The Modular Group

1. The Modular Group

The *modular group* Γ is the set of all 2×2 matrices with integral elements and determinant 1. That is, Γ is the special linear group of 2×2 matrices over the integers, $\Gamma = \mathrm{SL}(2, \mathbb{Z})$. It forms a group under matrix multiplication. If

$$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \Gamma$$

Then M defines a map

$$f_M(z) = \frac{az + b}{cz + d}$$

of the extended complex plane to itself. Here $f_M(-d/c) = \infty$, and $f_M(\infty) = a/c$. Suppose that $N = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}$. Then by direct calculation we find that

$$f_N(f_M(z)) = \frac{(\alpha a + \beta c)z + (\alpha b + \beta d)}{(\gamma a + \delta c)z + (\gamma b + \delta d)} = f_{NM}(z).$$

While it might seem surprising that the composition of one such rational function with another would be connected with matrix multiplication, the mystery can be dispelled by considering how M and N transform the 2-dimensional vectors $\mathbb{C}^2$. Suppose that

$$M \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \qquad N \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}.$$

Then

$$NM \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}.$$

Now suppose we consider these vectors in terms of projective geometry. Two vectors are then considered to be the same if their coordinates are proportional (i.e., the vectors are colinear). In other words, if c is a non-zero complex number, then $\left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} cz_1 \\ cz_2 \end{smallmatrix}\right]$ are considered to be the same. Thus a projective point $z_1 : z_2$ is determined by the ratio $z = z_1/z_2$ if its coordinates, and the image $w_1 : w_2$ is determined by the ratio $w = w_1/w_2$ of its coordinates. But then

$$w = \frac{w_1}{w_2} = \frac{az_1 + bz_2}{cz_1 + dz_2} = \frac{az_1/z_2 + b}{cz_1/z_2 + d} = \frac{az + b}{cz + d},$$

so the map from z to w reflects a linear transformation in projective coordinates.

The modular group

Let $\mathcal{H}$ denote the upper half-plane of the complex plane, $\mathcal{H} = \{z \in \mathbb{C} : \Im z > 0\}$. Write $z = x + iy$. Then

$$w = \frac{az+b}{cz+d} = \frac{az+b}{cz+d}\,\frac{c\overline{z}+d}{c\overline{z}+d} = \frac{ac(x^2+y^2)+(bc+ad)x+bd}{(cx+d)^2+(cy)^2} + i\frac{(ad-bc)y}{(cx+d)^2+(cy)^2}.$$

Since $ad - bc = 1$, we deduce that $w \in \mathcal{H}$ if and only if $z \in \mathcal{H}$. We call two points $z \in \mathcal{H}$ and $w \in \mathcal{H}$ *equivalent* if there is an $M \in \Gamma$ such that $w = f_M(z)$. Since the identity matrix I takes z to itself, it follows that $z \sim z$. If M takes z to w, then M^{-1} takes w to z. That is, $z \sim w$ implies that $w \sim z$. Finally, if $z \sim w$ and $w \sim t$, then NM takes z to t, so that $z \sim t$. Hence this is an equivalence relation, and the upper half-plane $\mathcal{H}$ is partitioned into equivalence classes. In analogy with complete systems of residues (mod m), we say that a set $\mathcal{S} \subseteq \mathcal{H}$ is a *fundamental region* of Γ if for every $z \in \mathcal{H}$ there is exactly one $w \in \mathcal{S}$ such that $z \sim w$.

Theorem 1. *Let $\mathcal{R}$ be the set of those $z = x + iy \in \mathcal{H}$ such that either $-1/2 \le x < 1/2$ and $|z| > 1$ or else $-1/2 \le x \le 0$ and $|z| = 1$. Then $\mathcal{R}$ is a fundamental region for Γ.*

Proof. We show first that if $z \in \mathcal{H}$, then there is a $w \in \mathcal{R}$ that is equivalent to z. Let $S = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ and $T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, and note that

$$\begin{aligned} S\mathcal{R} &= \{z \in \mathcal{H} : |z+1| > 1, |z-1| \ge 1, |z| < 1\} \cup \{z \in \mathcal{H} : |z| = 1, 0 \le x \le 1/2\}, \\ ST^{-1}\mathcal{R} &= \{z \in \mathcal{H} : |z-1| < 1, |z-1/3| \ge 1/3, x < 1/2\} \\ &\qquad \cup \{z \in \mathcal{H} : x = 1/2, \sqrt{3}/6 \le y \le 1/2\}, \\ ST\mathcal{R} &= \{z \in \mathcal{H} : |z+1| \le 1, |z+1/3| > 1/3, x > -1/2\} \\ &\qquad \cup \{z \in \mathcal{H} : x = -1/2, 1/2 \le y \le \sqrt{3}/2\}. \end{aligned}$$

Suppose that $z \in \mathcal{H}$. Choose an integer m so that $w = z + m$ satisfies $-1/2 \le \Re w < 1/2$. If $w \in \mathcal{R}$, then we are done. If $w \in ST^{-1}\mathcal{R}$, then $TSw \in \mathcal{R}$, and we are done. If $w \in ST\mathcal{R}$, then $T^{-1}Sw \in \mathcal{R}$, and we are done. Otherwise, $|w| \le 1/\sqrt{3}$, so that $|Sw| \ge 3|w|$. In this case we begin again with Sw. Since the imaginary part increases by a factor of at least 3 upon each repetition, the process eventually terminates with an equivalent member of $\mathcal{R}$.

To complete the proof we show that no member of $\mathcal{R}$ is equivalent to a different member of $\mathcal{R}$. Suppose that $z \in \mathcal{R}$, and that M takes z to $w \in \mathcal{R}$. If $c = 0$, then the condition $\det(M) = 1$ implies that $ad = 1$, so that without loss of generality $a = c = 1$. Then $w = z + b$, and so the condition $-1/2 \le \Re w < 1/2$ implies that $b = 0$. That is, $M = I$, $z = w$. Now suppose that $c \ne 0$. By direct calculation we find that

$$cw - a = \frac{-1}{cz+d}.$$

Thus if $|cz + d| > 1$, then $|cw - a| < 1$. Suppose that $|c| > 1$. Then $z \in \mathcal{R}$ implies that $|cz + d| > 1$, which implies that $|cw - a| < 1$, which implies that $w \notin \mathcal{R}$. Suppose that

$|c| = 1$. By replacing M by $-M$ we may suppose that $c = 1$. The disc $|z + d| < 1$ does not intersect $\mathcal{R}$, and the closed disc $|z + d| \le 1$ has non-empty intersection with $\mathcal{R}$ only when $d = 0$ or $d = 1$. Similarly, the disc $|w - a| < 1$ does not intersect $\mathcal{R}$, and the closed disc $|w - a| \le 1$ has non-empty intersection with $\mathcal{R}$ only when $a = 0$ or $a = -1$. Hence if $z \in \mathcal{R}$ and $w \in \mathcal{R}$, then $|z + d| = 1$ and $|w - a| = 1$, and we have four cases:
1. $a = d = 0$. In order that $\det(M) = 1$, we must have $b = -1$, which is to say that $M = S$ and $f(z) = -1/z$. But $\Re z < 0$ implies $\Re w > 0$, and then $w \notin \mathcal{R}$, since $|w| = 1$. If $z = i$, then $w = i$, so $z = w$, even though $M \ne \pm I$.
2. $a = 0$, $d = 1$. Since $\det(M) = 1$, we have $b = -1$. The only point $z \in \mathcal{R}$ such that $|z + 1| = 1$ is $z = \rho = -1/2 + i\sqrt{3}/2$. But $f(\rho) = -1/(\rho + 1) = \rho$, so again we have $w = z$.
3. $a = -1$, $d = 0$. Since $\det(M) = 1$, we must have $b = -1$. This determines M, and we note that M^{-1} is the matrix treated in the preceding case, so again $z \in \mathcal{R}$ and $w \in \mathcal{R}$ only when $z = w = \rho$.
4. $a = -1$, $d = 1$. In order that $\det(M) = 1$ we must have $b = -2$. Now $|z + 1| = 1$ implies that $z = \rho$. But $f(\rho) = (-\rho - 2)/(\rho + 1) = -3/2 + i\sqrt{3}/2 \notin \mathcal{R}$.

In all cases, if $z \in \mathcal{R}$, $w \in \mathcal{R}$, and $z \sim w$, then $z = w$, so the proof is complete.

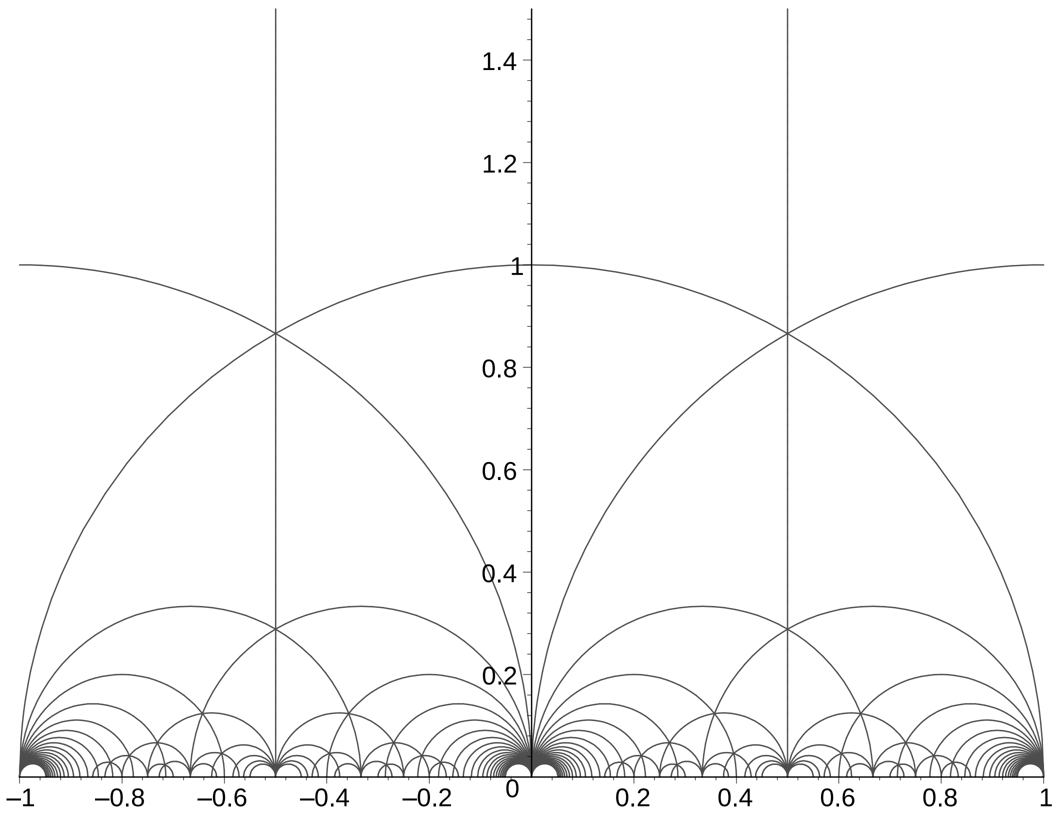

Figure 1. The Fundamental Domain and some of its images.

By examining the above proof, we may observe that $z \in \mathcal{R}$ and $f(z) = z$ only when $M = \pm I$, with the exceptions that $f(i) = i$ when M is one of the four matrices $\pm I$, $\pm S$, and $f(\rho) = \rho$ when M is one of the six matrices $\pm I$, $\pm \begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix}$, $\pm \begin{bmatrix} -1 & -1 \\ 1 & 0 \end{bmatrix}$.

2. Positive definite binary quadratic forms

Write

$$\begin{aligned} f(x,y) &= ax^2 + bxy + cy^2 = a(x - ry)(x - \overline{r}y), \\ g(x,y) &= Ax^2 + Bxy + Cy^2 = A(x - Ry)(x - \overline{R}y) \end{aligned}$$

where r and R are in the upper half-plane $\mathcal{H}$. Since $b = -2a\Re r$, to say that $-a < b \le a$ is equivalent to $-1/2 \le \Re r < 1/2$. Since $c = a|r|^2$, the inequality $c > a$ is equivalent to $|r| > 1$. Similarly, the inequalities $0 \le b \le a = c$ are equivalent to $-1/2 \le \Re r \le 0$, $|r| = 1$. Thus we see that f is reduced if and only if $r \in \mathcal{R}$. Moreover, if $g(x,y) = f(m_{11}x + m_{12}y, m_{21}x + m_{22}y)$, then

$$\begin{aligned} g(x,y) &= a(m_{11}x + m_{12}y - r(m_{21}x + m_{22}y))(m_{11}x + m_{12}y - \overline{r}(m_{21}x + m_{22}y)) \\ &= a((-m_{21}r + m_{11})x - (m_{22}r - m_{12})y)((-m_{21}\overline{r} + m_{11})x - (m_{22}\overline{r} - m_{12})y) \\ &= a(m_{11} - rm_{21})(m_{11} - \overline{r}m_{21})\left(x - \frac{m_{22}r - m_{12}}{-m_{21}r + m_{11}}y\right)\left(x - \frac{(m_{22}\overline{r} - m_{12}}{-m_{21}\overline{r} + m_{11}}y\right). \end{aligned}$$

Here we see that $A = a(m_{11} - rm_{21})(m_{11} - \overline{r}m_{21}) = f(m_{11}, m_{21})$, which we already knew. More importantly, we see that

$$R = \frac{m_{22}r - m_{12}}{-m_{21}r + m_{11}},$$

which is to say that

$$r = \frac{m_{11}R + m_{12}}{m_{21}R + m_{22}}.$$

That is, M takes R to r, so that $r \sim R$. Since each $r \in \mathcal{H}$ is equivalent to a unique $R \in \mathcal{R}$, every positive definite binary quadratic form is equivalent to a unique reduced quadratic form.

The coefficients a, b, c of f determine a unique $r \in \mathcal{H}$, provided that $d = b^2 - 4ac < 0$. Conversely, once $d < 0$ is fixed, the coefficients a, b, c can be recovered from r by the relations

$$a^2(r - \overline{r})^2 = d, \qquad a > 0, \qquad b = -a(r + \overline{r}), \qquad c = ar\overline{r}.$$

Thus if $r \sim R$, then we have associated f and g with $f \sim g$.