MODULAR FORMS AND SUMS OF SQUARES

BY CONNOR HALLECK-DUBÉ

Lagrange proved in 1770 that every natural number can be expressed as a sum of four squares, which naturally begs the enumerative question. Define the following function for counting ways to express $n \geq 0$ as a sum of $k \geq 0$ squares.

$$r(n,k) := \#\{(m_1,\ldots,m_k) \in \mathbb{Z}^k : n = m_1^2 + \cdots + m_k^2\}.$$

This note presents a scenic development of some of the basic theory of modular forms on subgroups of $\mathrm{SL}_2(\mathbb{Z})$, with the aim of proving the following formula.

Theorem 1. *For every $n \geq 1$ we have*

$$r(n,4) = 8 \sum_{\substack{0<d|n \\ 4\nmid d}} d.$$

The assumed prerequisites for this work are somewhat uneven and reflect only the author's comfort. The primary reference for the first two sections is [1]. The third section emphasizes geometry more heavily, following [3] and [5]. The content on modular forms on $\mathrm{SL}_2(\mathbb{Z})$ can also be found in [4]. We mostly follow the notation of [1].

I. Modular Forms on $\mathrm{SL}_2(\mathbb{Z})$

Recall that the modular group $\mathrm{SL}_2(\mathbb{Z})$ naturally acts on the upper half-plane $\mathbb{H}$ by Mobius transformations:

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

Modular forms are holomorphic functions on the half-plane satisfying the following generalization of $\mathrm{SL}_2(\mathbb{Z})$-invariance.

Definition. A meromorphic function $f : \mathbb{H} \to \mathbb{C}$ is called *weakly modular of weight k* if for every $g \in \mathrm{SL}_2(\mathbb{Z})$ as above and $z \in \mathbb{H}$ we have

$$f(g(z)) = (cz+d)^k f(z).$$

In particular, weak modularity of weight 0 is just $\mathrm{SL}_2(\mathbb{Z})$-invariance.

Since $\mathrm{SL}_2(\mathbb{Z})$ is generated by the standard generators

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

weak modularity of weight k is equivalent to the conditions

$$f(z+1) = f(z), \quad f(-1/z) = z^k f(z).$$

For such functions, we can make use of the periodicity in the real direction by applying a logarithm to identify them with meromorphic functions $g : \mathbb{D}' \to \mathbb{C}$ on the punctured disk. For $q = e^{2\pi i z}$, we define

$$g(q) := f(\log(q)/2\pi i).$$

This is well-defined independently of our choice of branch of the logarithm. Since our identification $\mathbb{H} \to \mathbb{D}'$ is locally a biholomorphism, g is meromorphic on the punctured disk.

Definition. We say a weakly modular function f of weight k is a *modular form of weight k* (with respect to $\mathrm{SL}_2(\mathbb{Z})$) if f is holomorphic on $\mathbb{H}$ and is *holomorphic at ∞*, in the sense that g extends to a holomorphic function on the full disk.

That is, there should be a power series expansion for g (the Fourier series of f) around the origin:

$$g(q) = \sum_{n=0}^{\infty} a_n(f) q^n.$$

We denote by $M_k(\mathrm{SL}_2(\mathbb{Z}))$ the graded ring of modular forms of weight k.

To check that a function is holomorphic at ∞ requires only showing that $\lim_{\mathrm{Im}(z)\to\infty} f(z)$ is bounded, because then $f(q)$ is bounded in a neighborhood of the puncture and must therefore extend across it.

Definition. We say that a modular form is a *cusp form* if $a_0 = 0$, or equivalently if $\lim_{\mathrm{Im}(z)\to\infty} f(z) = 0$. We denote by $S_k(\mathrm{SL}_2(\mathbb{Z}))$ the graded ideal of weight-k cusp forms.

I.1. **Eisenstein Series.** The following are our first non-constant examples of modular forms, and play a central role in our argument, so we devote some space to checking their basic properties.

Definition. For $k > 2$ even, the *Eisenstein series of weight k* is defined as the *lattice function* on $\mathbb{H}$ given by

$$G_k(z) = \sum_{(c,d)\neq(0,0)} \frac{1}{(cz+d)^k},$$

where the sum is over pairs $(c,d) \in \mathbb{Z}^2 \setminus \{(0,0)\}$.

Proposition 2. *The series above converges absolutely (and uniformly on compact subsets of $\mathbb{H}$) and defines a modular form of weight k.*

Proof. Consider the absolute value of the series corresponding to $G_k(z)$ and pull out the terms with $c = 0$, which sum to twice the usual Riemann zeta function at k:

$$\sum_{(c,d)\neq(0,0)} \frac{1}{|cz+d|^k} = 2\zeta(k) + \sum_{\substack{(c,d)\neq(0,0)\\ c\neq 0}} \frac{1}{|c|^k \left|z+\frac{d}{c}\right|^k}.$$

We will find appropriate bounds for $|c|$ and for $|z+\delta|$ on the compact "triangular" region

$$\Omega = \{z \in \mathbb{H} : |\operatorname{Re}(z)| \le A, \operatorname{Im}(z) \ge B\}, \quad A, B > 0.$$

First, we will show there is a constant $C > 0$ such that $|z+\delta| > C \sup\{1, |\delta|\}$ for all $z \in \Omega$ and $\delta \in \mathbb{R}$. This is because first, if $|\delta| < 1$ then

$$|z+\delta| \ge |z| - 1 > B \ge B \sup\{1, |\delta|\}.$$

On the other extreme, if $|\delta| > 3A$ then comparing to the real part gives

$$|z+\delta| \ge |\delta| - A \ge \frac{2}{3}|\delta|.$$

In the intermediate region $1 \le |\delta| \le 3A$ (if this is nonempty), if $\operatorname{Im}(z) > A$ then

$$|z+\delta| > A \ge \frac{1}{3}|\delta|,$$

which leaves only a compact region $B \le \operatorname{Im}(z) \le A$. On this final region, the function $|z+\delta|/|\delta|$ achieves some positive minimum value m, so

$$|z+\delta| > C \sup\{1, |\delta|\}$$

for any C less than $\inf\{\frac{1}{3}, B, m\}$. Putting this into our expression for the absolute sum gives

$$\begin{aligned}\sum_{(c,d)\neq(0,0)} \frac{1}{|cz+d|^k} &< 2\zeta(k) + \sum_{\substack{(c,d)\neq(0,0)\\ c\neq 0}} \frac{1}{|c|^k C^k \sup\left\{1, \left|\frac{d}{c}\right|^k\right\}} \\ &\le 2\zeta(k) + \frac{1}{C^k} \sum_{\substack{(c,d)\neq(0,0)\\ c\neq 0}} \frac{1}{\sup\{|c|,|d|\}^k}.\end{aligned}$$

It suffices therefore to show that

$$\sum_{(c,d)\neq(0,0)} \sup(\{|c|,|d|\})^{-k}$$

is bounded. The partial sum over the square $S_l = [-l,l] \times [-l,l]$ can be bounded by splitting it into four triangles:

$$\sum_{(c,d)\in[-l,l]^2} \sup\{|c|,|d|\}^{-k} \le 4 \sum_{\substack{(c,d)\in[-l,l]^2\\ c\ge|d|\ge 0}} |c|^{-k} = 4\sum_{c\ge 1}(2c+1)|c|^{-k},$$

which converges since $k > 2$. We conclude $G_k(z)$ converges absolutely and uniformly on Ω. Sets of the form Ω cover $\mathbb{H}$ so $G_k(z)$ converges absolutely on all of $\mathbb{H}$. It is an easy application of Morera's

theorem that a limit of holomorphic functions which is uniform on compact subsets is holomorphic. The summands are holomorphic, so the partial sums are holomorphic, so G_k is holomorphic on $\mathbb{H}$.

It remains to check that G_k is weakly modular of weight k: for $\gamma = \begin{pmatrix} a & b \\ e & f \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$,

$$G_k(\gamma(z)) = (ez+f)^k \sum_{(c,d)\neq(0,0)} \frac{1}{((ca+de)z+(cb+df))^k}.$$

The coefficients in the denominator are now given by $\gamma^T(c,d)$, and $\gamma^T \in \mathrm{SL}_2(\mathbb{Z})$ induces a bijection on $\mathbb{Z}^2$, so this is a rearrangement of the same sum and we have shown weak modularity.

Finally, the uniformity argument from above implies that G_k is bounded on each Ω, and G_k is $\mathbb{Z}$-periodic since it is weakly modular, so $G_k(z)$ is bounded as $\mathrm{Im}(z) \to \infty$ and thus holomorphic at infinity. □

We also want a concrete description of the Fourier coefficients of $G_k(z)$. We define the k-th divisor counting function by

$$\sigma_k(n) := \sum_{\substack{m|n \\ m>0}} m^k.$$

Then we have the following result.

Proposition 3. *The Fourier expansion of G_k is given by*

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

Proof. Following [1], we first want to compute the Fourier coefficients of

$$f(z) = \sum_{d\in\mathbb{Z}} \frac{1}{(z+d)^k}.$$

We obtain this via a pair of trigonometric identities

$$\sum_{d\in\mathbb{Z}} \frac{1}{z+d} = \frac{1}{z} + \sum_{d=1}^{\infty}\left(\frac{1}{z+d}+\frac{1}{z-d}\right) = \pi\cot\pi x = \pi i - 2\pi i\sum_{n=0}^{\infty} e^{2\pi i z n}.$$

The first equality with the cotangent can be obtained by noting that the functions are both invariant under $z \mapsto z+1$ and agree close to their poles at integer points. But away from the integer points both sides of the equation are bounded, hence their difference is globally bounded and constant. But the two agree at some point, say $z = 1$, so are equal everywhere. The second equality follows from expressing cotangent in terms of the complex exponential.

The composite identity is frequently useful for the computation of Fourier series associated to modular forms, as it allows interchange between sums of lattice functions and power series. In particular, we can here differentiate both sides of

$$\sum_{d\in\mathbb{Z}} \frac{1}{z+d} = \pi i - 2\pi i\sum_{m=0}^{\infty} e^{2\pi i z n}$$

$(k-1)$ times term-by-term to give an expression for the Fourier coefficients of f:

$$\sum_{d\in\mathbb{Z}} \frac{1}{(z+d)^k} = \frac{(-2\pi i)^k}{(k-1)!}\sum_{n=1}^{\infty} n^{k-1}q^n.$$

Now that we have an expression for these 1-dimensional sums, we partition the Eisenstein series into vertical strips and apply the identity to each to obtain (now for $k > 2$ even)

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

Collecting the terms based on exponent of q gives the result. □

Remark. If one normalizes the Eisenstein series by dividing out $2\zeta(k)$ to get a series with leading term 1, the resulting series has rational coefficients because $\zeta(k)/\pi^k \in \mathbb{Q}$ for even positive integers.

Remark. There is an equivalent characterization of modular forms as certain functions on lattices in the complex plane. From this perspective, Eisenstein series are the simplest possible modular forms, so it is sensible for them to be central in the theory.

Remark. We will also need to consider the weight 2 Eisenstein series

$$G_2(z) = \sum_{c\in\mathbb{Z}} \sum_{\substack{d\in\mathbb{Z} \\ (c,d)\neq(0,0)}} \frac{1}{(cz+d)^2}.$$

This does NOT converge absolutely, but does converge conditionally when ordered this way. We will consider in the next section modifications of this function that are modular forms.

II. Congruence Subgroups

Applications in both geometry and number theory require the relaxation of the modularity condition to certain finite-index subgroups of $\mathrm{SL}_2(\mathbb{Z})$.

Definition. The *principal congruence subgroup of level* N, $\Gamma(N) \trianglelefteq \mathrm{SL}_2(\mathbb{Z})$, is the kernel of the entry-wise reduction homomorphism $\pi : \mathrm{SL}_2(\mathbb{Z}) \to \mathrm{SL}_2(\mathbb{Z}/n\mathbb{Z})$. We call a subgroup of $\mathrm{SL}_2(\mathbb{Z})$ a *congruence subgroup of level* N if it contains $\Gamma(N)$.

There are two other classes of congruence subgroups we will need.

Definition. Let π the same mod N reduction map as above. We define the *Hecke congruence subgroup* $\Gamma_0(N)$ as the preimage of the upper triangular matrices:

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

Similarly define $\Gamma_1(N)$ as the preimage of the unipotent matrices:

$$\Gamma_1(N) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : c \equiv 0,\ a, d \equiv 1 \mod N \right\}.$$

We have a natural filtration $\Gamma(N) \trianglelefteq \Gamma_1(N) \trianglelefteq \Gamma_0(N)$.

We can naturally extend our definition of weak modularity to congruence subgroups.

Definition. For notational convenience, define the *weight k operator*

$$(f[\gamma]_k)(z) := (cz+d)^{-k} f(\gamma(z)).$$

A meromorphic function $f : \mathbb{H} \to \mathbb{C}$ is *weakly modular of weight k with respect to* Γ if

$$f[\gamma]_k = f$$

for all $\gamma \in \Gamma$.

It takes a bit of work to organically determine the condition corresponding to "holomorphic at infinity." As motivation, we reconsider the definition of a modular form for $\mathrm{SL}_2(\mathbb{Z})$ above more geometrically. The quotient $Y := \mathrm{SL}_2(\mathbb{Z}) \setminus \mathbb{H}$ is very nearly a compact Riemann surface (and indeed, a complex algebraic curve) except that all the fundamental domains are missing a single point on the boundary, either the point at infinity or in $\mathbb{Q} \subset \mathbb{R}$. In order to make this compactification natural, we want to extend $\mathbb{H}$ to include a point at infinity. But we want the extended space to still have a natural action by $\mathrm{SL}_2(\mathbb{Z})$, so we also need to include the $\mathrm{SL}_2(\mathbb{Z})$-orbit of ∞ along the boundary of $\mathbb{H}$. This action includes inversion, translation by integers, and scaling by rational numbers, so we have to include all of

$$\mathbb{H}^\star = \mathbb{H} \cup \{\infty\} \cup \mathbb{Q}.$$

Since the $\mathrm{SL}_2(\mathbb{Z})$ action is transitive, the quotient $X := \mathrm{SL}_2(\mathbb{Z}) \setminus \mathbb{H}^\star$ is naturally a compact Riemann surface given by the one-point compactification of Y. We call the additional point the *cusp* of X.

Remark. It would be reasonable to wonder what the space X has to do with modular forms. After all, modular forms of weight $k \neq 0$ are not well-defined functions on X. However, the *quotient* of two such does. There is a close analogy to the algebraic geometry of projective space, where one studies homogeneous polynomials (which are not well-defined as functions) because the rational functions are given by quotients of two homogeneous polynomials of the same degree.

Given a congruence subgroup $G \subset \mathrm{SL}_2(\mathbb{Z})$ of level N, the boundary $\mathbb{Q} \cup \{\infty\}$ may consist of multiple $\mathrm{SL}_2(\mathbb{Z})$ orbits, though the containment $\Gamma(N) \subset G$ ensures there are only finitely many such. We can still build Riemann surfaces X and Y as above, with the caveat that now the compactification may add multiple cusps, one for each orbit of the action of $\Gamma(N)$ on $\mathbb{Q} \cup \{\infty\} = \mathbb{P}^1(\mathbb{Q})$.

$$Y(\Gamma) := \Gamma \setminus \mathbb{H}$$
$$X(\Gamma) := \Gamma \setminus \mathbb{H}^\star$$

Modular forms (at last!) are defined in such a way as to make them objects (though not functions) living on $X(\Gamma)$. First notice that if a function is weight k invariant with respect to a congruence subgroup, then it satisfies $f(z+N) = f(z)$ for some N, and so we still have a well-defined map to the punctured disk via $z \mapsto q = e^{2\pi i z/N}$ and we can still make sense of the notion of being holomorphic at infinity.

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

(1) f is holomorphic,
(2) f is weakly modular of weight k with respect to Γ, and
(3) for all $\gamma \in \mathrm{SL}_2(\mathbb{Z})$, the function $f[\gamma]_k$ is holomorphic at infinity.

If in addition all the Fourier expansions of $f[\gamma]_k$ for all $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ have $a_0 = 0$, then we say f is a *cusp form of weight k with respect to* Γ. We denote the set of k-modular forms $M_k(\Gamma)$ and the k-cusp forms $S_k(\Gamma)$. Again these combine to give a graded ring $M(\Gamma)$ and a homogeneous ideal $S(\Gamma)$.

The third condition is the natural way to ask the function to be holomorphic at every point of $\mathbb{P}^1(\mathbb{Q})$, i.e. at all of the cusps. The weight k operator $[\gamma]_k$ should be thought of as performing a change of coordinates to move a specific boundary point to ∞, which we can then test for a holomorphic extension in the punctured disk.

Remark. It is clear from the discussion above that it suffices to check the holomorphic cusp and vanishing cusp conditions only for one boundary point in each orbit, or for γ a set of coset representatives of Γ in $\mathrm{SL}_2(\mathbb{Z})$. We also have the following proposition.

Proposition 4. *Let $\Gamma \supset \Gamma(N)$ a congruence subgroup. Assume f is holomorphic and weight k invariant. If in the Fourier expansion*

$$f(z) = \sum_{n=0} a_n q^{n/N}$$

the coefficients are bounded by a polynomial: $|a_n| \leq Cn^r$, then f is holomorphic at the cusps.

Proof omitted. □

We consider now modifications of the conditionally convergent Eisenstein series G_2 which give weight 2 modular forms on the subgroup $\Gamma_0(N)$.

Proposition 5. *For each positive integer N, define the modified Eisenstein series*

$$G_{2,N}(z) := G_2(z) - NG_2(Nz).$$

Then $G_{2,N} \in M_2(\Gamma_0(N))$.

Proof Sketch. Despite converging only conditionally, the order of terms chosen in the definition of G_2 are such that the Fourier series computation for G_k in the previous section remains valid, i.e.

$$G_2(z) = 2\zeta(2) - 8\pi^2 \sum_{n=1}^{\infty} \sigma(n) q^n.$$

One works carefully with the conditionally convergent series for G_2 to check that the steps performed in our Fourier series derivation are valid.

Then one checks that for all $\gamma \in \mathrm{SL}_2(\mathbb{Z})$,

$$(G_2[\gamma]_2)(z) = G_2(z) - \frac{2\pi i c}{cz+d}.$$

One checks this for the standard generators, and then checks that the property is preserved under products and inverses.

With this in hand, we can prove weak modularity of $G_{2,N}$. Then for any $\gamma = \begin{pmatrix} a & b \\ Nc & d \end{pmatrix} \in \Gamma_0(N)$, notice that

$$N\gamma(z) = \gamma'(Nz), \qquad \gamma' = \begin{pmatrix} a & Nb \\ c & d \end{pmatrix}.$$

Applying this we can compute:

$$\begin{aligned}
(G_{2,N}[\gamma]_2)(z) &= (Ncz+d)^{-2}\left(G_2(\gamma(z)) - NG_2(N\gamma(z))\right) \\
&= G_2(z) - \frac{2\pi icN}{cNz+d} - N\left((c(Nz)+d)^{-2}G_2(\gamma'(Nz))\right) \\
&= G_2(z) - \frac{2\pi icN}{cNz+d} - N\left(G_2(Nz) - \frac{2\pi ic}{cNz+d}\right) \\
&= G_2(z) - NG_2(Nz).
\end{aligned}$$

So $G_{2,N}$ is weakly modular of weight k with respect to $\Gamma_0(N)$. They are holomorphic on $\mathbb{H}$ since G_2 is. That they are holomorphic at infinity will be a corollary of the previous proposition and the Fourier series we compute below, whose terms are bounded by

$$a_n \leq 8\pi^2\sigma(n) \leq 8\pi^2 n^2.$$

□

Proposition 6. *The Fourier series for the modified Eisenstein polynomials $G_{2,2}$ and $G_{2,4}$ are given by*

$$G_{2,2}(z) = -\frac{\pi^2}{3}\left(1 + 24\sum_{n=1}^{\infty}\left(\sum_{\substack{d|n \\ 2\nmid d}} d\right)q^n\right),$$

$$G_{2,4}(z) = -\pi^2\left(1 + 8\sum_{n=1}^{\infty}\left(\sum_{\substack{d|n \\ 4\nmid d}} d\right)q^n\right).$$

Proof. Applying the Fourier series for G_2 gives (for $\sigma(n) = \sigma_1(n) = \sum_{d|n} d$) the expression

$$\begin{aligned}
G_{2,2}(z) &= 2\zeta(2) - 8\pi^2\sum_{n=1}^{\infty}\sigma(n)q^n - 2\left(2\zeta(2) - 8\pi^2\sum_{n=1}^{\infty}\sigma(n)q^{2n}\right) \\
&= -2\zeta(2) - 8\pi^2\sum_{n\geq 1 \text{ odd}}\sigma(n)q^n - 8\pi^2\sum_{n\geq 1 \text{ even}}\left(\sigma(n) - 2\sigma\left(\frac{n}{2}\right)\right)q^n.
\end{aligned}$$

But for even n division by 2 induces a bijection between even divisors of n and all divisors of $n/2$, which gives

$$\sum_{\substack{d|n \\ 2\nmid d}} d = \sigma(n) - 2\sigma\left(\frac{n}{2}\right)$$

so pulling out $-2\zeta(2) = -\frac{\pi^2}{3}$ gives the result. The case of $G_{2,4}$ is similar. □

III. Differential forms and Dimension Formulae

This section mostly follows [3], and assumes a background in basic algebraic geometry. An important milestone in the theory is to compute the dimension of the space of modular forms $M_k(\Gamma)$. We will do this by interpreting modular forms as meromorphic differential forms on $X(\Gamma)$ twisted by an appropriate divisor and then applying Riemann-Roch. We work in substantially more generality than we will need for our application.

Let Γ a congruence subgroup and $\Omega^1_{X(\Gamma)}$ denote the sheaf of holomorphic differentials on $X(\Gamma)$. Notice that for $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$ we have

$$\frac{d(\gamma(z))}{dz} = \frac{d}{dz}\left(\frac{az+b}{cz+d}\right) = \frac{1}{(cz+d)^2},$$

so modular forms f of weight 2 give rise to Γ-invariant differentials $f(z)dz$ on the plane, which naturally descend to differentials on $Y(\Gamma)$. The condition at infinity implies that since $dz = \frac{1}{2\pi i}\frac{dq}{q}$,

$$f(z)dz = \left(\frac{1}{2\pi i}\sum_{n\geq 0} a_n(f)q^n\right)\frac{dq}{q}.$$

So modular forms of weight 2 do not necessarily give rise to differential 1-forms on all of $X = X(\Gamma)$, but can have simple poles at each of the cusps. Weight 2 *cusp* forms, correspondingly, cannot have poles at the cusps of X.

We might hope based on the above discussion that twisting by the divisor D on X which is the sum of the cusps is the correct thing to do, i.e. that

$$M_2(X) \stackrel{?}{=} H^0(X, \Omega^1_X(D))$$

and more generally that

$$M_{2k}(X) \stackrel{?}{=} H^0(X, \Omega_X^{\otimes k}(kD))$$

by identifying a weight $2k$ modular form with the differential k-form $f(z)(dz)^k$ on the plane. This is true if Γ is, for example, torsion free, but is complicated in general by the existence of certain degenerate points on $X(\Gamma)$ called *elliptic points* on X. These are points with nontrivial stabilizer subgroups, in the following sense.

Definition. For $z \in \mathbb{H}$, let $\Gamma_z := \{\gamma \in \Gamma : \gamma(z) = z\}$ denote the stabilizer subgroup. We say that a point $z \in \mathbb{H}$ is *elliptic for* γ if $\{\pm I\}\Gamma_z \supsetneq \{\pm I\}$.

Notice that if $-I \in \Gamma$ then it acts trivially on all of $\mathbb{H}$. The definition is simply a convenient way of excluding this. Equivalently, if $-I \in \Gamma$, then an elliptic point is one with $|\Gamma_z| > 2$, and if $-I \notin \Gamma$, then the condition is $|\Gamma_z| > 1$.

Example. For $\Gamma = \Gamma(1) = \mathrm{SL}_2(\mathbb{Z})$, the only elliptic points are i, ζ_6, ζ_6^2 for ζ_6 the primitive 6th root of unity in $\mathbb{H}$. The latter two are equivalent, so $X(1)$ has two elliptic points.

For a congruence subgroup $\Gamma \subset \mathrm{SL}_2(\mathbb{Z})$, the elliptic points on $X(\Gamma)$ are all contained in the image of the orbits $\mathrm{SL}_2(\mathbb{Z})i$ or $\mathrm{SL}_2(\mathbb{Z})\zeta_6$. The action of Γ on these orbits may not be transitive, but there are at most as many as cosets of Γ in $\mathrm{SL}_2(\mathbb{Z})$, so finitely many.

Remark. Consider the point i in $\mathbb{H}$ under the natural map to $X(1)$. We can see that a small disk around i consists of two copies of the standard fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$. As a result, the complex structure on $X(1)$ is chosen so that the quotient $\mathbb{H}^\star \to X(1)$ is locally isomorphic to the standard ramified double cover $z \mapsto z^2$. We say this elliptic point has *ramification index 2.*[1] Similarly, the quotient map above ζ_6 looks like a ramified triple cover, so we say it is an elliptic point of *ramification index 3.* Every elliptic point on any $X(\Gamma)$ is equivalent to one of these two. We denote e_P for the ramification index of an elliptic point P.

Pulling back functions and differential forms along ramified covers distorts the orders of vanishing at ramification points, as the following result shows.

Proposition 7. *Let $\varphi : D \to D$ be the map of the disk given by $z \mapsto z^e$. If ω is a meromorphic differential k-form on D and $\omega^\star$ is its pullback along φ, then*

$$\mathrm{ord}_0(\omega^\star) = e\,\mathrm{ord}_0(\omega) + k(e-1).$$

Proof. First consider the case $\omega = f$ is a meromorphic function with $m = \mathrm{ord}_0(f)$. The Laurent series expansion for f around 0 is

$$az^m + bz^{m+1} + \cdots,$$

so the corresponding expansion for $f^\star$ is

$$a(z^e)^m + \cdots$$

and $f^\star$ vanishes to order em. For $\omega = f(z)(dz)^k$ a general differential k-form on D, we have

$$\omega^\star = f(z^e)(dz^e)^k = f(z^e)\left(ez^{e-1}dz\right)^k = f(z^e)e^k z^{k(e-1)}(dz)^k,$$

which gives the formula above. □

[1] This is also the cardinality of the image of the stabilizer in $\mathrm{SL}_2(\mathbb{Z})/\{\pm I\} \cong \mathrm{PSL}_2(\mathbb{Z})$.

At points which are neither elliptic points nor cusps, orders of vanishing are preserved. Combining these observations gives the following result.

Lemma 8. *Let ω_0 be an arbitrary k-fold differential (a choice of basis), so that $\operatorname{div}(\omega_0) = kK$ for K a canonical divisor on $X(\Gamma)$. Then we can identify k-fold differentials $\omega = h\omega_0$ (i.e. $M_{2k}(\Gamma)$) with meromorphic functions h satisfying vanishing conditions at elliptic points and cusps. Concretely, if we denote $[\cdot]$ for the function that rounds to integers towards zero and define the divisor*

$$D := \sum_{P_i \text{ cusps}} kP_i + \sum_{Q_i \text{ elliptic}} \left[k(1 - 1/e_{Q_i})\right],$$

then we have

$$M_{2k}(\Gamma) = H^0(X(\Gamma), \Omega^{\otimes k}_{X(\Gamma)}(D)) = H^0(X(\Gamma), L(D + \operatorname{div}(\omega_0)))$$

for the usual Riemann-Roch sheaf $L(D)$.

Now we can apply Riemann-Roch to give a satisfying geometric formula for the dimension of the spaces of modular and cusp forms.

Theorem 9. *For Γ a congruence subgroup, we have*

$$\dim(M_{2k}(\Gamma)) = \begin{cases} 0 & \text{if } k \leq -1 \\ 1 & \text{if } k = 0 \\ (2k-1)(g-1) + kv_\infty + \sum_{Q_i \text{ elliptic}} \left[k\left(1 - \frac{1}{e_{Q_i}}\right)\right] & \text{if } k \geq 1 \end{cases}$$

where

(1) g is the genus of $X(\Gamma)$,
(2) v_∞ is the number of cusps on $X(\Gamma)$,
(3) the sum is taken over all elliptic points Q_i.

Proof. We prove only the case of $k \geq 1$. The canonical divisor K has degree $2g-2$, and so any k-fold differential form has degree $k(2g-2)$. Hence $D + \operatorname{div}(\omega_0) = D + kK$ has degree more than $2g-2$ and the correction term in Riemann-Roch vanishes. We conclude

$$\begin{aligned} \dim(M_{2k}(\Gamma)) &= 1 - g + \deg(D + \operatorname{div}(\omega_0)) \\ &= 1 - g + k(2g-2) + kv_\infty + \sum_{Q_i \text{ elliptic}} \left[k\left(1 - \frac{1}{e_{Q_i}}\right)\right] \end{aligned}$$

as desired. □

In order for this to be useful, we need to be able to compute the genus and ramification indices practically.

Proposition 10. *The genus of $X(\Gamma)$ is given by the formula*

$$g = 1 + \frac{m}{12} - \frac{v_2}{4} - \frac{v_3}{3} - \frac{v_\infty}{2},$$

where v_2, v_3 are the number of elliptic points of order 2,3 respectively, and v_∞ is the number of cusps.

Proof Sketch. Riemann-Hurwitz gives an equation relating genera under a ramified covering. There is such a ramified covering $X(\Gamma) \to X(1) = X(\mathrm{SL}_2(\mathbb{Z}))$, and since $X(1)$ is topologically a sphere it becomes

$$g = 1 - m + \sum_{Q_i \text{ elliptic}} (e_{Q_i} - 1)/2.$$

Doing some analysis of the ramification above each type of elliptic point, we can tally the formula above. □

We have reduced determining the dimension to essentially group-theoretic data, which is tedious but viable to compute. We suppress the details and demonstrate the techniques on $\Gamma_0(N)$.

Proposition 11. *The period 2 (resp. period 3) elliptic points of $\Gamma_0(N)$ are in bijective correspondence with the ideals of $\mathbb{Z}[i]$ (resp. $\mathbb{Z}[\zeta_6]$) such that as groups $\mathbb{Z}[i]/J \cong \mathbb{Z}/N\mathbb{Z}$ (resp. $\mathbb{Z}[\zeta_6]/J \cong \mathbb{Z}/N\mathbb{Z}$).*

Proof omitted. See [1, 3.7 and 3.8], or [5]. □

Corollary 12. *The congruence subgroup $\Gamma_0(4)$ has no elliptic points.*

Proof. We make use of the splitting behavior of primes in $\mathbb{Z}[i]$ and $\mathbb{Z}[\zeta_6]$. For the former, we can only obtain a 2-group quotient if we quotient by some power of the ramified prime $(1+i)$ over (2). There are two cases depending on whether the power is odd or even, but neither gives a cyclic quotient. For $\mathbb{Z}[\zeta_6]$, we have to look over the prime 2, which is inert, so quotients by $(2)^e$ are group-isomorphic to $(\mathbb{Z}/2^e\mathbb{Z})^2$. This will never give $\mathbb{Z}/4\mathbb{Z}$ either. □

Proposition 13. *Let Γ_s be the stabilizer subgroup of any cusp s. Examples include the stabilizer of infinity (the translations) or of zero (the transposes of the translations). Then the cusps of $X(\Gamma)$ are in natural bijection with the double coset*

$$\Gamma \setminus \mathrm{SL}_2(\mathbb{Z})/\Gamma_s.$$

Proof. Inside $\mathrm{PSL}_2(\mathbb{Z}) = \mathrm{SL}_2(\mathbb{Z})/\{\pm I\}$, the image $\overline{\Gamma_s}$ is still the stabilizer subgroup of s. So by orbit-stabilizer we have an isomorphism of Γ-sets

$$\mathrm{SL}_2(\mathbb{Z})/\Gamma_s \cong \mathrm{PSL}_2(\mathbb{Z})/\overline{\Gamma_s} \cong (\mathbb{Q} \cup \{\infty\}).$$

So the double coset is in bijection with the quotient of $\mathbb{P}^1(\mathbb{Q})$ by Γ, i.e. the set of cusps. □

This double quotient is easy to count if Γ is normal, but can be tricky in general. Generally it is computed in examples by working explicitly with the action on $\mathbb{P}^1(\mathbb{Q})$, as in the example here.

Corollary 14. *The number of cusps of $\Gamma_0(N)$ is given by*

$$v_\infty = \sum_{0<d|N} \varphi(\gcd(d, N/d)).$$

Proof. Take $s = 0$ above, so that $\Gamma_s = \{\pm \begin{pmatrix} 1 & 0 \\ m & 1 \end{pmatrix}$. We make use of a common trick for generating elements of $\mathrm{SL}_2(\mathbb{Z})$. For each pair $\{c, d\}$ of positive integers with $\gcd(c, d) = 1$, $d|N$, and $0 < c \leq N/d$, we choose a, b such that $ad - bc = 1$. We claim that the corresponding elements $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ form a full set of representatives for $\Gamma_0(N) \setminus \mathrm{SL}_2(\mathbb{Z})$.

Certainly the pairs must give distinct representatives because if any two pairs were equivalent under the action of $\Gamma_0(N)$, it would imply that some pair was equivalent to the pair $\{1, N\}$. Taking that pair and left-multiplying by an element of $\Gamma_0(N)$ gives

$$\begin{pmatrix} \star & \star \\ Nn' & m' \end{pmatrix} \begin{pmatrix} a & b \\ 1 & N \end{pmatrix} = \begin{pmatrix} \star & \star \\ Nm'a + m' & Nn'b + m'N \end{pmatrix}.$$

For this to be another pair we would need $m' \equiv 0 \mod N$, but then the elements of the pair are not relatively prime, a contradiction. So the pairs give distinct representatives of the quotient. One computes that the number of pairs is

$$N \prod_{p|N} (1 + p^{-1}),$$

which is also the index of $\Gamma_0(N)$ in $\mathrm{SL}_2(\mathbb{Z})$, so we have found them all.

Now the double coset is the set of orbits of these pairs $\{c, d\}$ modulo the equivalence

$$\begin{pmatrix} \star & \star \\ c' & d' \end{pmatrix} = \begin{pmatrix} \star & \star \\ c & d \end{pmatrix} \begin{pmatrix} 1 & 0 \\ m & 1 \end{pmatrix} = \begin{pmatrix} \star & \star \\ c + dm & d \end{pmatrix}.$$

Fixing d, therefore, the associated c are identified if they are equal up to some multiple of d, so we can by Bezout's identity take representatives $0 < c \leq \gcd(d, N/d)$ and relatively prime to both. Hence the number of such is $\varphi(\gcd(d, N/d))$ and summing over all $d|N$ gives the formula. □

Example. For $\Gamma_0(4)$, we have $v_2 = v_3 = 0$ and $v_\infty = 1 + 1 + 1$. The index is

$$m = 4 \cdot \frac{3}{2} = 6.$$

So the genus is

$$g = 1 + \frac{1}{2} - 0 - 0 - \frac{3}{2} = 0,$$

and

$$\dim(M_{2k}(\Gamma_0(4))) = 1 - 2k + 3k + 0 = k + 1.$$

IV. Counting Sums of Four Squares

Recall the function $r(n,k)$ defined in the introduction. Notice that we distinguish order of terms and allow the m_i to be negative. For example, we have $r(1,4) = 8, r(2,4) = 24$. This function satisfies the following natural recursion: for fixed $k_1 + k_2 = k$, we have

$$r(n,k) := \sum_{n_1+n_2=n} r(n_1,k_1)r(n_2,k_2).$$

Motivated by the fact that this looks like the convolution rule for multiplying power series, we construct a complex-valued generating function on the upper half-plane,

$$\theta(z,k) = \sum_{n=0}^{\infty} r(n,k)e^{2\pi izn} = \sum_n r(n,k)q^n.$$

Proposition 15. *For each $k \geq 0$ the series $\theta(z,k)$ converges absolutely and defines a holomorphic function on $\mathbb{H}$.*

Proof. We first re-index the series slightly, as

$$\theta(z,k) = \sum_{\mathbf{v}\in\mathbb{Z}^k} e^{2\pi i|\mathbf{v}|^2 z}.$$

We will show this converges absolutely and uniformly to a holomorphic function on $\mathbb{H}$.[2]

Consider a compact region $K \subset \mathbb{H}$, on which $\mathrm{Im}(z) \geq y_0$ for some $y_0 > 0$. A tail of the function is given by summing over the vectors with $|\mathbf{v}| \geq m$. We group this sum by the parameter m:

$$\begin{aligned}\sum_{\substack{\mathbf{v}\in\mathbb{Z}^k \\ |\mathbf{v}|\geq m}} \left|e^{2\pi i|\mathbf{v}|^2 z}\right| &= \sum_{m=M}^{\infty} \sum_{\substack{\mathbf{v}\in\mathbb{Z}^k \\ |\mathbf{v}|=m}} \left|e^{2\pi i m^2 z}\right| \\ &\leq \sum_{m=M}^{\infty} \sum_{\substack{\mathbf{v}\in\mathbb{Z}^k \\ |\mathbf{v}|=m}} e^{-2\pi m^2 y_0}.\end{aligned}$$

What remains is to bound the number of vectors $\mathbf{v}$ on the sphere $|\mathbf{v}| = m$. We choose the extremely weak bound $(2m+1)^l$ and get

$$\sum_{\substack{\mathbf{v}\in\mathbb{Z}^k \\ \max_i v_i \geq m}} \left|e^{2\pi i|\mathbf{v}|^2 z}\right| \leq \sum_{m=M}^{\infty} (2m+1)^l e^{-2\pi m^2 y_0}.$$

For M large enough, we can ensure all the $m \geq M$ satisfy

$$(2m+1)^l \leq e^m \quad \text{and} \quad 2\pi m^2 y_0 \geq m,$$

and thus get

$$\sum_{\substack{\mathbf{v}\in\mathbb{Z}^k \\ \max_i v_i \geq m}} \left|e^{2\pi i|\mathbf{v}|^2 z}\right| \leq \sum_{m=M}^{\infty} e^{-m} \leq e^{1-M}.$$

This gets arbitrarily small as $M \to \infty$ and does not depend on z, so the series converges absolutely and uniformly on K. So $\theta(z,k)$ is holomorphic on $\mathbb{H}$. □

Absolute convergence allows us to freely reorder terms of the Taylor series, so that

$$\begin{aligned}\theta(z,k_1)\theta(z,k_2) &= \left(\sum_{n_1=0}^{\infty} r(n_1,k_1)q^{n_1}\right)\left(\sum_{n_2=0}^{\infty} r(n_2,k_2)q^{n_2}\right) \\ &= \sum_{n=0}^{\infty}\left(\sum_{n_1+n_2=n} r(n_1,k_1)r(n_2,k_2)\right)q^n \\ &= \theta(z,k_1+k_2).\end{aligned}$$

[2] If you are nervous about this rearranging a series before we prove absolute convergence, don't be. Since the two series are rearrangements of one another, one converges absolutely if and only if the other does, so the equality of the two series is a corollary of our argument rather than an assumption.

By definition, $\theta(z,k)$ is automatically $\mathbb{Z}$-periodic in the real direction. It is not weakly modular for all of $\mathrm{SL}_2(\mathbb{Z})$, but is for the appropriate congruence subgroup.

Proposition 16. *The function $\theta(z,4)$ is modular of weight 2 for the congruence subgroup $\Gamma_0(4)$ defined above.*

Proof. First we check weak modularity. The group $\Gamma_0(4)$ is generated by $\pm\begin{pmatrix}1&1\\0&1\end{pmatrix}$ and $\pm\begin{pmatrix}1&0\\4&1\end{pmatrix}$. We know $\theta(z,k)$ are $\mathbb{Z}$-invariant, so it suffices to understand the action of $\gamma=\begin{pmatrix}1&0\\4&1\end{pmatrix}$ on $\theta(z,4)$.

We will study the transformation properties of

$$\theta(z/2,1)=\sum_{d\in\mathbb{Z}} e^{\pi i d^2 z}.$$

First we prove a relation for $z=2it$ on the positive imaginary axis, so $\theta(it,1)=\sum_{d\in\mathbb{Z}} e^{-\pi d^2 t}$. This is closely related to the Gaussian $f(d)=e^{-\pi d^2}$, for which we have $\widehat{f}(m)=e^{-\pi m^2}$ by a simple contour integration argument. Then the time dilation property of Fourier transforms tells us the transform of $f(d\sqrt{t})$ is $\frac{1}{\sqrt{t}}\widehat{f}(m/\sqrt{t})$.

The Poisson summation formula now tells us that for appropriate $h(x)=e^{-\pi x^2 t}$, we have

$$\theta(it,1)=\sum_{d\in\mathbb{Z}} h(d)=\sum_{m\in\mathbb{Z}}\widehat{h}(m)=\sum_{m\in\mathbb{Z}}\frac{1}{\sqrt{t}}f\left(\frac{m}{\sqrt{t}}\right)=\frac{1}{\sqrt{t}}\theta(i/t,1).$$

We rewrite this as

$$\theta(-1/(4z),1)=\sqrt{-2iz}\,\theta(z,1)$$

for all z on the positive imaginary axis. These are two holomorphic functions of z, defined on all of $\mathbb{H}$, which agree on a set with an accumulation point, hence the identity holds on all of $\mathbb{H}$.

Now the composition

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

lets us get a transformation law for $\theta(z,1)$ under γ:

$$\begin{aligned}\theta\left(\frac{z}{4z+1}\right)&=\sqrt{2i\left(\frac{1}{4z}+1\right)}\,\theta\left(-\frac{1}{4z}-1\right)\\&=\sqrt{2i\left(\frac{1}{4z}+1\right)}\,\theta\left(-\frac{1}{4z}\right)\\&=\sqrt{2i\left(\frac{1}{4z}+1\right)}\,\sqrt{-2iz}\;\theta(z)\\&=\sqrt{4z+1}\;\theta(z).\end{aligned}$$

The square root product (in the principal branch) is valid because both parts are in the right half-plane. Applying this to $\theta(z,4)=\theta(z,1)^4$ gives

$$\theta(\gamma(z),4)=(4z+1)^2\;\theta(z,4).$$

Hence $\theta(z,4)$ is weakly modular of weight 2 with respect to $\Gamma_0(4)$.

To show $\theta(z,4)$ is holomorphic at the cusps, we apply Proposition 4 and observe that the coefficients $r(n,k)$ are bounded by $(2n)^k$ since we need only consider numbers of absolute value $\le n$. Hence $\theta(z,4)\in M_2(\Gamma_0(4))$. □

Proof of Theorem 1. We have shown $M_2(\Gamma_0(4))$ is 2-dimensional, and have two Eisenstein series $G_{2,2}$ and $G_{2,4}$ defined above which are clearly linearly independent based on their first terms. Although

$$\begin{aligned}\theta(z,4)&=aG_{2,2}+bG_{2,4}\\&=-a\frac{\pi^2}{3}\left(1+24q+\cdots\right)-b\pi^2\left(1+8q+\cdots\right).\end{aligned}$$

So looking at the first two terms of $\theta(z,4) = 1 + 8q + \cdots$, it is clear that $a = 0$ and $b = -\frac{1}{\pi^2}$. Equating the terms of the sequences gives our desired formula,

$$r(n,4) = 8 \sum_{\substack{0<d|n, \\ 4\nmid d}} d.$$

□

We close with a few remarks. This proof generalizes nicely to the cases of $k = 6, 8$. One uses a variation of the argument to show $\theta(z,k) \in M_{k/2}(\Gamma_1(4))$. There is a general theory of Eisenstein series which gives rise to a canonical decomposition $M_k(\Gamma) = S_k(\Gamma) \oplus E_k(\Gamma)$ (the latter factor the space of Eisenstein series). For those values of k, one can check that the space of cusp forms $S_{k/2}(\Gamma_1(4))$ has dimension zero using a dimension formula, and thereby determine that $\theta(z,k)$ is a sum of Eisenstein series and derive an analogous formula to the one above. This is omitted from this note for the sake of avoiding notational baggage rather than any real difficulty. There is also an exact formula for $k = 2$ which is easy to show directly but hard using modular forms theory because defining a convergent weight 1 Eisenstein series is difficult. For higher even k, there are nontrivial cusp forms which deny the possibility of an exact formula of this type. One can however always recover a similar formula asymptotically, as the Fourier coefficients of Eisenstein series can be shown to dominate those of cusp forms.

For k odd, modular forms are ill-suited to studying the sums-of-squares problem, and the problem is overall harder. The case of $k = 3$ can be handled using quadratic forms in three variables or the arithmetic of imaginary quadratic fields. Hardy-Mordell-Ramanujan developed analytic techniques for the general odd case, [2] though they lack clear arithmetic formulas like can be achieved here.

References

[1] Fred Diamond and Jerry Michael Shurman. *A first course in modular forms*, volume 228. Springer, 2005.

[2] GH Hardy. On the representation of a number as the sum of any number of squares, and in particular of five. *Transactions of the American Mathematical Society*, 21(3):255–284, 1920.

[3] James S Milne. Modular functions and modular forms. 1997.

[4] Jean-Pierre Serre. *A course in arithmetic*, volume 7. Springer Science & Business Media, 2012.

[5] Goro Shimura. *Introduction to the arithmetic theory of automorphic functions*, volume 1. Princeton university press, 1971.