k} + \frac{1}{\delta^k}\sum_{m+|n|>N}\frac{1}{|mi+n|^k}.\end{aligned}$$

Since $\sum_{m\geq 1, n\in\mathbf{Z}} 1/|m+ni|^k$ converges, for any $\varepsilon > 0$ the tail $\sum_{m+|n|>N} 1/|mi+n|^k$ is less than ε if N is sufficiently large and this doesn't involve τ. For such a choice of N, the finite series $\sum_{m+|n|\leq N} 1/|m\tau+n|^k$ is less than ε if $\operatorname{Im}\tau$ is sufficiently large. Thus the double series in (4.1) is less than 2ε if $\operatorname{Im}\tau$ is sufficiently large. □

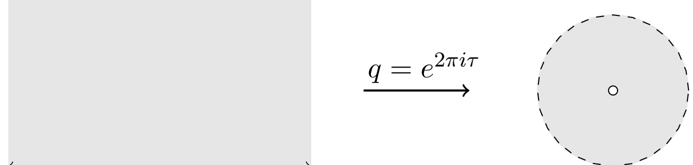
We saw in Example 1.4 that every modular form satisfies $f(\tau+1) = f(\tau)$. The function $e^{2\pi i\tau}$ also satisfies this periodicity relation, and the standard way to write down modular forms is through a power series in $e^{2\pi i\tau}$.

Theorem 4.5. *If $f\colon H \to \mathbf{C}$ is holomorphic, $f(\tau+1) = f(\tau)$ for all τ, and f is bounded as $\tau \to \infty$ then there are $a_n \in \mathbf{C}$ for $n \geq 0$ such that*

$$f(\tau) = \sum_{n\geq 0} a_n e^{2\pi i n\tau}$$

for all $\tau \in \mathfrak{h}$. In particular, $f(\tau)$ has a limit as $\tau \to i\infty$.

Proof. For $\tau \in \mathfrak{h}$ set $q(\tau) = e^{2\pi i\tau}$. Writing $\tau = x + iy$, we have $q(\tau) = e^{-2\pi y}e^{2\pi i x}$, so $|q(\tau)| = e^{-2\pi y} \in (0,1)$. Thus $q(\tau)$ lies in the punctured unit disc $D' = \{q \in \mathbf{C} : 0 < |q| < 1\}$, and conversely each point in D' can be written as $e^{2\pi i\tau}$ for a discrete set of values $\tau \in \mathfrak{h}$. The mapping $\mathfrak{h} \to D'$ given by $q(\tau)$ is surjective and locally invertible: if we write $q_0 \in D'$ as $e^{2\pi i\tau_0}$ then any q sufficiently close to q_0 can be written as $e^{2\pi i\tau}$ for a unique τ near τ_0. This mapping is pictured below. Note $\tau \to i\infty$ in $\mathfrak{h}$ corresponds to $q \to 0$ in D'.



Convert the function $f\colon \mathfrak{h} \to \mathbf{C}$ into a function $\widetilde{f}\colon D' \to \mathbf{C}$ by defining $\widetilde{f}(q) = f(\tau)$ for any $\tau \in \mathfrak{h}$ that makes $e^{2\pi i\tau} = q$. This is well-defined because if $e^{2\pi i\tau'} = q$ then $\tau' = \tau + n$ for some $n \in \mathbf{Z}$, so $f(\tau') = f(\tau+n) = f(\tau)$ due to the relation $f(\tau+1) = f(\tau)$ for all $\tau \in \mathfrak{h}$. Since f is holomorphic, we can prove $\widetilde{f}$ is holomorphic by computing the derivative of $\widetilde{f}$:

for each $q_0 \in D'$, write $q_0 = e^{2\pi i\tau_0}$. Then any q near q_0 is $e^{2\pi i\tau}$ for a unique τ near τ_0, and $q \to q_0$ is equivalent to $\tau \to \tau_0$. Thus

$$\frac{\widetilde{f}(q) - \widetilde{f}(q_0)}{q - q_0} = \frac{f(\tau) - f(\tau_0)}{q - q_0} = \frac{f(\tau) - f(\tau_0)}{\tau - \tau_0}\frac{\tau - \tau_0}{e^{2\pi i\tau} - e^{2\pi i\tau_0}}.$$

As $\tau \to \tau_0$, the right side tends to $f'(\tau_0)/(2\pi i e^{2\pi i\tau_0}) = f'(\tau_0)/(2\pi i q_0)$. (This formula for $f'(q_0)$ is intuitive by the chain rule: $d\widetilde{f}/dq = (df/d\tau)(d\tau/dq) = f'(\tau)(d\tau/d(e^{2\pi i\tau})) = f'(\tau)/(2\pi i q)$.)

The boundedness of $f(\tau)$ as $\tau \to i\infty$ implies boundedness of $\widetilde{f}(q)$ as $q \to 0$. An important theorem in complex analysis, Riemann's removable singularities theorem, says a holomorphic function on a punctured neighborhood $\{z : 0 < |z - a| < r\}$ of a point a that is bounded on a small neighborhood of a (i) has a limit as $z \to a$ and (ii) the extended function set equal to the limit at $z = a$ is holomorphic at a. Therefore the boundedness of $\widetilde{f}(q)$ as $q \to 0$ implies $\widetilde{f}$ is holomorphic at 0. Thus $\widetilde{f}$ has a power series expansion at 0, say $\sum_{n\geq 0} a_n q^n$. Since $\widetilde{f}$ is holomorphic on the whole open unit disc $D = \{q \in \mathbf{C} : |q| < 1\}$, another basic theorem from complex analysis guarantees that $\sum_{n\geq 0} a_n q^n$ converges on all of D: a holomorphic function on an open disc has its series at the center converge on the whole disc. Therefore

$$f(\tau) = \widetilde{f}(e^{2\pi i\tau}) = \sum_{n\geq 0} a_n e^{2\pi i n\tau}$$

for all $\tau \in \mathfrak{h}$. □

Definition 4.6. The q-*expansion* of a modular form $f(\tau)$ is the series $\sum_{n\geq 0} a_n q^n$ for which $f(\tau) = \sum_{n\geq 0} a_n e^{2\pi i n\tau}$. The coefficients a_n in the q-expansion are called the *Fourier coefficients* of f.

A q-expansion is not merely a formal object: the equation $f(\tau) = \sum_{n\geq 0} a_n e^{2\pi i n\tau}$ is analytic on both sides, with the right side convergent for every $\tau \in \mathfrak{h}$. When writing a modular form $f(\tau)$ as its q-expansion, it is a common abuse of notation to write the function as $f(q)$, using the same letter f with the new variable $q = e^{2\pi i\tau}$.

The constant term a_0 in the q-expansion is $f(i\infty)$. While the q-expansion of f encodes the relation $f(\tau + 1) = f(\tau)$, the other relation $f(-1/\tau) = \tau^k f(\tau)$ is not visible at all in terms of q. If we are given a new power series converging on the open unit disc, there is usually no simple way to show if it is the q-expansion of a modular form without further information. The definition of a modular form is awkward to formulate directly in terms of q-expansions.

For the rest of this section we will work out the q-expansion of the Eisenstein series G_k. We already saw in the proof of Theorem 4.4 that the constant term of the q-expansion is $2\sum_{n\geq 1} 1/n^k$. For every complex number s with $\mathrm{Re}(s) > 1$, the *Riemann zeta-function* at s is $\zeta(s) := \sum_{n\geq 1} 1/n^s$. This series is absolutely and uniformly convergent on compact subsets of $\{s : \mathrm{Re}(s) > 1\}$, so $\zeta(s)$ is holomorphic on $\{s : \mathrm{Re}(s) > 1\}$. The constant term of $G_k(\tau)$ is $2\zeta(k)$, and long before Riemann worked with $\zeta(s)$ Euler showed $\zeta(k)$ is a rational multiple of π^k when k is a positive even integer, *e.g.*, $\zeta(2) = \pi^2/6$ and $\zeta(4) = \pi^4/90$.

Theorem 4.7. *For even $k \geq 4$, the q-expansion of $G_k(\tau)$ is*

$$G_k(\tau) = 2\zeta(k) + \frac{2(2\pi i)^k}{(k-1)!}\sum_{n\geq 1}\sigma_{k-1}(n)q^n,$$

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

Proof. Rewrite (4.1) as

$$G_k(\tau) = 2\zeta(k) + 2\sum_{m\geq 1}\left(\sum_{n\in\mathbf{Z}} \varphi_{m\tau}(n)\right), \tag{4.2}$$

where $\varphi_w(x) = 1/(w+x)^k$ for $w \in \mathbf{C} - \mathbf{R}$ and $x \in \mathbf{R}$. On the right side, the inner sum has the form $\sum_{n\in\mathbf{Z}} f(n)$, and there is a beautiful result in Fourier analysis that expresses the sum of one function over $\mathbf{Z}$ as the sum of another function over $\mathbf{Z}$: the Poisson summation formula. It says that if $f\colon \mathbf{R} \to \mathbf{C}$ is a suitably nice function then

$$\sum_{n\in\mathbf{Z}} f(n) = \sum_{n\in\mathbf{Z}} \widehat{f}(n),$$

where $\widehat{f}\colon \mathbf{R} \to \mathbf{C}$ is the *Fourier transform* of f:

$$\widehat{f}(y) = \int_{-\infty}^{\infty} f(x)e^{2\pi i xy}\,dx.$$

If we can justify applying Poisson summation to the inner sum, then

$$G_k(\tau) = 2\zeta(k) + 2\sum_{m\geq 1}\left(\sum_{n\in\mathbf{Z}} \widehat{\varphi_{m\tau}}(n)\right), \tag{4.3}$$

and a calculation of the Fourier transform of $\varphi_{m\tau}$ will produce the desired q-expansion of the Eisenstein series.

Not all functions have Fourier transforms, but since $|e^{2\pi ixy}| = 1$ for all x and y, if f is absolutely integrable on $\mathbf{R}$, meaning $\int_{-\infty}^{\infty} |f(x)|\,dx < \infty$, then the Fourier transform of f is defined. Writing $m\tau = a + bi$, we have

$$|\varphi_{m\tau}(x)| = \frac{1}{|(a+x)+bi|^k} = \frac{1}{((a+x)^2+b^2)^{k/2}},$$

so $\varphi_{m\tau}$ is absolutely integrable for $k \geq 2$. Thus $\widehat{\varphi_{m\tau}}(y)$ makes sense for all $y \in \mathbf{R}$.

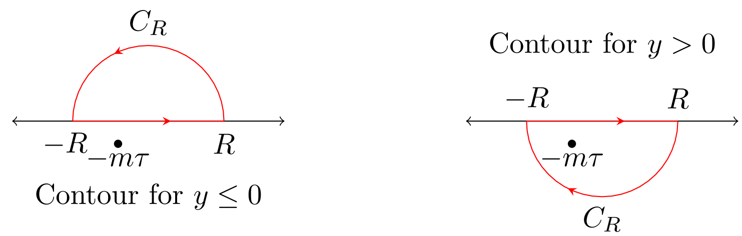
The Poisson summation formula is valid for any function $f\colon \mathbf{R} \to \mathbf{C}$ for which f and its Fourier transform $\widehat{f}$ are both continuous and absolutely integrable on $\mathbf{R}$. Clearly $\varphi_{m\tau}$ is continuous, and we showed it is absolutely integrable. It remains to compute the Fourier transform of $\varphi_{m\tau}$ to check it is continuous and absolutely integrable.

By definition,

$$\widehat{\varphi_{m\tau}}(y) = \int_{\mathbf{R}} \frac{e^{-2\pi ixy}}{(m\tau+x)^k}\,dx = \lim_{R\to\infty}\int_{-R}^{R} \frac{e^{-2\pi ixy}}{(m\tau+x)^k}\,dx. \tag{4.4}$$

We will calculate this integral using the residue theorem from complex analysis. Complexify the integrand to $h(z) = e^{-2\pi izy}/(m\tau+z)^k$ for $z \in \mathbf{C}$. This has a kth order pole at $-m\tau$, which is a point in the lower half-plane since $m \geq 1$ and $\tau \in \mathfrak{h}$.

The numerator $e^{-2\pi izy}$ in $h(z)$ has absolute value $e^{2\pi \operatorname{Im}(z)y}$, so if $y \leq 0$ we want to integrate $h(z)$ along $[-R,R]$ and then counterclockwise along the semicircle in the *upper* half plane with R and $-R$ as endpoints (figure below on the left), since in the upper half-plane $|e^{2\pi\operatorname{Im}(z)y}| \leq 1$. If $y > 0$, we want to integrate $h(z)$ along $[-R,R]$ and then clockwise along the semicircle in the *lower* half-plane connecting R to $-R$ since $|e^{2\pi\operatorname{Im}(z)y}| \leq 1$ on this semicircle.



By the residue theorem the first contour integral is 0 for all $R > 0$, and it is left to the reader to check the integral of $h(z)$ along the semicircle tends to 0 as $R \to \infty$, so $\widehat{\varphi_{m\tau}}(y) = 0$. Check the integral along the semicircle in the second picture also tends to 0 as $R \to \infty$. For R large enough the pole of $h(z)$ is inside the contour, so the residue theorem tells implies

$$\widehat{\varphi_{m\tau}}(y) = -2\pi i \operatorname{Res}_{z=-m\tau} \frac{e^{-2\pi i z y}}{(m\tau + z)^k} = -2\pi i e^{2\pi i m \tau y} \operatorname{Res}_{z=0} \frac{e^{-2\pi i z y}}{z^k}.$$

There is a minus sign out front because we integrate clockwise instead of counterclockwise in order to be integrating in the natural direction along the real axis so that we match (4.4). For any $a \in \mathbf{C}$, $\operatorname{Res}_{z=0}(e^{az}/z^k) = a^{k-1}/(k-1)!$. Our calculation of (4.4) can be summarized as

$$\widehat{\varphi_{m\tau}}(y) = \begin{cases} 0, & \text{if } y \le 0, \\ \frac{(-2\pi i)^k}{(k-1)!} y^{k-1} e^{2\pi i m \tau y}, & \text{if } y > 0. \end{cases} \tag{4.5}$$

Since k is even in our application, we can ignore the minus sign in $(-2\pi i)^k$.

As a function of y (4.5) is continuous[7] and, up to a constant, the integral of $|\widehat{\varphi_{m\tau}}(y)|$ over $\mathbf{R}$ is bounded above by $\int_0^\infty y^{k-1} e^{-2\pi m \operatorname{Im}(\tau) y}\, dy$, which is finite.

It is therefore legal to apply Poisson summation to the inner sum in (4.2):

$$\sum_{n \in \mathbf{Z}} \frac{1}{(m\tau + n)^k} = \sum_{n \in \mathbf{Z}} \widehat{\varphi_{m\tau}}(n) = \frac{(2\pi i)^k}{(k-1)!} \sum_{n \ge 1} n^{k-1} e^{2\pi i m \tau n}. \tag{4.6}$$

Therefore (4.3) becomes

$$\begin{aligned} G_k(\tau) &= 2\zeta(k) + 2 \sum_{m \ge 1} \frac{(2\pi i)^k}{(k-1)!} \sum_{n \ge 1} n^{k-1} e^{2\pi i \tau (mn)} \\ &= 2\zeta(k) + \frac{2(2\pi i)^k}{(k-1)!} \sum_{m \ge 1} \sum_{n \ge 1} n^{k-1} q^{mn}. \end{aligned}$$

Writing mn as N,

$$G_k(\tau) = 2\zeta(k) + \frac{2(2\pi i)^k}{(k-1)!} \sum_{N \ge 1} \left(\sum_{n \mid N} n^{k-1} \right) q^N = 2\zeta(k) + \frac{2(2\pi i)^k}{(k-1)!} \sum_{N \ge 1} \sigma_{k-1}(N) q^N.$$

□

[7]There is a general theorem in Fourier analysis that a function that is absolutely integrable has a Fourier transform that is continuous, so the continuity of $\widehat{\varphi_{m\tau}}$ was predictable.

Remark 4.8. In most treatments of modular forms, the q-expansion of $G_k(\tau)$ is derived not using Poisson summation, but using a more elementary method involving the partial fraction decomposition of $\pi\cot(\pi z)$. We use the technique of Poisson summation since it's good to get familiar with it. We'll use Poisson summation later to construct a special modular form of weight 12.

Euler's formula for $\zeta(k)$ when $k \geq 2$ is even is

$$\zeta(k) = \frac{(2\pi)^k(-1)^{k/2+1}}{k!}\frac{B_k}{2} = -\frac{(2\pi i)^k}{(k-1)!}\frac{B_k}{2k}, \tag{4.7}$$

where B_k is the kth Bernoulli number: it is a rational number appearing in the power series

$$\frac{x}{e^x-1} = \sum_{k\geq 0}\frac{B_k}{k!}x^k = 1 - \frac{1}{2}x + \frac{1}{12}x^2 - \frac{1}{720}x^4 + \cdots$$

The table below lists the first few Bernoulli numbers. The early data suggest $B_k = 0$ for odd $k > 1$, which is true. The early data also suggest $|B_k|$ is small, but $|B_k| \to \infty$ as $k \to \infty$.

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
B_k	1	$-\frac{1}{2}$	$\frac{1}{6}$	0	$-\frac{1}{30}$	0	$\frac{1}{42}$	0	$-\frac{1}{30}$	0	$\frac{5}{66}$	0	$-\frac{691}{2730}$	0	$\frac{7}{6}$

By Theorem 4.7 and (4.7),

$$G_k(\tau) = 2\zeta(k) - \frac{4k\zeta(k)}{B_k}\sum_{n\geq 1}\sigma_{k-1}(n)q^n.$$

For arithmetic applications it is convenient to scale $G_k(\tau)$ so that its constant term is 1.

Definition 4.9. For even $k \geq 4$, define the *normalized Eisenstein series of weight k* to be

$$E_k(\tau) = E_k(q) := \frac{G_k(\tau)}{2\zeta(k)} = 1 - \frac{2k}{B_k}\sum_{n\geq 1}\sigma_{k-1}(n)q^n. \tag{4.8}$$

Using the table of values of Bernoulli numbers, some special cases of (4.8) are

$$\begin{aligned}
E_4(\tau) &= 1 + 240q + 2160q^2 + 6720q^3 + \dots\\
E_6(\tau) &= 1 - 504q - 16632q^2 - 122976q^3 - \dots\\
E_8(\tau) &= 1 + 480q + 61920q^2 + 1050240q^3 + \dots\\
E_{10}(\tau) &= 1 - 264q - 135432q^2 - 5196576q^3 - \dots\\
E_{12}(\tau) &= 1 + \frac{65520}{691}q + \frac{134250480}{691}q^2 + \frac{11606736960}{691}q^3 + \dots\\
E_{14}(\tau) &= 1 - 24q - 196632q^2 - 38263776q^3 - \dots
\end{aligned}$$

Since $2k/B_k \in \mathbf{Z}$ for $k = 4, 6, 8, 10$, and 14, all Fourier coefficients of $E_k(\tau)$ are integers for these k.

The product of modular forms of weight k and ℓ is easily seen to be a modular form of weight $k+\ell$, and we can find the q-expansion of the product by multiplying the q-expansions of the two modular forms. For example,

$$\begin{aligned} E_4(\tau)^2 &= 1 + 480q + 61920q^2 + 1050240q^3 + \dots \text{ has weight } 8, \\ E_4(\tau)E_6(\tau) &= 1 - 264q - 135432q^2 - 5196576q^3 + \dots \text{ has weight } 10, \\ E_4(\tau)^3 &= 1 + 720q + 179280q^2 + 16954560q^3 + \dots \text{ has weight } 12, \\ E_6(\tau)^2 &= 1 - 1008q + 220752q^2 + 16519104q^3 + \dots \text{ has weight } 12. \end{aligned}$$

From the initial parts of q-expansions, it looks like $E_8 = E_4^2$ and $E_{10} = E_4E_6$. In weight 12, the modular forms E_{12}, E_4^3, and E_6^2 are all different and are not scalar multiples of each other since their constant terms all equal 1.

The explanation for identities like $E_8 = E_4^2$ and $E_{10} = E_4E_6$ will come from the fact that the modular forms of a fixed weight are a complex vector space that is *finite-dimensional*, whose proof is the main goal of Section 5.

While the original definition of $G_k(\tau)$ for even $k \geq 4$ makes no sense when $k = 2$, the q-expansion of $G_k(\tau)$ in Theorem 4.7 does make sense!

Definition 4.10. For $\tau \in \mathfrak{h}$, define

$$G_2(\tau) = 2\zeta(2) + \frac{2(2\pi i)^2}{(2-1)!}\sum_{n\geq 1}\sigma_1(n)q^n = \frac{\pi^2}{3} - 8\pi^2\sum_{n\geq 1}\sigma_1(n)q^n$$

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

The series $G_2(\tau)$ converges for all q in the open unit disc on account of the weak bound $\sigma_1(n) = \sum_{d|n} d \leq \sum_{k=1}^n k \sim n^2/2$. It is holomorphic in q (as all convergent power series are in a disc of convergence) and thus also in τ by composition. Trivially $G_2(\tau+1) = G_2(\tau)$ and $G_2(\tau) \to \pi^2/3$ as $\tau \to i\infty$. Could $G_2(-1/\tau) = \tau^2 G_2(\tau)$? No.

Theorem 4.11. *For all $\tau \in \mathfrak{h}$, $G_2(-1/\tau) = \tau^2 G_2(\tau) - 2\pi i\tau$. Equivalently, $E_2(-1/\tau) = \tau^2 G_2(\tau) - (6i/\pi)\tau$.*

Proof. This is a project. □

We will see in Section 5 that the only modular form of weight 2 for $\mathrm{SL}_2(\mathbf{Z})$ is 0.

Exercises.

1. In the proof of Lemma 4.2, for each half-strip $S_{a,b} = \{x+iy : |x| \leq a, y \geq b\}$ in $\mathfrak{h}$, where $a > 0$ and $b > 0$, show there is a $\delta > 0$ such that $|m\tau + n| \geq \delta|mi + n|$ for all $\tau \in S_{a,b}$ and all $m, n \in \mathbf{Z}$. That is, δ in Lemma 4.2 can be chosen uniformly in $S_{a,b}$.
2. For even $k \geq 4$, show
$$G_k(\tau) = \zeta(k)\sum_{(m,n)=1}\frac{1}{(m\tau+n)^k},$$
so $E_k(\tau) = (1/2)\sum_{(m,n)=1}(m\tau+n)^{-k}$.
3. Let M be a positive integer and $k \geq 4$ an even integer. Show
$$\sum_{\substack{(m,n)\in M\mathbf{Z}\times\mathbf{Z} \\ (m,n)\neq(0,0)}}\frac{1}{(m\tau+n)^k} = 2\zeta(k) + 2\frac{(2\pi i)^k}{(k-1)!}\sum_{n\geq 1}\sigma_{k-1}(n)e^{2\pi iMn\tau}.$$

5. Dimensions of Spaces of Modular Forms

Let M_k denote the set of all weight k modular forms for $\mathrm{SL}_2(\mathbf{Z})$. It is a vector space over $\mathbf{C}$. In this section, we show each M_k is finite-dimensional and write down an explicit dimension formula.

The proof will fall into four parts:

(1) Prove $M_k = \{0\}$ for $k < 0$,
(2) Construct a modular form $\Delta(\tau)$ of weight 12 that is nonvanishing on $\mathfrak{h}$.[8]
(3) Use (1) and (2) to compute $\dim M_k$ for $0 \leq k \leq 10$.
(4) Use (2) and (3) to compute $\dim M_k$ for $k \geq 12$.

Theorem 5.1. *If $k < 0$ then $M_k = \{0\}$.*

Proof. Pick $f \in M_k$ and write its q-expansion as $\sum_{n\geq 0} a_n q^n$. We will prove each Fourier coefficient is 0, so $f = 0$.

The modulariity condition for f and the imaginary part formula (2.2) raised to the $k/2$-power say

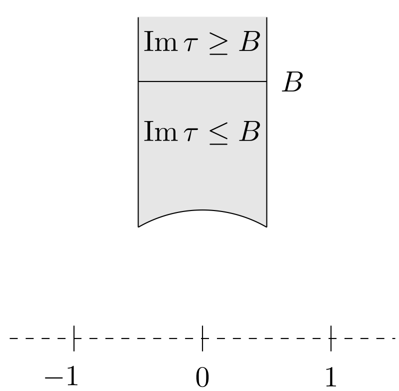
$$f\left(\frac{a\tau+b}{c\tau+d}\right) = (c\tau+d)^k f(\tau), \quad \left(\mathrm{Im}\left(\frac{a\tau+b}{c\tau+d}\right)\right)^{k/2} = \frac{(\mathrm{Im}\,\tau)^{k/2}}{|c\tau+d|^k},$$

for all $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbf{Z})$. Therefore if we take the absolute value of f and multiply,

$$\left|f\left(\frac{a\tau+b}{c\tau+d}\right)\right| \left(\mathrm{Im}\left(\frac{a\tau+b}{c\tau+d}\right)\right)^{k/2} = |c\tau+d|^k |f(\tau)| \frac{(\mathrm{Im}\,\tau)^{k/2}}{|c\tau+d|^k} = |f(\tau)|(\mathrm{Im}\,\tau)^{k/2}.$$

This says the continuous real-valued function $|f(\tau)|(\mathrm{Im}\,\tau)^{k/2}$ on $\mathfrak{h}$ is $\mathrm{SL}_2(\mathbf{Z})$-invariant. (So far we have not used $k < 0$.)

Any $\mathrm{SL}_2(\mathbf{Z})$-invariant function on $\mathfrak{h}$ has all of its values achieved on the fundamental domain $\mathcal{F}$ from Section 3. Break up $\mathcal{F}$ into two parts: that with $\mathrm{Im}\,\tau \leq B$ and that with $\mathrm{Im}\,\tau \geq B$ for B to be determined. See the picture below.



As $\tau \to i\infty$ in $\mathcal{F}$, $|f(\tau)|$ is bounded and $(\mathrm{Im}\,\tau)^{k/2} \to 0$ *because* $k < 0$. Therefore $|f(\tau)|(\mathrm{Im}\,\tau)^{k/2} \to 0$ as $\tau \to i\infty$, so there is some $B > 0$ such that $|f(\tau)|(\mathrm{Im}\,\tau)^{k/2} \leq 1$ for $\mathrm{Im}\,\tau \geq B$. On $\{\tau \in \mathcal{F} : \mathrm{Im}\,\tau \leq B\}$ the function $|f(\tau)|(\mathrm{Im}\,\tau)^{k/2}$ is bounded above since

[8] The minimal positive weight for a modular form that doesn't vanish on $\mathfrak{h}$ is 12 because if $f \in M_k$ and $k \not\equiv 0 \bmod 3$ then $f(\omega) = 0$, and if $k \not\equiv 0 \bmod 4$ then $f(i) = 0$. See Exercise 3.3 for special cases.

a continuous real-valued function on a compact set is bounded. Putting these two parts together, there is some $C > 0$ such that

$$|f(x+iy)|y^{k/2} \leq C \tag{5.1}$$

for all $x + iy \in \mathcal{F}$ and thus also for all $x + iy \in \mathfrak{h}$ by $\mathrm{SL}_2(\mathbf{Z})$-invariance.

Pick $y > 0$. In the q-expansion $f(x+iy) = \sum_{n\geq 0} a_n q^n = \sum_{n \geq 0} a_n e^{-2\pi n y} e^{2\pi i n x}$, multiply both sides by $e^{-2\pi i m x}$ and integrate from 0 to 1:

$$\int_0^1 f(x+iy)e^{-2\pi imx}\,dx = \sum_{n\geq 0} a_n e^{-2\pi ny}\int_0^1 e^{2\pi inx}e^{-2\pi imx}\,dx.$$

(Why is termwise integration of the series justified?) Since $\int_0^1 e^{2\pi inx}e^{-2\pi imx}\,dx$ is 0 for $n \neq m$ and is 1 for $n = m$,

$$\int_0^1 f(x+iy)e^{-2\pi imx}\,dx = a_m e^{-2\pi my},$$

so

$$a_m = e^{2\pi my}\int_0^1 f(x+iy)e^{-2\pi imx}\,dx \overset{(5.1)}{\Longrightarrow} |a_m| \leq e^{2\pi my}\int_0^1 Cy^{-k/2}\,dx = \frac{Ce^{2\pi my}}{y^{k/2}}.$$

This holds for all $y > 0$. Letting $y \to 0^+$, the factor $e^{2\pi my}$ tends to 1 and the factor $y^{k/2}$ tends to ∞ since $k < 0$. Therefore $|a_m| = 0$, so $a_m = 0$ for all m. Thus $f = 0$. $\square$

Theorem 5.2. *There is a modular form $\Delta(\tau) \in M_{12}$ that is nonvanishing on $\mathfrak{h}$ and it has a simple zero at $i\infty$: its q-expansion starts out as $q + b_2 q^2 + \cdots$.*

Using Eisenstein series it is easy to construct a modular form of weight 12 whose q-expansion starts out with q: since $E_4^3 = 1 + 720q + \cdots$ and $E_6^2 = 1 - 1008q + \cdots$, the difference $(E_4^3 - E_6^2)/1728$ has first term q in its q-expansion. What is not easy to see is that this modular form vanishes nowhere on $\mathfrak{h}$. The way we will prove Theorem 5.2 is by building a modular form of weight 12 in a different way. The argument is rather technical (it will use a version of Poisson summation), so for now we will accept Theorem 5.2 as proved and see how to use it to compute the dimensions (and bases) of every M_k for $k \geq 0$. At the end of this section we will return to prove Theorem 5.2.

Theorem 5.3. *For $k = 0, 2, 4, 6, 8, 10$, $\dim M_k$ is given in the following table.*

k	0	2	4	6	8	10
$\dim M_k$	1	0	1	1	1	1

Proof. First we treat the cases $k = 4, 6, 8, 10$. Let $f \in M_k$ and $a_0 = f(i\infty)$. The difference $f(\tau) - a_0 E_k(\tau)$ lies in M_k and its q-expansion has constant term $a_0 - a_0 = 0$.

The ratio $(f - a_0 E_k)/\Delta$ lies in M_{k-12}: it is holomorphic on $\mathfrak{h}$ since $\Delta(\tau) \neq 0$ for all $\tau \in \mathfrak{h}$, it easily satisfies the modularity condition for weight $k - 12$. and as $q \to 0$ the ratio has a finite limit since $f - a_0 E_k$ has a zero at $q = 0$ and Δ has a simple zero at $q = 0$. By Theorem 5.1, $M_{k-12} = \{0\}$ since $k - 12 < 0$, so $f - a_0 E_k = 0$. Thus $f = a_0 E_k$, so $M_k = \mathbf{C}E_k$ is one-dimensional.

If $k = 0$, the constant function 1 lies in M_0 and reasoning as above with 1 in place of E_k shows $f = a_0 \cdot 1 = a_0$, so $M_0 = \mathbf{C}$.

Finally, we will prove $M_2 = 0$. Let $f \in M_2$, so $f(-1/\tau) = \tau^2 f(\tau)$ for all $\tau \in \mathfrak{h}$. Setting $\tau = i$ we get $f(i) = -f(i)$, so $f(i) = 0$. The square f^2 lies in M_4, and we already proved

$M_4 = \mathbf{C}E_4(\tau)$, so $f(\tau)^2 = cE_4(\tau)$ for some $c \in \mathbf{C}$ and all τ. Setting $\tau = i$ on both sides and using the q-expansion of E_4,

$$0 = cE_4(i) = c\left(1 + 240\sum_{n\geq 1}\sigma_3(n)e^{-2\pi n}\right).$$

The sum on the right is positive, so $c = 0$ and thus $f = 0$. □

Theorem 5.4. *Every space M_k is finite-dimensional. For even $k \geq 0$,*

$$\dim M_k = \begin{cases} [k/12]+1, & \text{if } k \not\equiv 2 \bmod 12, \\ [k/12], & \text{if } k \equiv 2 \bmod 12. \end{cases}$$

Proof. We have verified the theorem for $k = 0, 2, 4, 6, 8$, and 10.

For even $k \geq 12$ and $f \in M_k$ with constant term a_0, $(f - a_0E_k)/\Delta$ lies in M_{k-12} by the reasoning used in the proof of Theorem 5.3. Therefore $f = a_0E_k + \Delta g$ where $g \in M_{k-12}$, so the $\mathbf{C}$-linear map $\mathbf{C} \oplus M_{k-12} \to M_k$ given by $(c, g) \mapsto cE_k + \Delta g$ is surjective. To show it is injective we show the kernel is 0: if $cE_k + \Delta g = 0$ in M_k then looking at the constant term of the q-expansion on the left implies $c = 0$, so $\Delta g = 0$, and thus $g = 0$.

Since $\mathbf{C} \oplus M_{k-12} \cong M_k$ as $\mathbf{C}$-vector spaces for $k \geq 12$, M_k is finite-dimensional with dimension $1 + \dim M_{k-12}$. The dimension formula in the theorem satisfies the same recursion, so we are done by induction on k. □

Here is an initial list of the dimensions of M_k for even $k \geq 0$. Note in particular that $\dim M_k = 1$ exactly for $k = 0, 4, 6, 8, 10$, and 14.

k	0	2	4	6	8	10	12	14	16	18	20	22	24	26
$\dim M_k$	1	0	1	1	1	1	2	1	2	2	2	2	3	2

Example 5.5. The equations $E_8 = E_4^2$ and $E_{10} = E_4E_6$ follow from M_8 and M_{10} being one-dimensional; just check the constant terms on both sides agree.

Example 5.6. The space M_{12} has dimension 2, so E_4^3 and E_6^2 must be a basis since they are nonzero and are not scalar multiples (look at the q-expansions).

Since $E_{12} \in M_{12}$, there are complex numbers a and b such that $E_{12} = aE_4^3 + bE_6^2$. We can find a and b by looking at the constant and linear Fourier coefficients on both sides as the first and second components of a vector equation:

$$\begin{pmatrix} 1 \\ 655020/691 \end{pmatrix} = a\begin{pmatrix} 1 \\ 720 \end{pmatrix} + b\begin{pmatrix} 1 \\ -1008 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 720 & -1008 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix}.$$

Using linear algebra, $a = 441/691$ and $b = 250/691$. For example, if we look at the coefficients of q^2 in E_{12}, E_4^3, and E_6^2 then

$$\frac{134250480}{691} = 179280a + 220752b.$$

Since modular forms lie in finite-dimensional spaces but their q-expansions have infinitely many Fourier coefficients, there is some redundancy in the coefficients: knowing a suitable finite list of Fourier coefficients is enough to determine the modular form. The following theorem is one version of this idea.

Theorem 5.7. *For each even $k \geq 0$ there is an $R \geq 0$ such that the first R Fourier coefficients of any weight k modular form for $\mathrm{SL}_2(\mathbf{Z})$ determine the form.*

Proof. Let $L_j\colon M_k \to \mathbf{C}^j$ by sending each modular form to the vector of its first j Fourier coefficients:

$$L_j(f) = (a_0, \ldots, a_{j-1}).$$

The kernels $\ker(L_j)$ are a decreasing sequence of subspaces of M_k: $\ker(L_{j+1}) \subset \ker(L_j)$. Since M_k is finite-dimensional, the kernel subspaces must eventually stabilize, say $\ker(L_R) = \ker(L_j)$ for all $j \geq R$.

This implies $\ker(L_R) = 0$, by contradiction. If this kernel were nonzero, there is a nonzero $f \in M_k$ whose first R coefficients vanish. Some later coefficient is nonzero, say the R'-th coefficient, so $\ker(L_{R'})$ is a proper subspace of $\ker(L_R)$, which contradicts the stabilization. Thus $\ker(L_R) = 0$, so L_R is injective and that means each $f \in M_k$ is determined by its first R Fourier coefficients. □

Clearly R has to be at least as large as the dimension of M_k. It turns out that this minimal choice always works, but that is not obvious and we omit the proof.

Now we return to the proof of Theorem 5.2.

Proof. To construct a weight 12 modular form that is nonvanishing on $\mathfrak{h}$ with a simple zero at $i\infty$, we will use a "twisted" version of Poisson summation. The usual Poisson summation formula says

$$\sum_{n\in\mathbf{Z}} f(n) = \sum_{n\in\mathbf{Z}} \widehat{f}(n)$$

for suitably nice functions $f\colon \mathbf{R} \to \mathbf{C}$ (*e.g.*, f and $\widehat{f}$ are both continuous and absolutely integrable). A twisted version is this:

$$\sum_{\substack{n\in\mathbf{Z}\\ n \text{ odd}}} (-1)^{(n-1)/2} f(n) = \frac{i}{2} \sum_{\substack{n\in\mathbf{Z}\\ n \text{ odd}}} (-1)^{(n-1)/2} \widehat{f}(n/4). \tag{5.2}$$

This formula can be proved using the ordinary Poisson summation formula on a suitable auxiliary function (Exercise 5.5c). We will use (5.2) for the function $f(x) = xe^{-\pi a x^2}$, where $a > 0$. Its Fourier transform is $\widehat{f}(y) = (-iy/a^{3/2})e^{-\pi y^2/a}$ (Exercise 5.5b). Both $f(x)$ and $\widehat{f}(y)$ are continuous and absolutely integrable, which suffices to justify using (5.2). Thus

$$\begin{aligned} \sum_{\substack{n\in\mathbf{Z}\\ n \text{ odd}}} (-1)^{(n-1)/2} n e^{-\pi a n^2} &= \frac{i}{2} \sum_{\substack{n\in\mathbf{Z}\\ n \text{ odd}}} (-1)^{(n-1)/2} \frac{-i(n/4)}{a^{3/2}} e^{-\pi n^2/16a} \\ &= \frac{1}{8a^{3/2}} \sum_{\substack{n\in\mathbf{Z}\\ n \text{ odd}}} (-1)^{(n-1)/2} n e^{-\pi n^2/16a}. \end{aligned}$$

Replacing a with $a/4$ throughout,

$$\sum_{\substack{n\in\mathbf{Z}\\ n \text{ odd}}} (-1)^{(n-1)/2} n e^{-\pi a n^2/4} = \frac{1}{a^{3/2}} \sum_{\substack{n\in\mathbf{Z}\\ n \text{ odd}}} (-1)^{(n-1)/2} n e^{-\pi n^2/4a}.$$

In each sum, the terms at n and $-n$ are equal, so combine the terms and divide by 2:

$$\sum_{\substack{n\geq 1\\ n \text{ odd}}} (-1)^{(n-1)/2} n e^{-\pi a n^2/4} = \frac{1}{a^{3/2}} \sum_{\substack{n\geq 1\\ n \text{ odd}}} (-1)^{(n-1)/2} n e^{-\pi n^2/4a}. \tag{5.3}$$

For $\tau \in H$, define

$$\theta(\tau) = \sum_{\substack{n\geq 1 \\ n \text{ odd}}} (-1)^{(n-1)/2} n e^{\pi i n^2 \tau/4} = e^{\pi i\tau/4} - 3e^{\pi i 9\tau/4} + 5e^{\pi i 25\tau/4} - \cdots .$$

Writing $\tau = x + iy$,

$$\theta(x+iy) = \sum_{\substack{n\geq 1 \\ n \text{ odd}}} (-1)^{(n-1)/2} n e^{-\pi n^2 y/4} e^{\pi i n^2 x/4},$$

which converges very rapidly; it is holomorphic on $\mathfrak{h}$ (as the series converges uniformly on compact subsets of $\mathfrak{h}$) and $\theta(i\infty) = 0$. Along the imaginary axis

$$\begin{aligned}
\theta(iy) &= \sum_{\substack{n\geq 1 \\ n \text{ odd}}} (-1)^{(n-1)/2} n e^{-\pi n^2 y/4} \\
&= \frac{1}{y^{3/2}} \sum_{\substack{n\geq 1 \\ n \text{ odd}}} (-1)^{(n-1)/2} n e^{-\pi n^2/4y} \text{ by (5.3)} \\
&= \frac{1}{y^{3/2}} \theta(i/y) \\
&= \frac{1}{y^{3/2}} \theta\left(-\frac{1}{iy}\right).
\end{aligned}$$

Therefore $\theta(-1/iy) = y^{3/2}\theta(iy)$. Raise both sides to the 8th power:

$$\theta\left(\frac{-1}{iy}\right)^8 = y^{12}\theta(iy)^8 = (iy)^{12}\theta(iy)^8.$$

It follows from this that $\theta(-1/\tau)^8 = \tau^{12}\theta(\tau)^8$ on $\mathfrak{h}$ since both sides are holomorphic and we proved they are equal on the imaginary axis in $\mathfrak{h}$, so they must be equal everywhere.

It is left to the reader to check $\theta(\tau+1) = ((1+i)/\sqrt{2})\theta(\tau)$ (Exercise 5.6). Since $(1+i)/\sqrt{2}$ is an 8th root of unity, $\theta(\tau+1)^8 = \theta(\tau)^8$.

Define

$$\Delta(\tau) = \theta(\tau)^8.$$

Then Δ is holomorphic on $\mathfrak{h}$, $\Delta(\tau+1) = \Delta(\tau)$, and $\Delta(-1/\tau) = \tau^{12}\Delta(\tau)$. Since $\theta(i\infty) = 0$, also $\Delta(i\infty) = 0$, so $\Delta \in M_{12}$.

To prove Δ is nonvanishing on $\mathfrak{h}$, we will prove θ is novanishing on $\mathfrak{h}$. Suppose $\theta(\tau_0) = 0$ for some τ_0. Then $\theta(\gamma\tau_0) = 0$ for all $\gamma \in \mathrm{SL}_2(\mathbf{Z})$, so we may assume $\tau_0 \in \mathcal{F}$ (the fundamental domain for $\mathrm{SL}_2(\mathbf{Z})$). In the equation

$$0 = \theta(\tau_0) = \sum_{\substack{n\geq 1 \\ n \text{ odd}}} (-1)^{(n-1)/2} n e^{\pi i n^2 \tau_0/4}$$

bring the term at $n = 1$ over to the left side and take absolute values:

$$|e^{\pi i\tau_0/4}| \leq \sum_{\substack{n\geq 3 \\ n \text{ odd}}} n|e^{\pi i n^2\tau_0/4}| \tag{5.4}$$

Set $\tau_0 = x_0 + iy_0$, so $y_0 \geq \sqrt{3}/2$ because $\tau_0 \in \mathcal{F}$. Then (5.4) becomes

$$e^{-\pi y_0/4} \leq \sum_{\substack{n\geq 3\\ n \text{ odd}}} ne^{-\pi n^2 y_0/4},$$

so

$$1 \leq \sum_{\substack{n\geq 3\\ n \text{ odd}}} ne^{-\pi(n^2-1)y_0/4} \leq \sum_{\substack{n\geq 3\\ n \text{ odd}}} ne^{-\pi(n^2-1)\sqrt{3}/8}.$$

The sum on the right is rapidly convergent and without caring about error estimates the sum of the first few terms is approximately .013, which is much less than 1, so it appears we have a contradiction.

To make that last step rigorous, we will prove $\sum_{\text{odd } n\geq 3} ne^{-\pi(n^2-1)\sqrt{3}/8} < 1$. Writing odd $n \geq 3$ as $2m+1$ for $m \geq 1$ and doing some algebra,

$$\sum_{\text{odd } n\geq 3} ne^{-\pi(n^2-1)\sqrt{3}/8} = \sum_{m\geq 1}(2m+1)e^{-\pi(m^2+m)\sqrt{3}/2}.$$

Since $m^2 + m \geq 2m$,

$$\sum_{m\geq 1}(2m+1)e^{-\pi(m^2+m)\sqrt{3}/2} \leq \sum_{m\geq 1}(2m+1)e^{-\pi m\sqrt{3}}.$$

This upper bound is a power series in $e^{-\pi\sqrt{3}} \approx .0043$. For $0 < x < 1$, set

$$f(x) = \sum_{m\geq 1}(2m+1)x^m = 2\sum_{m\geq 1} mx^m + \sum_{m\geq 1} x^m = \frac{2x}{(1-x)^2} + \frac{x}{1-x}.$$

This rational function is strictly increasing on $(0,1)$ (its derivative is $(3+x)/(1-x)^3$). The unique number in $(0,1)$ where f has value 1 is $(5-\sqrt{17})/4 \approx .218$, which is greater than $e^{-\pi\sqrt{3}} \approx .0043$, so $f(e^{-\pi\sqrt{3}}) < 1$, and therefore the sum we care about is also less than 1. More precisely, $f(e^{-\pi\sqrt{3}}) \approx .01309$, so the sum we care about is less than .01309. □

Our method of proving finite-dimensionality of the spaces M_k depended in a crucial way on the existence of a modular form that is nonvanishing on $\mathfrak{h}$ with a simple 0 at $i\infty$. The modular forms of positive weight for groups other than $\mathrm{SL}_2(\mathbf{Z})$ do not typically include a form that is nowhere zero on $\mathfrak{h}$, so proving finite-dimensionality of spaces of modular forms in general requires more conceptual arguments, such as using the Riemann-Roch theorem.

Exercises.

1. From $E_8 = E_4^2$ deduce for $n \geq 1$ that $\sigma_7(n) = \sigma_3(n) + 120\sum_{m=1}^{n-1}\sigma_3(m)\sigma_3(n-m)$.
2. Let $f \in M_k$ and $g \in M_\ell$. Show $\ell f'(\tau)g(\tau) - kf(\tau)g'(\tau) \in M_{k+\ell+2}$, where the differentiation is with respect to τ.
3. Show the ratio E_6/E_4 satisfies the modularity condition for weight 2. Why doesn't this contradict $M_2 = \{0\}$?
4. If $f \in M_k$ is nonvanishing on $\mathfrak{h}$ then prove $12 \mid k$ and f equals $\Delta^{k/12}$ up to multiplication by a nonzero constant.
5. (a) Let $f\colon \mathbf{R} \to \mathbf{C}$ be an absolutely integrable function. For $a > 0$ and $b \in \mathbf{R}$, set $f_{a,b}(x) = f(ax+b)$. Prove the Fourier transform of $f_{a,b}$ is

$$\widehat{f_{a,b}}(y) = \frac{e^{2\pi ib/a}}{a}\widehat{f}\left(\frac{y}{a}\right).$$

(b) Prove $xe^{-\pi x^2}$ has Fourier transform $-iye^{\pi y^2}$ and then use part (a) to find the Fourier transform of $xe^{-\pi a x^2}$ for $a > 0$.

(c) Prove (5.2). (Hint: Write $\sum_{\text{odd } n}(-1)^{(n-1)/2}f(n)$ as

$$\sum_{m\in\mathbf{Z}} f(4m+1) - \sum_{m\in\mathbf{Z}} f(4m-1)$$

and apply the usual Poisson summation formula to the functions $f(4x+1)$ and $f(4x-1)$, whose Fourier transforms are described by part (a).)

6. For $\theta(\tau) = \sum_{\text{odd } n\geq 1}(-1)^{(n-1)/2}ne^{\pi i n^2\tau/4}$, show $\theta(\tau+1) = \frac{1+i}{\sqrt{2}}\theta(\tau)$.
7. Use the fact that $f(x) = e^{-\pi a x^2}$ has Fourier transform $\widehat{f}(y) = (1/\sqrt{a})e^{\pi y^2/a}$ to prove that the function $\widetilde{\theta}(\tau) := \sum_{n\in\mathbf{Z}} e^{\pi i n^2\tau} = 1 + \sum_{n\geq 1} 2e^{\pi i n^2\tau}$ satisfies $\widetilde{\theta}(-1/\tau)^4 = -\tau^2\widetilde{\theta}(\tau)^4$.

6. The Eisenstein basis

We computed $\dim M_k$ without writing down a basis (when $k > 12$). In this section we describe an explicit basis built out of Eisenstein series, and more precisely built from E_4 and E_6.

How did Eisenstein series play a role leading up to Theorem 5.4? We used $E_4(i) > 0$ in the proof that $M_2 = \{0\}$, and when we showed $(f - a_0E_k)/\Delta$ is a modular form the only property we needed of E_k is that it lies in M_k and has constant term 1. For every even $k \geq 4$ we can write $k = 4a + 6b$ for some nonnegative integers a and b, so $E_4^aE_6^b$ is in M_k with constant term 1. Therefore we can prove Theorem 5.4 using only the Eisenstein series E_4 and E_6: no E_k for $k > 6$ is needed for the proof.

Theorem 6.1. *For $k \geq 0$, the set $\{E_4^aE_6^b : a, b \geq 0, 4a + 6b = k\}$ is a basis of M_k.*

Proof. We may suppose k is even.

Let N_k be the number of solutions to $4a + 6b = k$ in nonnegative integers a and b. By a direct check, $N_k = \dim M_k$ for $k \leq 12$. Since $N_k = 1 + N_{k-12}$ for $k \geq 12$, $N_k = \dim M_k$ for all k. So the proposed basis $\{E_4^aE_6^b : a, b \geq 0, 4a + 6b = k\}$ has the right size.

To show this set is linearly independent, we may suppose $k \geq 14$. Let

$$\sum_{\substack{4a+6b=k\\ a,b\geq 0}} c_{a,b}E_4(\tau)^aE_6(\tau)^b = 0$$

for all τ. If there is a pure E_4 term, say $c_{A,0}E_4(\tau)^A$, then setting $\tau = i$ shows $c_{A,0}E_4(i)^A = 0$ since $E_6(i) = 0$ (Exercise 3.3). Since $E_4(i) > 0$, $c_{A,0} = 0$. Therefore all nonzero terms in the sum have $b \geq 1$. As E_6 is not identically 0, we may divide by it and get

$$\sum c_{a,b}E_4(\tau)^aE_6(\tau)^{b-1} = 0,$$

a linear relation in weight $k - 6$. By induction the remaining coefficients are 0. □

Definition 6.2. The basis $\{E_4^aE_6^b : 4a + 6b = k\}$ of M_k will be called the *Eisenstein basis*.

The following application of the Eisenstein basis depends on E_4 and E_6 having all rational Fourier coefficients.

Theorem 6.3. *If $k > 0$ and $f \in M_k$ has q-expansion $\sum_{n\geq 0} a_nq^n$ with $a_n \in \mathbf{Q}$ for all $n \geq 1$, then $a_0 \in \mathbf{Q}$.*

Proof. Before we do anything with modular forms, we will prove a result from abstract algebra that describes rational numbers using field automorphisms of the complex numbers.

There are two known field automorphisms of $\mathbf{C}$: the identity and complex conjugation. Many additional field automorphisms of $\mathbf{C}$ exist, since Zorn's lemma (the axiom of choice) can be used to prove for any subfield $F \subset \mathbf{C}$ that any field automorphism of F can be extended (somehow, usually in many ways) to a field automorphism of $\mathbf{C}$. For a proof, see Corollary 4 of http://www.math.uconn.edu/~kconrad/blurbs/zorn2.pdf.

As an example, if $F = \mathbf{Q}(\sqrt{2})$ then the automorphism $a + b\sqrt{2} \mapsto a - b\sqrt{2}$ on F extends (in infinitely many ways in fact) to an automorphism of $\mathbf{C}$. Such an extension is neither the identity nor complex conjugation, since the extension does not fix $\sqrt{2}$ but the identity and complex conjugation both fix $\sqrt{2}$. No automorphism of $\mathbf{C}$ besides the identity or complex conjugation is continuous, and the extra field automorphisms can't be written down using explicit formulas, so their existence really needs Zorn's lemma.

Field automorphisms of $\mathbf{C}$ can tell us whether or not a complex number is rational.

Claim: If $a \in \mathbf{C}$ is not rational then there is a field automorphism $\sigma\colon \mathbf{C} \to \mathbf{C}$ that does not fix a.

Proof of claim: We take cases depending on if a is algebraic or transcendental over $\mathbf{Q}$. If a is algebraic over $\mathbf{Q}$ and $a \notin \mathbf{Q}$, let F be the splitting field of $\mathbf{Q}(a)$ over $\mathbf{Q}$. By Galois theory, there is a field automorphism of F that does not fix a. Any extension σ of this automorphism to $\mathbf{C}$ will not fix a. If a is instead transcendental over $\mathbf{Q}$, let $F = \mathbf{Q}(a)$. Then F is isomorphic to the rational function field $\mathbf{Q}(X)$ for an indeterminate X, so $a \mapsto 1/a$ (or $a \mapsto -a$) defines a field automorphism of F not fixing a. This automorphism of F extends to an automorphism of $\mathbf{C}$ and does not fix a. This concludes the proof of the claim.

Now we turn to the part of the proof that involves modular forms.

Each modular form for $\mathrm{SL}_2(\mathbf{Z})$ is determined by its q-expansion, so we can embed the vector space M_k into the ring of *formal* power series $\mathbf{C}[[q]]$ by thinking about each modular form as its q-expansion

$$\sum_{n\geq 0} a_n q^n, \qquad a_n \in \mathbf{C},$$

viewed purely formally in $\mathbf{C}[[q]]$. For example, the two Eisenstein series E_4 and E_6 are viewed as series in $\mathbf{C}[[q]]$ that both have all coefficients in $\mathbf{Q}$.

For any field automorphism σ of $\mathbf{C}$ we can define a ring automorphism r_σ of $\mathbf{C}[[q]]$ by mapping every formal power series $\sum a_n q^n$ to the formal power series $\sum \sigma(a_n)q^n$. If $f = \sum_{n\geq 1} a_n q^n$ is in M_k, is $r_\sigma(f) = \sum \sigma(a_n)q^n$ the q-expansion of a modular form?

Yes! To prove this, we can assume k is even and at least 4, since otherwise M_k is $\{0\}$ or $\mathbf{C}$. Write f as a $\mathbf{C}$-linear combination of the Eisenstein basis for M_k:

$$f = \sum_{4a+6b=k} c_{ab} E_4^a E_6^b$$

for some complex numbers c_{ab}. Viewing both sides in $\mathbf{C}[[q]]$ and applying r_σ to this equation,

$$r_\sigma(f) = r_\sigma\left(\sum_{4a+6b=k} c_{ab} E_4^a E_6^b\right) = \sum_{4a+6b=k} \sigma(c_{ab}) r_\sigma(E_4)^a r_\sigma(E_6)^b.$$

Since the q-expansion coefficients of E_4 and E_6 are rational, $r_\sigma(E_4) = E_4$ and $r_\sigma(E_6) = E_6$. Thus

$$r_\sigma(f) = \sum_{4a+6b=k} \sigma(c_{ab}) E_4^a E_6^b.$$

This is a **C**-linear combination of the q-expansions of modular forms of weight k, so $r_\sigma(f)$ is the q-expansion of a modular form of weight k (the same weight as f).

Now suppose all the q-expansions coefficients of f are in **Q** except perhaps for its constant term a_0. Then the q-expansion coefficients of f and $r_\sigma(f)$ agree everywhere except possibly in their constant terms, which are a_0 and $\sigma(a_0)$. Since f and $r_\sigma(f)$ are both in M_k, their difference $f - r_\sigma(f)$ is a constant function in M_k. The only constant function of weight $k > 0$ is 0. Therefore $r_\sigma(f) - f = 0$, so $r_\sigma(f) = f$, which implies $\sigma(a_0) = a_0$ for all automorphisms σ of **C**. Thus, by the claim at the start of this proof, $a_0 \in \mathbf{Q}$. □

If $\sum a_n q^n$ is the q-expansion of a modular form and $a_n \in \mathbf{Z}$ for $n \geq 1$, it is generally false that $a_0 \in \mathbf{Z}$. An example is

$$\frac{1}{240}E_4 = \frac{1}{240} + \sum_{n\geq 1} \sigma_3(n)q^n = \frac{1}{240} + q + 9q^2 + 28q^3 + \dots$$

We can now use modular forms to prove a property of the Riemann zeta-function.

Theorem 6.4. *For even $k \geq 8$, $\zeta(k)$ is a rational multiple of π^k.*

Proof. Apply Theorem 6.3 to the Eisenstein series

$$\frac{G_k(\tau)}{2(2\pi i)^k/(k-1)!} = \frac{\zeta(k)}{(2\pi i)^k/(k-1)!} + \sum_{n\geq 1} \sigma_{k-1}(n)q^n,$$

whose q-expansion does not depend on prior knowledge of zeta-values at even integers $k \geq 8$. Since all the higher-degree Fourier coefficients $\sigma_{k-1}(n)$ are rational, the constant term is also rational, so $\zeta(k)/\pi^k$ is rational. □

The proof of Theorem 6.4 depends on Theorem 6.3, whose proof in turns depends on rationality of all the Fourier coefficients of the Eisenstein basis. The rationality of the Fourier coefficients of E_4 and E_6 requires knowing $\zeta(4)/\pi^4$ and $\zeta(6)/\pi^6$ are rational. Therefore our proof using modular forms that $\zeta(k)/\pi^k \in \mathbf{Q}$ for even integers $k \geq 8$ needs this result to be known already for $k = 4$ and $k = 6$ (the case $k = 2$ does not matter). You can check we never relied on the Eisenstein series with weight > 6 for anything but examples, so deducing the rationality of $\zeta(k)/\pi^k$ for even $k \geq 8$ from the cases $k = 4$ and 6 is not a circular argument.

This method of deducing rationality properties of zeta-values from their appearance in the constant term of a modular form can be generalized to zeta-values of all totally real number fields at positive even integers, by constructing modular forms in which the zeta-values appear in the constant term. This is the Klingen–Siegel theorem.

Are there any linear relations between forms of different weights?

Lemma 6.5. *Modular forms with different weights are linearly independent over* **C**.

Proof. Let $f_1, f_2, \dots, f_m$ be nonzero modular forms with respective weights $k_1 < k_2 < \cdots < k_m$. Assume they admit a nontrivial linear relation:

$$\alpha_1 f_1(\tau) + \alpha_2 f_2(\tau) + \cdots + \alpha_m f_m(\tau) = 0 \tag{6.1}$$

for all $\tau \in \mathfrak{h}$, where not all α_j equal 0. We may assume this is an example with $m \geq 2$ minimal, so all α_j are nonzero.

Pick γ in $\mathrm{SL}_2(\mathbf{Z})$ with lower left entry $c \neq 0$ (i.e., $\gamma \neq \pm T^n$ for any $n \in \mathbf{Z}$). Replacing τ with $\gamma\tau$ in (6.1),

$$\alpha_1(c\tau + d)^{k_1} f_1(\tau) + \alpha_2(c\tau + d)^{k_2} f_2(\tau) + \dots \alpha_r(c\tau + d)^{k_m} f_m(\tau) = 0 \tag{6.2}$$

for all τ.

Let $f_j(\tau)$ have q-expansion $\sum_{n\geq 0} a_n^{(j)} e^{2\pi i n\tau}$, so

$$\sum_{n\geq 0}(\alpha_1(c\tau+d)^{k_1}a_n^{(1)}+\cdots+\alpha_r(c\tau+d)^{k_m}a_n^{(m)})e^{2\pi i n\tau}=0.$$

Look at this along the imaginary axis: for $\tau = iy$ with $y > 0$,

$$\sum_{n\geq 0}(\alpha_1(ciy+d)^{k_1}a_n^{(1)}+\cdots+\alpha_m(ciy+d)^{k_m}a_n^{(m)})e^{-2\pi n y}=0. \tag{6.3}$$

For $n > 0$, $y^r e^{-2\pi n y} \to 0$ as $y \to \infty$ for any $r \geq 0$, so if we divide through (6.3) by $e^{-2\pi N y}$ for the smallest N such that some $a_N^{(j)}$ is nonzero and then let $y \to 0$, we are left with

$$\lim_{y\to\infty} \alpha_1(c(iy)+d)^{k_1}a_N^{(1)}+\cdots+\alpha_m(c(iy)+d)^{k_m}a_N^{(m)}=0$$

All α_j are nonzero, some $a_N^{(j)}$ is nonzero, and the weights k_j are distinct, so the left side is the limit of a nonconstant polynomial in y. We have a contradiction. □

Note the proof hardly used the modularity condition for $\mathrm{SL}_2(\mathbf{Z})$; only one $\gamma \neq \pm T^n$ was required.

Since $M_k M_\ell \subset M_{k+\ell}$, the $\mathbf{C}$-linear combinations of modular forms of all weights for $\mathrm{SL}_2(\mathbf{Z})$ is not only a vector space, but a ring containing $M_0 = \mathbf{C}$. (The sum of modular forms of different weights is is not a modular form.) By Theorem 6.1, the forms $E_4^a E_6^b$ for general $a, b \geq 0$ span the $\mathbf{C}$-algebra generated by all modular forms, so the ring generated over $\mathbf{C}$ by modular forms for $\mathrm{SL}_2(\mathbf{Z})$ is $\mathbf{C}[E_4, E_6]$.

Theorem 6.6. *The modular forms E_4 and E_6 are algebraically independent over $\mathbf{C}$.*

Proof. If $P(E_4(\tau), E_6(\tau)) = 0$ for all τ, where $P(X, Y)$ is a polynomial in two variables over $\mathbf{C}$, then Lemma 6.5 reduces us to the case that $P(E_4, E_6)$ is a sum of modular forms of the same weight. Then $P = 0$ by Theorem 6.1. □

Corollary 6.7. *The ring generated over $\mathbf{C}$ by all modular forms for $\mathrm{SL}_2(\mathbf{Z})$ is isomorphic to the polynomial ring $\mathbf{C}[X, Y]$.*

Proof. The ring is $\mathbf{C}[E_4, E_6]$, so the algebraic independence of E_4 and E_6 over $\mathbf{C}$ implies $\mathbf{C}[E_4, E_6] \cong \mathbf{C}[X, Y]$. □

Exercises.

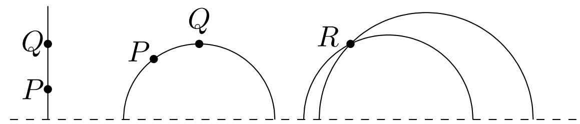
1. (a) Express E_{18} as a linear combinations of E_6^3 and $E_4^3 E_6$.
   (b) Express each of E_{12}^2, $E_6 E_8 E_{10}$, and E_{24} as linear combinations of E_4^6, $E_4^3 E_6^2$, and E_6^4.

Appendix A. The Hyperbolic Plane

The hyperbolic plane is the upper half-plane $\mathfrak{h}$ with a definition of lines (also called geodesics) and distances that differ from the usual meaning of these notions in the Euclidean plane $\mathbf{R}^2$.

Lines in $\mathfrak{h}$ are the vertical lines in $\mathfrak{h}$ or the semicircles in $\mathfrak{h}$ that meet the x-axis in a 90-degree angle (the x-axis is the diameter of the semicircle). In the picture below, if P and Q have the same x-coordinate then the line $\overline{PQ}$ through P and Q is the part of the usual

Euclidean (vertical) line through P and Q that is in $\mathfrak{h}$. If P and Q do not have the same x-coordinate then $\overline{PQ}$ is the unique Euclidean semicircle through P and Q with diameter on the x-axis.



On the right side of the picture two lines drawn through a point R not on $\overline{PQ}$ don't intersect $\overline{PQ}$. This contradicts the parallel postulate of Euclidean geometry, which says a point not on a line L has exactly one line through it that doesn't meet L. In $\mathbf{R}^2$ the parallel postulate is true, but in $\mathfrak{h}$ it is not.

The hyperbolic distance between two points P and Q in $\mathfrak{h}$ is defined using integration along $\overline{PQ}$:

$$d_H(P,Q) = \int_P^Q \frac{\sqrt{(dx/dt)^2 + (dy/dt)^2}}{y(t)}\, dt,$$

where the integral is taken along the hyperbolic line in $\mathfrak{h}$ from P to Q using any smooth parametrization $(x(t), y(t))$ of the segment in $\overline{PQ}$ from P to Q.

Example A.1. To compute the distance between $y_0 i$ and $y_1 i$, parametrize the vertical line between them as $(x(t), y(t)) = (0, (1-t)y_0 + ty_1)$ for $0 \leq t \leq 1$. Then

$$d_H(y_0 i, y_1 i) = \int_0^1 \frac{\sqrt{0^2 + (y_1 - y_0)^2}}{(1-t)y_0 + ty_1}\, dt = |\log y_1 - \log y_0| = |\log(y_1/y_0)|.$$

For example, $d_H(yi, i) = |\log y|$ and the midpoint between $y_0 i$ and $y_1 i$ when $y_0 \neq y_1$ is $\sqrt{y_0 y_1}\, i$, which is (always) different from the Euclidean midpoint.

The action of $\mathrm{SL}_2(\mathbf{R})$ on $\mathfrak{h}$ by linear fractional transformations preserves hyperbolic distances: for each $A \in \mathrm{SL}_2(\mathbf{R})$, $d_H(A(P), A(Q)) = d_H(P,Q)$ for all P and Q in $\mathfrak{h}$. A function $\mathfrak{h} \to \mathfrak{h}$ that preserves distances is called an isometry, and $\mathrm{SL}_2(\mathbf{R})$ acting by linear fractional transformation is the group of all orientation-preserving isometries of the hyperbolic plane.[9] An example of an isometry of $\mathfrak{h}$ that reverses orientation is $\tau \mapsto -\overline{\tau}$, or equivalently $x + yi \mapsto -x + yi$, and every orientation-reversing isometry is this example composed with the action by a matrix in $\mathrm{SL}_2(\mathbf{R})$.

Appendix B. A lattice sum

This section proves a result used in Section 4 to show Eisenstein series of weight $k \geq 4$ are absolutely convergent.

From calculus, the series $\sum_{n \geq 1} 1/n^s$ converges for $s > 1$ and diverges for $0 < s \leq 1$. We will generalize this result to a sum over the d-dimensional integral lattice $\mathbf{Z}^d$ for any $d \geq 1$. For any $\mathbf{x} = (x_1, \ldots, x_d)$ in $\mathbf{R}^d$, set $||\mathbf{x}|| = \sqrt{x_1^2 + \cdots + x_d^2}$. This is the length of $\mathbf{x}$.

Theorem B.1. *For $s > 0$, the infinite series* $\displaystyle\sum_{\mathbf{a} \in \mathbf{Z}^d - \{\mathbf{0}\}} \frac{1}{||\mathbf{a}||^s}$ *converges for $s > d$ and diverges for $0 < s \leq d$.*

[9] Strictly speaking, since A and $-A$ act in the same way, the group of orientation-preserving isometries is $\mathrm{SL}_2(\mathbf{R})/\{\pm I_2\}$.

Proof. We will first collect together all the terms of the same size (that is, all vectors in $\mathbf{Z}^d$ with the same length), and then use an identity called summation by parts, which is a discrete analogue of integration by parts. Then we will rewrite the desired sum as an integral, and our problem will be reduced to the fact that $\int_1^\infty dx/x^t$ converges for $t > 1$ and diverges for $0 < t \leq 1$.

The squared length $||\mathbf{a}||^2$ of any $\mathbf{a} \in \mathbf{Z}^d$ is a positive integer. For $n \geq 1$, set

$$r_d(n) = |\{\mathbf{a} \in \mathbf{Z}^d : ||\mathbf{a}||^2 = n\}|.$$

Then

$$\sum_{\mathbf{a}\in\mathbf{Z}^d-\{\mathbf{0}\}} \frac{1}{||\mathbf{a}||^s} = \sum_{\mathbf{a}\in\mathbf{Z}^d-\{\mathbf{0}\}} \frac{1}{(||\mathbf{a}||^2)^{s/2}} = \sum_{n\geq 1} \frac{r_d(n)}{n^{s/2}} = \lim_{N\to\infty} \sum_{n=1}^{N} \frac{r_d(n)}{n^{s/2}}.$$

For $n \geq 1$ set $S(n) = r_d(1) + \cdots + r_d(n) = |\{\mathbf{a} \in \mathbf{Z}^d : ||\mathbf{a}||^2 \leq n\}|$ and $S(0) = 0$, so $r_d(n) = S(n) - S(n-1)$ for $n \geq 1$. Then

$$\sum_{n=1}^{N} \frac{r_d(n)}{n^{s/2}} = \sum_{n=1}^{N} \frac{S(n) - S(n-1)}{n^{s/2}}.$$

For a sum of the form $\sum_{n=1}^{N} u_n(v_n - v_{n-1})$, which resembles $\int u\,dv$, there is the following analogue of integration by parts, called summation by parts:

$$\sum_{n=1}^{N} u_n(v_n - v_{n-1}) = u_N v_N - u_1 v_0 - \sum_{n=1}^{N-1} v_n(u_{n+1} - u_n).$$

Using $u_n = 1/n^{s/2}$ and $v_n = S(n)$, so $v_0 = 0$, summation by parts implies

$$\text{(B.1)} \qquad \sum_{n=1}^{N} \frac{S(n) - S(n-1)}{n^{s/2}} = \frac{S(N)}{N^{s/2}} - \sum_{n=1}^{N-1} S(n) \left(\frac{1}{(n+1)^{s/2}} - \frac{1}{n^{s/2}} \right).$$

Write the difference $1/(n+1)^{s/2} - 1/n^{s/2}$ as an integral using the Fundamental Theorem of Calculus:

$$\frac{1}{(n+1)^{s/2}} - \frac{1}{n^{s/2}} = \int_n^{n+1} \frac{d}{dx}\left(\frac{1}{x^{s/2}}\right) dx = -\frac{s}{2}\int_n^{n+1} \frac{1}{x^{s/2+1}}\, dx.$$

Substituting this into (B.1),

$$\begin{aligned} \sum_{n=1}^{N} \frac{S(n) - S(n-1)}{n^{s/2}} &= \frac{S(N)}{N^{s/2}} + \frac{s}{2}\sum_{n=1}^{N-1} S(n)\int_n^{n+1} \frac{dx}{x^{s/2+1}} \\ &= \frac{S(N)}{N^{s/2}} + \frac{s}{2}\sum_{n=1}^{N-1} \int_n^{n+1} \frac{S(n)}{x^{s/2+1}}\, dx. \end{aligned}$$

For real $x \geq 1$, which need not be integers, set

$$S(x) = \sum_{1\leq n\leq x} r_d(n) = |\{\mathbf{a} \in \mathbf{Z}^d : 1 \leq ||\mathbf{a}||^2 \leq x\}|,$$

so $S(x) = S(n)$ where $n \le x < n+1$. Then

$$\begin{aligned}\sum_{n=1}^{N} \frac{S(n) - S(n-1)}{n^{s/2}} &= \frac{S(N)}{N^{s/2}} + \frac{s}{2}\sum_{n=1}^{N-1}\int_{n}^{n+1} \frac{S(x)}{x^{s/2+1}}\,dx \\ &= \frac{S(N)}{N^{s/2}} + \frac{s}{2}\int_{1}^{N} \frac{S(x)}{x^{s/2+1}}\,dx.\end{aligned}$$

To determine how $S(N)/N^{s/2}$ and the integral from 1 to N behave as $N \to \infty$, we will estimate $S(x)$ for large x using geometry.

The number $S(x)$ counts nonzero integral points inside the ball $\{\mathbf{x} \in \mathbf{R}^d : ||\mathbf{x}|| \le \sqrt{x}\}$ with radius $\sqrt{x}$, and the number of such integral points is approximately the volume of that ball. A ball of radius r in $\mathbf{R}^d$ has volume $C_d r^d$ for some constant C_d depending only on d (for example, $C_2 = \pi$). Therefore there are positive constants A_d and B_d such that

$$A_d x^{d/2} \le S(x) \le B_d x^{d/2} \tag{B.2}$$

for all large x. Dividing through the inequality (B.2) by $x^{s/2+1}$, we get

$$\frac{A_d x^{(d-s)/2}}{x} \le \frac{S(x)}{x^{s/2+1}} \le \frac{B_d}{x^{(s-d)/2+1}} \tag{B.3}$$

If $0 < s \le d$ then the first inequality in (B.3) tells us $S(x)/x^{s/2+1} \ge A_d/x$ for large x, which implies $\int_1^N S(x)/x^{s/2+1}\,dx \to \infty$ as $N \to \infty$, and thus our original lattice sum diverges.

If $s > d$ then the second inequality in (B.3) tells us $0 \le S(x)/x^{s/2+1} \le B_d/x^{1+\varepsilon}$, where $\varepsilon = (s-d)/2 > 0$, so $\int_1^N S(x)/x^{s/2+1}\,dx$ converges as $N \to \infty$. Using (B.2), $S(N)/N^{s/2} \le B_d/N^{(s-d)/2} \to 0$, so our lattice sum converges and in fact

$$\sum_{\mathbf{a} \in \mathbf{Z}^d - \{\mathbf{0}\}} \frac{1}{||\mathbf{a}||^s} = \frac{s}{2}\int_1^{\infty} \frac{S(x)}{x^{s/2+1}}\,dx.$$

□