MODULAR FORMS

RAYHAAN PATEL

Abstract. This paper discusses modular forms, highly symmetric functions that satisfy an invariance relation with a group action by $\mathrm{SL}_2(\mathbb{Z})$. They are closely tied to many areas of math; they originate from complex analysis but appear in algebraic topology, sphere packing, number theory and string theory. Their ties to elliptic curves is the centerpiece of Wiles' proof of Fermat's Last Theorem. We examine the basics of modular forms, and use them to enumerate the number of distinct ways to write an integer as the sum of 4 squares.

1. Introduction

We know that all positive integers n can be written as the sum of four nonnegative integer squares, so we may wonder how many ways we can do this for any integer n. To start, we define $r(n)$ to be the number of distinct ways we can write n as the sum of 4 squares (where order matters, and the integer being squared matters so $(-1)^2+0+0+0$ and $1^2+0+0+0$ would be considered distinct). Then we have

Theorem 1.1 (Jacobi's four-square theorem)**.**

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

We will prove this by introducing modular forms, complex functions that are symmetric with respect to a group action by $\mathrm{SL}_2(\mathbb{Z})$.

2. Lattices

To motivate modular forms, we begin by exploring lattices.

Definition 2.1. We define a *complex lattice* Λ to be $\Lambda = \{n\omega_1 + m\omega_2 : m, n \in \mathbb{Z}\}$ such that $\frac{\omega_1}{\omega_2} \notin \mathbb{R}$.

We can transform lattices, Λ by scaling them by a nonzero complex number, λ to get $\lambda\Lambda = \{\lambda a : a \in \Lambda\}$. Now, we consider functions on lattices that generalize linear maps: $f : \mathcal{L} \to \mathbb{C}$, where $\mathcal{L}$ is the set of lattices over $\mathbb{C}$, where $f(\lambda\Lambda) = \lambda^{-k} f(\Lambda)$, for some integer k, all nonzero $\lambda \in \mathbb{C}$ and $\Lambda \in \mathcal{L}$. We call such a function *homogeneous* with *weight* k. It is often useful to express lattices in terms of their generators, so we can rewrite Λ as $\Lambda(\omega_1, \omega_2)$ where ω_1, ω_2 are generators of Λ, so now $f(\Lambda)$ becomes $f(\Lambda(\omega_1, \omega_2))$. But now, we can define a function directly in terms of ω_1 and ω_2: $F(\omega_1, \omega_2) = F(\omega_2, \omega_1) := f(\Lambda(\omega_1, \omega_2))$. Since f is homogeneous, $F(a\omega_1, a\omega_2) = a^{-k} F(\omega_1, \omega_2)$. If we have $a = \frac{1}{\omega_1}$, we get

$$F(1, \frac{\omega_2}{\omega_1}) = \omega_1^k F(\omega_1, \omega_2),$$

Date: February 2025.

so $F(\omega_1, \omega_2)$ is determined entirely by the values of $F(1, \tau)$, so instead we can work with the function $F(\tau) := F(1, \tau)$. Note that if $\frac{\omega_2}{\omega_1}$ has negative imaginary part, $\frac{\omega_1}{\omega_2}$ would have a positive imaginary part, so by swapping ω_1 and ω_2 if necessary, we can ensure $\Im\tau > 0$ (if $\Im\tau = 0$ then ω_1 and ω_2 are not linearly independent, so they don't generate a lattice). Now, we would expect F to have some strict symmetry; we can rescale each lattice to get many different τ depending on the choice of basis for those lattices, and lattices generated by 1 and τ could be generated by many other τ as well, so we will try to determine these symmetries. We start by finding all other bases of the lattice $\Lambda(1, \tau)$: we need 2 linearly independent vectors generated by τ and 1 that generate the whole lattice, so we get $a\tau + b$ and $c\tau + d$, such that the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ has determinant ± 1 and $a, b, c, d \in \mathbb{Z}$.

From here, given some arbitrary $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2(\mathbb{Z})$,

$$F(\tau) = F(1, \tau) = f(\Lambda(1, \tau)) = f(\Lambda(a\tau + b, c\tau + d)) = F(a\tau + b, c\tau + d).$$

If we scale $\Lambda(1, \tau) = \Lambda(a\tau + b, c\tau + d)$ by $\frac{1}{c\tau + d}$ we end up with

$$F(\tau) = F(a\tau + b, c\tau + d) = (c\tau + d)^{-k} F\left(\frac{a\tau + b}{c\tau + d}, 1\right) = (c\tau + d)^{-k} F\left(\frac{a\tau + b}{c\tau + d}\right)$$

or

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

We call meromorphic functions satisfying this property *weakly modular*.

Definition 2.2. A *weakly modular form* of weight $k \in \mathbb{Z}$ is a meromorphic function $f : \mathcal{H} \to \mathbb{C}$ that satisfies

$$f(\gamma(\tau)) = (c\tau + d)^k f(\tau)$$

for all $\gamma \in \mathrm{SL}_2(\mathbb{Z})$, where $\mathcal{H} = \{\tau \in \mathbb{C} : \Im(\tau) > 0\}$ is the *upper half plane*.

3. Series representation of modular forms

In our work above, a change of lattice basis, which is the result of multiplication by $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ corresponds to the map $\frac{a\tau + b}{c\tau + d}$. Thus, we might expect that the group $\mathrm{SL}_2(\mathbb{Z})$ acts on $\mathbb{C} \cup \{\infty\}$ with the map

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}(\tau) = \frac{a\tau + b}{c\tau + d}$$

where $\tau \in \mathbb{C} \cup \{\infty\}$, which is indeed the case.

However, we do have a few special cases involving ∞ or division by 0 that we need to deal with: if $c \neq 0$ then $\frac{-d}{c} \mapsto \infty$ and $\infty \mapsto \frac{a}{c}$, and if $c = 0$ then $\infty \mapsto \infty$. Note that A and $-A$ produce the same transformation of $\mathbb{C} \cup \{\infty\}$ for all $A \in \mathrm{SL}_2(\mathbb{Z})$.

Since $\mathrm{SL}_2(\mathbb{Z})$ is generated by $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, which act on $\tau \in \mathbb{C}$ by $\tau \mapsto \tau + 1$ and $\tau \mapsto \frac{-1}{\tau}$, respectively, we only need to check that

$$f(\tau + 1) = f(\tau)$$

and

$$f\left(\frac{-1}{\tau}\right) = \tau^k f(\tau)$$

rather than

$$f(\gamma(\tau)) = (c\tau + d)^k f(\tau).$$

Weakly modular forms can only have even weighx't: if $\gamma = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$ then we get $(-1)^k f(\tau) = f\left(\frac{-\tau}{-1}\right) = f(\tau)$, and if k is odd, we have a contradiction. Since weakly modular functions are periodic, we write down their Fourier series; we can take the holomorphic map $\tau \mapsto e^{2\pi i\tau} = q$, which maps $\mathcal{H}$ to the punctured unit disk $D' = \{q \in \mathbb{C} : 0 < |q| < 1\}$. Now, we get the function $g : D' \to \mathbb{C}$ defined by $g(e^{2\pi i\tau}) = f(\tau)$, that is $g(q) = f(\frac{\log(q)}{2\pi i})$ (which is only well defined because f is periodic). If f is holomorphic, then g is holomorphic, so we can write a Laurent expansion for g,

$$g(q) = \sum_{n\in\mathbb{Z}} a_n q^n$$

for $q \in D'$. We can extend g to be holomophic at 0, so it is defined on the entire nonpunctured unit disk $D = \{q \in \mathbb{C} : |q| < 1\}$. Since

$$|q| = \left|e^{2\pi i\tau}\right| = \left|e^{-2\pi\Im(\tau)}\right| \left|e^{2\pi i\Re(\tau)}\right| = e^{-2\pi\Im(\tau)},$$

we can see that as $\Im(\tau) \to \infty$, $q \to 0$. Thus, if we consider ∞ to by infinitely far in the imaginary direction, we can define f to be holomophic at ∞ if g has a holomorphic extension to 0. This would imply the Laurent series for g does not contain terms with negative exponents, and since the Laurent series corresponds to the the Fourier series of f, we get a Fourier series of the form

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

We define modular forms as:

Definition 3.1. A *modular form* for $\mathrm{SL}_2(\mathbb{Z})$ is a function $f : \mathcal{H} \to \mathbb{C}$ such that

(1) f is holomorphic on $\mathcal{H}$,
(2) f is holomorphic at ∞, and
(3) f is a weakly modular form; for all $\gamma = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2(\mathbb{Z})$, $f(\gamma(\tau)) = (c\tau + d)^k f(\tau)$.

If f is 0 at ∞, then it is called a *cusp* form.

We can generalize this definition to involve subgroups of $\mathrm{SL}_2(\mathbb{Z})$, which we will do later.

4. Eisenstein series and Congruence Subgroups

We now introduce an example of a modular form: the Eisenstein series.

Definition 4.1. Given a lattice, Λ, the Eisenstein series of weight k is

$$G_k(\Lambda) = \sum_{\omega\in\Lambda/\{0\}} \frac{1}{\omega^k}.$$

It resembles the Riemann zeta function, just defined on lattices instead.

We can see that $G_k(\Lambda)$ is homogeneous: for all $\lambda \in \mathbb{C}/\{0\}$,

$$G_k(\lambda\Lambda) = \sum_{\omega\in\Lambda/\{0\}} \frac{1}{(\lambda\omega)^k} = \lambda^{-k}G_k(\Lambda).$$

From our previous work with lattices, we can rewrite G_k as a single variable weakly modular function, where we set $\Lambda = \Lambda(1,\tau)$ and get

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

For odd k, all of the (m,n) and $(-m,-n)$ terms cancel, giving us 0 for all τ, and if $k=2$, the series is not absolutely convergent

Theorem 4.2. *When $k \geq 4$,*

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

where $\sigma_k(n)$ is the sum of the k^{th} powers of the divisors of n, and $q = e^{2\pi i\tau}$

We will not discuss the proof of this.

Theorem 4.3. *The series $G_k(\tau)$ is a weight k modular form.*

Proof. We can see that $G_k(\frac{\log(q)}{2\pi i})$, which converts the above series from a Fourier series to a power series, is analytic, and therefore holomorphic. Thus $G_k(\tau)$ is holomorphic. Since we can analytically extend

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

to $q=0$ and get $2\zeta(k)$, so $G_k(\infty) = 2\zeta(k)$ is holomorphic at $G_k(\infty)$.

We already know that $G_k(\tau)$ is weakly modular, so it is a modular form as claimed. 

This is already unexpected: we started by essentially extending the zeta function to lattices, and ended up with a modular form encoding the sum of powers of divisors of n in its Fourier coefficients (though given that its related to ζ it might not be so surprising). We also want to define a normalized version of the Eisenstein series:

$$E_k(\tau) = \frac{G_k(\tau)}{2\zeta(k)}.$$

Definition 4.4. Given any positive integer n, the *principal congruence subgroup of level n* is

$$\Gamma(n) = \left\{\gamma \in \mathrm{SL}_2(\mathbb{Z}) : \gamma \equiv \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} \pmod{n}\right\}.$$

Definition 4.5. A subgroup Γ of $\mathrm{SL}_2(\mathbb{Z})$ that contains $\Gamma(n)$ for some n is called a *congruence subgroup* of $\mathrm{SL}_2(\mathbb{Z})$ with *level n*.

We will use the congruence subgroup

$$\Gamma_0(n) = \left\{\begin{bmatrix}a & b\\ c & d\end{bmatrix} \in \mathrm{SL}_2(\mathbb{Z}) : c \equiv 0 \pmod{n}\right\}$$

to construct modular forms invariant under $\Gamma_0(n)$ rather than $\mathrm{SL}_2(\mathbb{Z})$.

5. Jacobi's four-square theorem

Definition 5.1. The *theta function* is

$$\theta(\tau) = \sum_{n\in\mathbb{Z}} q^{n^2},$$

where $q = e^{2\pi i\tau}$.

Note that θ is holomorphic on $\mathcal{H}$, and

$$\theta(\tau)^k = \sum_{n=0}^{\infty} r(n)q^n;$$

each combination of squares that sums to n gives us one q^n term, so they add to $r(n)q^n$. Also,

Theorem 5.2. *When k is an even positive integer, θ^k is a weight $\frac{k}{2}$ modular form for $\Gamma_0(4)$.*

We will not discuss the proof of this.

Definition 5.3. Given a congruence subgroup Γ of $\mathrm{SL}_2(\mathbb{Z})$, we call the vector space of modular forms of weight k over the field $\mathbb{C}$ for Γ $\mathcal{M}_k(\Gamma)$.

The vector spaces of modular forms are finite dimensional, which we will use to prove the sum of squares formula.

Theorem 5.4 (Jacobi's four-square theorem)**.**

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

Proof. Since $\dim \mathcal{M}_2(\Gamma_0(4)) = 2$ (a result we will not show here), and the modular forms E_2^2 and E_2^4 are linearly independent (and therefore a basis), we can write θ^4 as a linear combination of E_2^2 and E_2^4. For some α, β, we get

$$\theta^4 = \alpha E_2^2 + \beta E_2^4.$$

If we expand the first 3 terms of the Fourier series, we get

$$\theta^4 = 1 + 8q + 24q^2$$
$$E_2^2 = -1 - 24q - 24q^2$$
$$E_2^4 = -3 - 24q - 72q^2,$$

and from these coefficients, we can see that $\alpha = 0$ and $\beta = \frac{-1}{3}$.

Therefore,

$$\theta^4 = -\frac{-1}{3}E_2^4 = 1 + 8\sum_{n\geq 1}\left(\sigma(n) - 4\sigma\left(\frac{n}{4}\right)\right),$$

so comparing coefficients gives us

$$r(n) = 8\left(\sigma(n) - 4\sigma\left(\frac{n}{4}\right)\right) = 8\sum_{d|n,4\nmid d} d$$

which completes the proof.