Congruence Subgroups

Undergraduate Mathematics Society, Columbia University

S. M.-C.

24 June 2015

Contents

Abstract

First we expand the connection between the modular group and elliptic curves by defining certain subgroups of the modular group, known as "congruence subgroups", and stating their relation to "enhanced elliptic curves". Then we will carefully define modular forms for congruence subgroups. Finally, as a neat application, we will use modular forms to count the number of ways an integer can be expressed as a sum of four squares.

This talk covers approximately §1.2 and §1.5 of Diamond & Shurman.

1 First Properties

Recall the modular group $\mathrm{SL}_2\,\mathbb{Z}$ of 2×2 integer matrices with determinant one, and its action on the upper half-plane $\mathfrak{H} = \{\tau \in \mathbb{C} : \mathrm{im}(\tau) > 0\}$:

$$\gamma\tau = \frac{a\tau + b}{c\tau + d} \quad \text{for} \quad \gamma = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2\,\mathbb{Z} \quad \text{and} \quad \tau \in \mathfrak{H}. \tag{1}$$

Any subgroup of the modular group inherits this action on the upper half-plane. Our goal is to investigate certain subgroups of the modular group, defined as follows.

Let

$$\Gamma(N) = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2\,\mathbb{Z} : \begin{bmatrix} a & b \\ c & d \end{bmatrix} \equiv \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \bmod N \right\} \tag{2}$$

(where we simply reduce each entry $\bmod N$). As the kernel of the natural morphism $\mathrm{SL}_2\,\mathbb{Z} \to \mathrm{SL}_2\,\mathbb{Z}/N\mathbb{Z}$, we see $\Gamma(N)$ is a normal subgroup of $\mathrm{SL}_2\,\mathbb{Z}$, called the *principal congruence subgroup* of *level* N.

In general, a subgroup Γ of $\mathrm{SL}_2\,\mathbb{Z}$ is called a *congruence subgroup* if $\Gamma(N) \subset \Gamma$ for some N, and the least such N is called the *level* of Γ.

We are particularly interested in the principal congruence subgroups $\Gamma(N)$ and the following other types: let

$$\Gamma_1(N) = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2\,\mathbb{Z} : \begin{bmatrix} a & b \\ c & d \end{bmatrix} \equiv \begin{bmatrix} 1 & * \\ 0 & 1 \end{bmatrix} \bmod N \right\} \tag{3}$$

and

$$\Gamma_0(N) = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2\,\mathbb{Z} : \begin{bmatrix} a & b \\ c & d \end{bmatrix} \equiv \begin{bmatrix} * & * \\ 0 & * \end{bmatrix} \bmod N \right\}, \tag{4}$$

(where $*$ means "anything"). Then $\Gamma_1(N)$ and $\Gamma_0(N)$ are congruence subgroups of $\mathrm{SL}_2\,\mathbb{Z}$ of level N.

From the definitions it is clear that $\Gamma(N) \subset \Gamma_1(N) \subset \Gamma_0(N) \subset \mathrm{SL}_2\,\mathbb{Z}$. The morphisms

$$\begin{aligned} \mathrm{SL}_2\,\mathbb{Z} &\to \mathrm{SL}_2\,\mathbb{Z}/N\mathbb{Z} \\ \begin{bmatrix} a & b \\ c & d \end{bmatrix} &\mapsto \begin{bmatrix} a & b \\ c & d \end{bmatrix} \bmod N \end{aligned} \tag{5}$$

with kernel $\Gamma(N)$,

$$\begin{aligned} \Gamma_1(N) &\to \mathbb{Z}/N\mathbb{Z} \\ \begin{bmatrix} a & b \\ c & d \end{bmatrix} &\mapsto b \bmod N \end{aligned} \tag{6}$$

with kernel $\Gamma(N)$, and

$$\begin{aligned} \Gamma_0(N) &\to (\mathbb{Z}/N\mathbb{Z})^* \\ \begin{bmatrix} a & b \\ c & d \end{bmatrix} &\mapsto d \bmod N \end{aligned} \tag{7}$$

with kernel $\Gamma_1(N)$ show furthermore that $\Gamma(N) \lhd \mathrm{SL}_2\,\mathbb{Z}$ and $\Gamma_1(N) \lhd \Gamma_0(N)$ are normal subgroups, and allow us to make index computations

$$[\mathrm{SL}_2\,\mathbb{Z} : \Gamma(N)] = N^3 \prod_{p|N} \left(1 - \frac{1}{p^2}\right), \tag{8}$$

$$[\Gamma_1(N) : \Gamma(N)] = N, \text{ and} \tag{9}$$

$$[\Gamma_0(N) : \Gamma_1(N)] = \phi(N). \tag{10}$$

Any other index we might want among these subgroups follows from these and multiplicativity $[A : B][B : C] = [A : C]$.

2 The Modular Group and Elliptic Curves

Since the modular group acts on the upper half-plane, we may identify points in the same orbit to form the quotient space $\mathrm{SL}_2\,\mathbb{Z}\backslash\mathfrak{H}$. We have previously referred to the fact that this is the moduli space of elliptic curves. Let's look closer at what this means and why it's true.

Recall that every complex elliptic curve is the quotient $\mathbb{C}/\Lambda$ of the complex plane by a lattice. We've also seen that every lattice (and thus every elliptic curve) can be specified by a point in the upper half-plane, via

$$\tau \quad \longmapsto \quad \Lambda_\tau = \mathbb{Z} + \mathbb{Z}\tau \quad \longmapsto \quad E_\tau = \mathbb{C}/\Lambda_\tau. \tag{11}$$

Now, $E_\tau = \mathbb{C}/\Lambda_\tau$ and $E_{\tau'} = \mathbb{C}/\Lambda_{\tau'}$ are isomorphic elliptic curves (i.e. isomorphic simultaneously as Riemann surfaces and groups) precisely when $\Lambda_{\tau'} = m\Lambda_\tau$ for some $m \in \mathbb{C}$. Regarding τ as fixed, which $\tau' \in \mathfrak{H}$ can produce such a $\Lambda_{\tau'}$? Well $(1, \tau)$ and $(m, m\tau')$ are bases for the same lattice (i.e. $\Lambda_\tau = m\Lambda_{\tau'}$) precisely when

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} \tau \\ 1 \end{bmatrix} = \begin{bmatrix} m\tau' \\ m \end{bmatrix} \text{ for some } \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2\,\mathbb{Z}. \tag{12}$$

Equating each entry, we have $a\tau + b = m\tau'$ and $c\tau + d = m$, and dividing the first equation by the second, $\frac{a\tau+b}{c\tau+d} = \tau'$. This is the statement that τ, τ' are in the same $\mathrm{SL}_2\,\mathbb{Z}$-orbit. (In case it's not obvious, the first two equations together are equivalent to the third, because to go backwards we just define m to be $c\tau + d$).

Thus the $\mathrm{SL}_2\,\mathbb{Z}$-orbits in $\mathfrak{H}$ are in bijective correspondence with isomorphism classes of complex elliptic curves: $E_\tau \cong E_{\tau'}$ if and only if $\mathrm{SL}_2\,\mathbb{Z} \cdot \tau = \mathrm{SL}_2\,\mathbb{Z} \cdot \tau'$. This is the sense in which $\mathrm{SL}_2\,\mathbb{Z}\backslash\mathfrak{H}$ is the *moduli space of elliptic curves*: the points of this space are in bijection with isomorphism classes of complex elliptic curves.

We won't go into as much detail (details can be found in Diamond & Shurman), but congruence subgroups also produce moduli spaces, for so-called "enhanced elliptic curves". An enhanced elliptic curve is an elliptic curve with some distinguished torsion data. The congruence subgroups we've introduced correspond to the following examples.

An *enhanced elliptic curve for* $\Gamma_0(N)$ is an elliptic curve E together with a chosen subgroup C of order N. Two pairs (E, C) and (E', C') are equivalent if there is an isomorphism $E \xrightarrow{\sim} E'$ restricting to an isomorphism $C \xrightarrow{\sim} C'$. The quotient space $\Gamma_0(N)\backslash\mathfrak{H}$ is the moduli space of these enhanced elliptic curves; that is, the points of $\Gamma_0(N)\backslash\mathfrak{H}$ correspond precisely to equivalence classes (E, C).

Similarly, an *enhanced elliptic curve for* $\Gamma_1(N)$ is an elliptic curve E together with a chosen point Q of order N. Two pairs (E, Q) and (E', Q') are equivalent if there is an isomorphism $E \xrightarrow{\sim} E'$ taking $Q \mapsto Q'$. The quotient space $\Gamma_1(N)\backslash\mathfrak{H}$ is the moduli space of these enhanced elliptic curves; that is, the points of $\Gamma_1(N)\backslash\mathfrak{H}$ correspond precisely to equivalence classes (E, Q).

Finally, an *enhanced elliptic curve for* $\Gamma(N)$ is an elliptic curve E together with a pair of points P, Q which generate the N-torsion subgroup $E[N]$ of E and have Weil pairing $e_N(P,Q) = e^{2\pi i/N}$. Two tuples (E,P,Q) and (E',P',Q') are equivalent if there is an isomorphism $E \xrightarrow{\sim} E'$ taking $P \mapsto P'$ and $Q \mapsto Q'$. The quotient space $\Gamma(N)\backslash\mathfrak{H}$ is the moduli space of these enhanced elliptic curves; that is, the points of $\Gamma(N)\backslash\mathfrak{H}$ correspond precisely to equivalence classes (E,P,Q).

3 Modular Forms for Congruence Subgroups

Now we define modular forms for congruence subgroups, which is slightly more subtle than modular forms for the full modular group $\mathrm{SL}_2\mathbb{Z}$. First some notation: for $\gamma = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}_2\mathbb{Z}$ and $\tau \in \mathfrak{H}$, define the *factor of automorphy* $j(\gamma,\tau) = c\tau + d$. Also, we translate the left action of $\mathrm{SL}_2\mathbb{Z}$ on $\mathfrak{H}$ to a right action on functions $\mathfrak{H} \to \mathbb{C}$. For $\gamma \in \mathrm{SL}_2\mathbb{Z}$ and $k \in \mathbb{Z}$, define the *weight-k action* of $\mathrm{SL}_2\mathbb{Z}$ on functions $f : \mathfrak{H} \to \mathbb{C}$ (or the *weight-k operator* $[\gamma]_k$) by

$$(f[\gamma]_k)(\tau) = j(\gamma,\tau)^{-k} f(\gamma\tau). \tag{13}$$

Now for a congruence subgroup Γ we can simply define, as in the case of the modular group: a function $f : \mathfrak{H} \to \mathbb{C}$ is *weakly modular of weight k for* Γ if f is meromorphic and $f[\gamma]_k = f$ for all $\gamma \in \Gamma$.

To define a modular form for a congruence subgroup, we add holomorphy conditions. It is essential that the space of modular forms be finite dimensional; as we'll glimpse in §4, the finite-dimensionality accounts for much of the power of modular forms. For the modular group, we required that our modular forms be holomorphic on the upper half-plane, but also at ∞, which we add to "compactify" the domain. In the case of a general congruence subgroup Γ we need to add still more bits: all of $\mathbb{Q}$ (up to the action of Γ). The Γ-orbits in $\mathbb{Q} \cup \infty$ are called the *cusps* of Γ, for geometric reasons that we'll hopefully see in later lectures.

Since every congruence subgroup contains $\Gamma(N)$ for some N, it contains the translation $\begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix}$ for some least h (which may not be N, but must divide N). Thus if $f : \mathfrak{H} \to \mathbb{C}$ is weakly modular for Γ, it is $h\mathbb{Z}$-periodic, and thus has a Fourier expansion

$$f(\tau) = \sum_{n \in \mathbb{Z}} a_n q^{n/h}, \qquad q = e^{2\pi i \tau}. \tag{14}$$

We say f is *holomorphic at ∞* if $a_n = 0$ for $n < 0$ in the above Fourier expansion. Now if $s \in \mathbb{Q}$, there is some $\alpha \in \mathrm{SL}_2\mathbb{Z}$ with $\alpha\infty = s$. We say f is *holomorphic at s* if $f[\alpha]_k$ is holomorphic at ∞.

We are ready to define modular forms for congruence subgroups. Let $\Gamma \subset \mathrm{SL}_2\mathbb{Z}$ be a congruence subgroup and $f : \mathfrak{H} \to \mathbb{C}$ a function. Then f is a *modular form for* Γ if

(1) f is weakly modular for Γ, i.e. $f[\gamma]_k = f$ for all $\gamma \in \Gamma$;

(2) f is holomorphic on $\mathfrak{H}$; and
(3) f is holomorphic at the cusps, i.e. $f[\alpha]_k$ is holomorphic at ∞ for all $\gamma \in \mathrm{SL}_2 \mathbb{Z}$.

If in addition $a_0 = 0$ in the Fourier expansion of $f[\alpha]_k$ for all $\alpha \in \mathrm{SL}_2 \mathbb{Z}$, i.e. if f "vanishes at the cusps", then f is called a *cusp form*.

A useful alternative to condition (3), which is equivalent to (3) in the presence of (1) and (2), is the following:

(3') for a_n the coefficients of the Fourier expansion $f(\tau) = \sum_{n \geq 0} a_n q^{n/h}$, there are constants $C, r > 0$ such that $|a_n| \leq C n^r$ for $n > 0$.

Let's have some examples. Recall the Eisenstein series that appeared in a previous lecture: for a lattice $\Lambda \subset \mathbb{C}$, we defined

$$G_4(\Lambda) = \sum_{\lambda \in \Lambda^*} \frac{1}{\lambda^4} \quad \text{and} \quad G_6(\Lambda) = \sum_{\lambda \in \Lambda^*} \frac{1}{\lambda^6} \tag{15}$$

(where $\Lambda^* = \Lambda \backslash \{0\}$). We can translate these to functions on the upper half-plane via our friend $\tau \mapsto \Lambda_\tau$. In general, define the *Eisenstein series of weight* k $G_k : \mathfrak{H} \to \mathbb{C}$ for even $k > 2$ by

$$G_k(\tau) = \sum_{\substack{c,d \in \mathbb{Z} \\ (c,d) \neq (0,0)}} \frac{1}{(c\tau + d)^k}. \tag{16}$$

This is a modular form of weight k for $\mathrm{SL}_2 \mathbb{Z}$, and some analysis work, which can be found in Diamond & Shurman, shows that

$$G_k(\tau) = 2\zeta(k) + 2\frac{(2\pi i)^k}{(k-1)!} \sum_{n \geq 1} \sigma_{k-1}(n) q^n, \tag{17}$$

(where $\sigma_r(n) = \sum_{d|n} d^r$).

Why do we require $k > 2$ here? Because if $k = 2$, then the series does not converge absolutely. However, the analogue of (17) for $k = 2$ makes sense, and we can define

$$G_2(\tau) = 2\zeta(2) - 8\pi^2 \sum_{n \geq 1} \sigma_1(n) q^n. \tag{18}$$

The failure of absolute convergence prevents $G_2(\tau)$ from being a modular form, but it still plays a very important role in the theory. For any positive integer N,

$$G_{2,N}(\tau) = G_2(\tau) - N G_2(N\tau) \tag{19}$$

is a modular form of weight 2 for $\Gamma_0(N)$. In fact, the set $\{G_{2,d} : d|N\}$ is a basis for $\mathcal{M}_2(\Gamma_0(N))$ for some small N, including $N = 1, 2, 3, 4, 5, 6, 7, 8, 10$.

4 Sums of Four Squares

Consider the function $\theta(\tau) = \sum_{n \in \mathbb{Z}} q^{n^2}$. This function counts the number of ways to write an integer as the sum of *one* square. That is, if we write $\theta(\tau) =$

$\sum_{k\geq 0} a_k q^k$, then

$$a_k = \begin{cases} 0 & \text{if } k \text{ is not a square} \\ 1 & \text{if } k = 0 \\ 2 & \text{if } k \text{ is a non-zero square.} \end{cases} \tag{20}$$

Note we count $(r)^2$ and $(-r)^2$ as different expressions.

What about the function $\theta(\tau)^2$? In the same way, this function counts the number of ways to write an integer as the sum of *two* squares. Indeed, expanding the product

$$\left(\sum_{n\in\mathbb{Z}} q^{n^2}\right)\left(\sum_{n\in\mathbb{Z}} q^{n^2}\right) = \sum_{m,n\in\mathbb{Z}} q^{m^2+n^2} = \sum_{k\geq 0} a_k q^k, \tag{21}$$

we see that the coefficient a_k is the number of terms $q^{m^2+n^2}$ for which $m^2+n^2 = k$, i.e. the number of expressions of k as a sum of two squares. Note that sign and order matter, so e.g.

$$\begin{aligned} 13 &= (2)^2 + (3)^2 \\ &= (-2)^2 + (3)^2 \\ &= (2)^2 + (-3)^2 \\ &= (-2)^2 + (-3)^2 \\ &= (3)^2 + (2)^2 \\ &= (-3)^2 + (2)^2 \\ &= (3)^2 + (-2)^2 \\ &= (-3)^2 + (-2)^2 \end{aligned} \tag{22}$$

can be "expressed as a sum of two squares" in eight ways ($a_{13} = 8$). This may seem strange, since up to order and sign 13 can be expressed as a sum of two squares in "essentially one way"; but our count has the nice features that

1. k can be written as a sum of two squares if and only if $a_k > 0$,
2. a summand of 0 is distinguished from a positive summand because $0 = -0$, and
3. repeated summands are distinguished from distinct summands.

Thus we actually get a lot of information: in this case,

$$a_k = \begin{cases} 0 & \text{if } k \text{ is not a sum of two squares} \\ 1 & \text{if } k = 0 \\ 4 & \text{if } k \text{ can be expressed in essentially one way as a non-zero square or twice a non-zero square} \\ 8 & \text{if } k \text{ can be expressed in essentially one way as a sum of two distinct non-zero squares} \\ \text{etc.} \end{cases} \tag{23}$$

In precisely the same way, $\theta(\tau)^4$ counts the number of ways to write an integer as the sum of four squares. That is, writing $\theta(\tau)^4 = \sum_{k\geq 0} a_k q^k$, the

coefficient a_k is the number of ways to write k as a sum of four squares. We can determine these a_k by showing that $\theta(\tau)^4$ is a modular form, and then invoking our knowledge of modular forms. (There's a lot we won't prove, but it's all accessible as an exercise or a topic that will be covered in a later lecture; we're looking a bit ahead here).

Since we defined it in terms of $q = e^{2\pi i \tau}$, it's clear that $\theta(\tau + 1) = \theta(\tau)$. Using the Poisson summation formula and some analysis, we can show that

$$\theta(\tfrac{\tau}{4\tau+1}) = \sqrt{4\tau + 1}\,\theta(\tau). \tag{24}$$

Thus $\theta(\tau + 1)^4 = \theta(\tau)^4$ and $\theta(\frac{\tau}{4\tau+1})^4 = (4\tau + 1)^2\theta(\tau)$. This shows $\theta(\tau)^4$ is weakly modular of weight 2 for the subgroup of $\mathrm{SL}_2\,\mathbb{Z}$ generated by $\pm \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and $\pm \begin{bmatrix} 1 & 0 \\ 4 & 1 \end{bmatrix}$, which turns out to be $\Gamma_0(4)$.

Since $\theta(\tau)$ is a sufficiently convergent series of holomorphic functions, it is holomorphic on $\mathfrak{H}$. Now we know $\theta(\tau)^4$ satisfies conditions (1) and (2) for the definition of a modular form for $\Gamma_0(4)$, and it is easy to show that it satisfies condition (3'): a_k is the number of ways to write k as a sum of four squares, and this is trivially bounded by the number of ways to choose four numbers between $-k$ and k (inclusive), which is $(2k+1)^4$. Thus $\theta(\tau)^4$ is a modular form of weight 2 for $\Gamma_0(4)$.

Now, as we observed (without proof) above, the space of such modular forms $\mathcal{M}_2(\Gamma_0(4))$ has a basis

$$\begin{aligned} G_{2,2}(\tau) &= -\frac{\pi^2}{3}\left(1 + 24\sum_{n\geq 1}\left(\sum_{\substack{d|n \\ d \text{ odd}}} d\right) q^n\right) = -\frac{\pi^2}{3}\left(1 + 24q + \cdots\right), \\ G_{2,4}(\tau) &= -\pi^2\left(1 + 8\sum_{n\geq 1}\left(\sum_{\substack{d|n \\ 4\nmid d}} d\right) q^n\right) = -\pi^2\left(1 + 8q + \cdots\right). \end{aligned} \tag{25}$$

We can compute by hand that

$$\theta(\tau)^4 = 1 + 8q + \cdots, \tag{26}$$

and we conclude from linear algebra that $\theta(\tau)^4 = -\frac{1}{\pi^2}G_{2,4}(\tau)$. (Behold the power of finite-dimensional spaces of modular forms!) Comparing coefficients, we see that the number of ways to write a positive integer k as the sum of four squares is

$$a_k = 8\sum_{\substack{d|k \\ 4\nmid d}} d. \tag{27}$$

In particular this is greater than zero, because every integer has a divisor that is not divisible by 4, so every integer can be expressed as a sum of four squares.