Depicting groups

Master Course, Spring term 2025

Fernando Chamizo https://matematicas.uam.es/~fernando.chamizo/

Contents. The action on the upper half plane. The modular Riemann manifold and the modular Riemann surface. Fundamental domains. Cusps.

3.1 Looking for the action

One of the first appearances of $\mathrm{SL}_2(\mathbb{Z})$ in number theory was in Gauss' theory of *quadratic forms*. Suppose that we consider a *binary* quadratic form

$$Q(x,y)=Ax^2+Bxy+Cy^2 \qquad \text{with} \quad A,B,C\in\mathbb{Z} \tag{1}$$

and we are interested, as Gauss was, in the representation function

$$R_Q(k)=\#\big\{(m,n)\in\mathbb{Z}^2\,:\,Q(m,n)=k\big\}.$$

If the map

$$\begin{array}{ccc}\mathbb{Z}^2 & \longrightarrow & \mathbb{Z}^2\\ \vec{n} & \longmapsto & \gamma\vec{n}\end{array} \qquad \text{with} \quad \gamma=\begin{pmatrix}a&b\\c&d\end{pmatrix}\in\mathcal{M}_{2\times 2}(\mathbb{Z})$$

is a bijection then $R_Q=R_{Q'}$ with $Q'(x,y)=Q(ax+by,cx+dy)$. Following Gauss, Q and Q' are said to be *equivalent*. Plainly, the map is bijective if and only if $\det(\gamma)=1$ or $\det(\gamma)=-1$. Therefore we obtain $\mathrm{SL}_2(\mathbb{Z})$ and something that is not a group, but it does not add anything really new because swapping columns we get $\mathrm{SL}_2(\mathbb{Z})$ again. Although Gauss considered both signs of the determinant in many articles of his masterpiece [4] under the names *proper* and *improper equivalence*. When things became really hard, he only considered what we call today $\mathrm{SL}_2(\mathbb{Z})$.

For the representation function it is natural to assume that Q is positive definite (recall that Pell's equation has infinitely many integer solutions) and that Q is *primitive*. This means $\gcd(A,B,C)=1$ with the notation of (1). In this situation, Q is determined by the quadratic polynomial $P(X)=Q(X,1)$ obtained dehomogenizing and P is determined by any of their complex conjugate zeros z and $\bar z$, say $\Im(z)>0$ (note that they are in a quadratic imaginary field because P cannot factorize in $\mathbb{Q}[X]$). With simple algebraic manipulations it is deduced that the linear algebra action of $\mathrm{SL}_2(\mathbb{Z})$ on the set of binary primitive positive definite quadratic forms becomes the following action on the zeros:

$$z\longmapsto\frac{az+b}{cz+d} \qquad \text{with} \quad \gamma=\begin{pmatrix}a&b\\c&d\end{pmatrix}\in\mathrm{SL}_2(\mathbb{Z}).$$

Note that γ and $-\gamma$ act in the same way. Getting rid of this ambiguity is a reason to consider the inhomogeneous modular group $\mathrm{PSL}_2(\mathbb{Z}) = \mathrm{SL}_2(\mathbb{Z})/\{\pm I\}$ and the relation between Q and P justifies the terminology homogeneous/inhomogeneous employed in [10].

In this arithmetic context of quadratic forms, z is a quadratic algebraic number, but this is not the case for modular forms. In fact, extensions of modular forms lead to consider matrices γ with non integral entries. With this idea in mind we consider the action of the group $\mathrm{SL}_2(\mathbb{R})$ on the complex *upper half-plane* $\mathbb{H}$ as

$$\begin{array}{ccc} \mathbb{H} & \longrightarrow & \mathbb{H} \\ z & \longmapsto & \frac{az+b}{cz+d} \end{array} \qquad \text{with} \quad \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R}), \quad \mathbb{H} = \{x+iy \,:\, x \in \mathbb{R},\ y \in \mathbb{R}^+\}.$$

The usual notations for this action and for the denominator appearing in it are

$$\gamma z = \frac{az+b}{cz+d} \qquad \text{and} \qquad j_\gamma(z) = cz+d. \tag{2}$$

This j_γ plays an important role in the theory of modular forms and sometimes (e.g., in [8]) is called the *automorphy factor*.

It is important to note that the map $z \mapsto \gamma z$ determines $\gamma \in \mathrm{SL}_2(\mathbb{R})$ except for a sign, it determines it completely as an element of $\mathrm{PSL}_2(\mathbb{R})$. The unit determinant condition avoids ambiguities like $(2z+1)/(2z+3) = (z+1/2)(z+3/2)$. Only the second form is valid because γ is assumed to be in $\mathrm{SL}_2(\mathbb{R})$.

The following results contain calculations to keep in mind. The first one proves that the action is actually well defined. The second is called the *cocycle condition* (the name comes from the jargon of group cohomology). It is a manifestation of the chain rule noting that the derivative of γz is $\big(j_\gamma(z)\big)^{-2}$.

Lemma 3.1. *For any $\gamma, \gamma_1, \gamma_2 \in \mathrm{SL}_2(\mathbb{R})$ and $z \in \mathbb{H}$*

$$\Im(\gamma z) = \frac{\Im(z)}{\big|j_\gamma(z)\big|^2} \qquad \textit{and} \qquad \gamma_1(\gamma_2 z) = (\gamma_1\gamma_2)z.$$

Lemma 3.2. *We have*

$$j_{\gamma_1\gamma_2}(z) = j_{\gamma_1}(\gamma_2 z) j_{\gamma_2}(z) \qquad \textit{for any } \gamma_1, \gamma_2 \in \mathrm{SL}_2(\mathbb{R}) \textit{ and } z \in \mathbb{H}.$$

Although $\mathbb{H}$ is not a group, it can be considered as a quotient of $\mathrm{SL}_2(\mathbb{R})$. In some sense, $\mathrm{SL}_2(\mathbb{R})$ is like $\mathbb{H}$ with a unit vector at each point (this idea is developed in Exercise 3). This is less relevant in the classic theory of modular forms, but it becomes important for the higher dimensional extensions appearing for instance in the celebrated *Langland's program* [9, §8].

3.2 The Riemannian view

Poincaré considered $\mathbb{H}$ endowed with the Riemann metric

$$\frac{|dz|^2}{\big(\Im(z)\big)^2} = \frac{dx^2+dy^2}{y^2} \qquad \text{with} \quad z = x+iy. \tag{3}$$

This metric, called the *Poincaré metric*, is interesting for geometers because it can be proved that the resulting Riemannian manifold, called *Poincaré's half-plane*, has -1 constant curvature. Then it becomes a simple model for the *hyperbolic plane*. This metric is *conformal*, it means that the angles are as in the Euclidean plane.

There is even an explicit formula for the distance between two points (this is very unusual in Riemannian geometry). One of its simplest presentations is [5, (2.2)]

$$d(z,w) = 2\operatorname{arctanh}\Big|\frac{z-w}{z-\overline{w}}\Big|, \qquad \text{recall that} \quad \operatorname{arctanh} x = \frac{1}{2}\log\frac{1+x}{1-x}.$$

Note that $d(z,w_0) < R$ with $R > 0$ leads to $|z-w_0|^2 < C|z-\overline{w}_0|^2$ with $0 < C < 1$ and if $z = x+iy$ the difference between both sides can be written as $(1-C)(x^2+y^2)$ plus terms of smaller degree. This proves that the circles in Poincaré's half-plane are also Euclidean circles. Of course, the radius and the center differ greatly from their Euclidean analogs.

The maps associated to the action of $\mathrm{SL}_2(\mathbb{R})$ have an important geometrical meaning.

Proposition 3.3. *Each $\gamma \in \mathrm{SL}_2(\mathbb{R})$ induces a global isometry (i.e., a bijective isometry) on Poincaré's half-plane.*

Proof. Lemma 3.1 implies that $z \mapsto \gamma z$ is inverted by $z \mapsto \gamma^{-1}z$ then the map is bijective. Recalling that the derivative of γz is $\big(j_\gamma(z)\big)^{-2}$, it also implies $\big(\Im(\gamma z)\big)^{-1}|d\gamma z| = \big(\Im(z)\big)^{-1}|dz|$ and squaring it is deduced that the Poincaré metric remains invariant under the action of γ. □

In fact, it can be proved that $\mathrm{SL}_2(\mathbb{R})$ gives all direct isometries. All the isometries are obtained combining them with the reflection $x+iy \mapsto -x+iy$.

The map $z \mapsto \gamma z$ is a *Möbius transformation*, in particular, it passes straight lines and circles into straight lines and circles [1]. This is enough to get all the (images of the) geodesics.

Proposition 3.4. *A maximal geodesic (i.e., an inextensible geodesic) in Poincaré's half-plane is either a ray parallel to the positive Y axis or a semicircle orthogonal to the X-axis.*

Proof. The ray r_0 given by the positive Y axis is a geodesic. This follows from the minimizing property: The length of a curve $c(t) = x(t)+it$ joining two close points iy_1 and iy_2 in r_0 is

$$\int_{y_1}^{y_2} t^{-1}\sqrt{x'(t)^2+1}\,dt \geq \int_{y_1}^{y_2} t^{-1}\sqrt{0^2+1}\,dt.$$

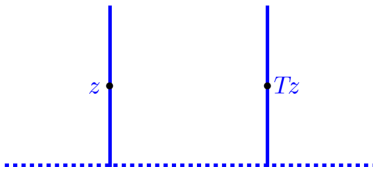
Then it is minimized when $x(t)$ vanishes. As a matter of fact, the arc length parametrization of this geodesic imposing $c(0) = iy_0$ is $c(t) = iy_0e^t$.

Let us name a, b, c and d the entries of $\gamma \in \mathrm{SL}_2(\mathbb{R})$ as in (2). Taking $a = d = 1$, $c = 0$, $b = x_0$, Proposition 3.3 assures that $\gamma r_0 = \{\Re(z) = x_0\} \cap \mathbb{H}$ is a geodesic. On the other hand, given $x_1 < x_2$, solving $a/c = x_2$, $b/d = x_1$, $ad-bc = 1$ we have $\gamma(0^+i) = x_1$ and $\gamma(+\infty i) = x_2$. As γz defines a Möbius transformation, γr_0 is an arc of circle joining x_1 and x_2. This arc is perpendicular to the X axis (in particular, it is a semicircle) because this Möbius transformation is conformal (it preserves angles) and leaves invariant the (compactification of the) X axis, which is perpendicular to r_0.

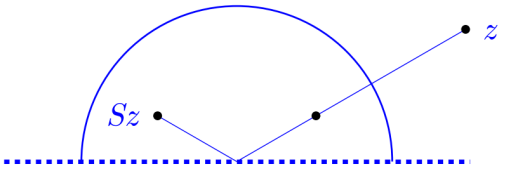
These are all the maximal geodesics because they are clearly inextensible and given a point in $\mathbb{H}$ and a direction there is one of these geodesics passing through the point and having a tangent vector in the chosen direction. □

As a side comment, in the geometric view of Gauss' theory of quadratic forms, an indefinite binary form is associated to the geodesic connecting the solutions of $P(X) = 0$.

Now, the matrices T and S from the first section can be understood in a geometric way: $Tz = z + 1$, so it is a unit *translation*, and $Sz = -1/z$, it applies $re^{i\theta}$ into $r^{-1}e^{i(\pi-\theta)}$. It is an *inversion* [13] followed by a mirror symmetry. It preserves the geodesic named r_0 in the previous proof.



Action of T



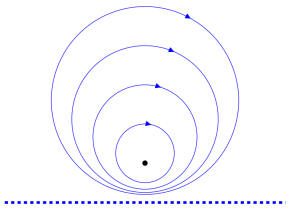
Action of S

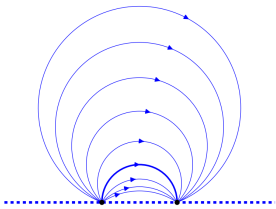
The elements in $\mathrm{SL}_2(\mathbb{R})$ other than $\pm I$ are classified according their fixed points. Namely, $\gamma \in \mathrm{SL}_2(\mathbb{R}) - \{\pm I\}$ is said to be

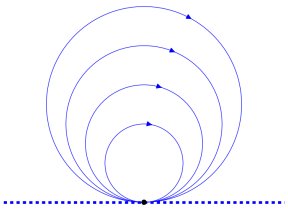
1. *elliptic* if γ has a fixed point in $\mathbb{H}$ (and its complex conjugate in $\overline{\mathbb{H}}$),
2. *hyperbolic* if γ has two fixed points on $\mathbb{R} \cup \{\infty\}$,
3. *parabolic* if γ has only a fixed point on $\mathbb{R} \cup \{\infty\}$.

For instance, S is elliptic because $Si = i$ (and $S(-i) = -i$), T is parabolic because $T\infty = \infty$ and $Tx \neq x$ for every $x \in \mathbb{R}$, and T^4S, which acts as $T^4S = 4 - 1/z$, is hyperbolic because $T^4S(2 \pm \sqrt{3}) = 2 \pm \sqrt{3}$.

In some way, the elliptic elements are a kind of rotations around the fixed point in $\mathbb{H}$, the hyperbolic elements are something like the extension of the translation along a geodesic and the parabolic elements are obtained a degeration of elliptic elements when the center of rotation goes to the boundary of $\mathbb{H}$. For more about the geometry interpretation of the action of $\mathrm{SL}_2(\mathbb{R})$, and the basic geometry of $\mathbb{H}$ see [6, §2].


Elliptic element


Hyperbolic element


Parabolic element

As shown in the figures, an elliptic element leaves invariant off-centered circles around the fixed point. The same is true for a parabolic element when the fixed point belongs to $\mathbb{R}$, but this time the circles are tangent to the X axis at that point. Finally, a hyperbolic element leaves invariant the circular arcs joining the fixed points (assumed in $\mathbb{R}$). The cases in which ∞ is a fixed point are easier to visualize (see the next proof).

Deciding the type of an element of $\mathrm{SL}_2(\mathbb{R}) - \{\pm I\}$ is utterly simple thanks to the following elementary result where $\mathrm{Tr}(\gamma)$ denotes the trace of γ.

Proposition 3.5. *Given $\gamma \in \mathrm{SL}_2(\mathbb{R}) - \{\pm I\}$, it is elliptic if and only if $\left|\mathrm{Tr}(\gamma)\right| < 2$, it is hyperbolic if and only if $\left|\mathrm{Tr}(\gamma)\right| > 2$ and it is parabolic if and only if $\left|\mathrm{Tr}(\gamma)\right| = 2$.*

In particular, any element of $\mathrm{SL}_2(\mathbb{R}) - \{\pm I\}$ is of one of the three types mentioned before.

We give a direct proof based on calculations. For a more elegant algebraic proof appealing to the Jordan canonical form, see [11, §1.2].

Proof. Clearing denominators, the fixed point equation $\gamma z = z$ is

$$cz^2 - (a-d)z - b = 0 \qquad \text{for} \quad \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

If $c = 0$, $\gamma\infty = \infty$ and there is not another fixed point if and only if $a = d$. Clearly, if $a = d$, $a = d = \pm 1$ giving $\left|\mathrm{Tr}(\gamma)\right| = 2$. If $a \neq d$, $\left|\mathrm{Tr}(\gamma)\right| = |a + d| > 2$ because $ad = 1$.

If $c \neq 0$, we have a quadratic equation. Its discriminant $\Delta = (a-d)^2 + 4bc$ can be written as $(a+d)^2 - 4$ using $ad - bc = 1$ and the result follows considering $\Delta < 0$, $\Delta > 0$ and $\Delta = 0$. □

3.3 Riemann again

There is another structure bearing Riemann's name and closely related to modular forms. A *Riemann surface* is a connected one-dimensional complex manifold. The difference with the 1-manifolds in differential geometry is that the chart changes must be given by holomorphic functions. Trivially, $\mathbb{H}$ is a Riemann surface considering the identity chart. A fundamental result called the *uniformization theorem* [5, §2.1] asserts that, except for *conformal equivalence*, the only simply connected Riemann surfaces are the Riemann sphere $\widehat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, the complex plane $\mathbb{C}$ and the upper half plane $\mathbb{H}$.

It turns out that the richness of the theory appears for Riemann surfaces having $\mathbb{H}$ as their universal cover and they can be understood as the action on $\mathbb{H}$ of a *Fuchsian group* [5, Th. 2.1], a discrete subgroup of $\mathrm{SL}_2(\mathbb{R})$. This concept includes the congruence groups Γ that we have considered. Following the standard notation, $\Gamma\backslash\mathbb{H}$ denotes the *orbits* of the (left) action of Γ on $\mathbb{H}$ endowed with the quotient topology. Each element of $\Gamma\backslash\mathbb{H}$ is the orbit of a $z \in \mathbb{H}$, the set $\{\gamma z : \gamma \in \Gamma\}$. This set with the quotient topology induced by the natural projection $\mathbb{H} \longrightarrow \Gamma\backslash\mathbb{H}$ becomes a Riemann surface. It turns out that adding the orbits of the fixed points by parabolic elements, which is indicated as $\Gamma\backslash\mathbb{H}^*$, gives a compact Riemann surface. To summarize it in a statement:

Theorem 3.6. *If Γ is a congruence subgroup then $\Gamma\backslash\mathbb{H}$ is a Riemann surface and $\Gamma\backslash\mathbb{H}^*$ is a compact Riemann surface.*

The interest of the compact Riemann surfaces is that they are the same as algebraic curves over $\mathbb{C}$ [5, §1.3]. We shall discuss the case of the modular group later.

We are not going to prove Theorem 3.6 here (see [11, §1.2]), just to sketch the idea. For the first part, if we forget the existence of elliptic elements the main point is to show that $\Gamma\backslash\mathbb{H}$ is Hausdorff, because then the natural projection $\mathbb{H} \longrightarrow \Gamma\backslash\mathbb{H}$ allows to lift the changes of chart to $\mathbb{H}$.

The elliptic elements cause problems because points arbitrarily close to the fixed point are glued together. The same applies for the parabolic elements in the second part. The points

$T^n(x+iy)$ are very close if y is very large. The solution is to employ functions like $z \mapsto z^n$ passing a sector to a circle for the the elliptic points and functions like $z \mapsto e^{2\pi i z}$ passing half bands to a punctured circle for the parabolic elements.

Let us see the proof of the remaining part.

Proposition 3.7. *With the indicated quotient topology, $\Gamma\backslash\mathbb{H}$ is a Hausdorff topological space.*

The following auxiliary result is fundamental for this purpose.

Lemma 3.8. *Let $z, w \in \mathbb{H}$ with $w \notin \Gamma z$. Then for any neighborhoods $\mathcal{U}_z$ and $\mathcal{U}_z$ of z and w with compact closure in $\mathbb{H}$, the set $\mathcal{S} = \{\gamma \in \Gamma : \gamma(\mathcal{U}_z) \cap \mathcal{U}_w \neq \emptyset\}$ is finite.*

Proof. For any $s \in \mathbb{H}$, using the notation in (2), $|j_\gamma(z)| \to \infty$ when $|c| \to \infty$ or $|d| \to \infty$ and Lemma 3.1 implies $\Im\big(\gamma(s)\big) \to 0$. Then there exists a finite number of possibilities for the lower row of the matrices in $\mathcal{S}$.

If $\gamma_1, \gamma_2 \in \Gamma$ have the same lower row, computing $\gamma_1\gamma_2^{-1}$ it is deduced $\gamma_1 = \pm T^k\gamma_2$ for some $k \in \mathbb{Z}$. Fixing $\gamma_2 \in \mathcal{S}$ corresponding to one of the possibilities for the lower row, it is clear, by the boundedness of $\mathcal{U}_z$ and $\mathcal{U}_w$, that k can only have a finite number of values if $\gamma_1 \in \mathcal{S}$. □

Proof of Proposition 3.7. Take z and w as in Lemma 3.8, they correspond to different orbits in $\Gamma\backslash\mathbb{H}$ and the Hausdorff property follows if we find $\mathcal{U}'_z \subset \mathcal{U}_z$ and $\mathcal{U}'_w \subset \mathcal{U}_w$ such that $\gamma(\mathcal{U}'_z) \cap \mathcal{U}'_w = \emptyset$ for every $\gamma \in \Gamma$. In fact, by Lemma 3.8 we have to consider only a finite number of them. we can find small neighborhoods of z and w satisfying the property and it is enough to take $\mathcal{U}'_z$ and $\mathcal{U}'_w$ as the intersection of these neighborhoods. □

3.4 Fundamental domains

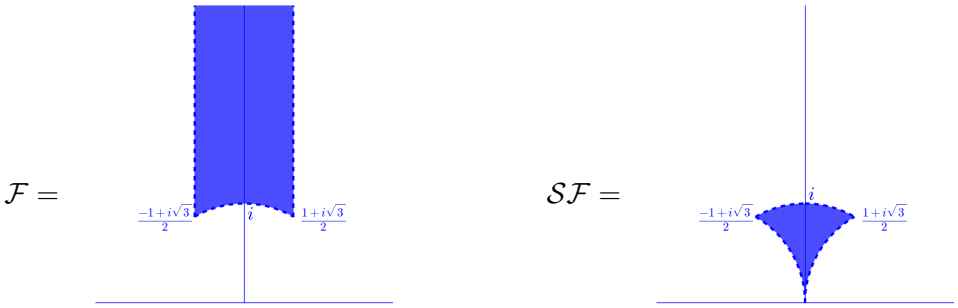
At last we are going to depict groups, using handy regions to see the action of a congruence subgroup (in general, of a Fuchsian group) or rather the topology of the associated Riemann surface. Roughly speaking, a *fundamental domain* is a set $\mathcal{D}$ such that $\bigcup_{\gamma\in\Gamma} \gamma\mathcal{D}$ tessellates $\mathbb{H}$ without substantial overlapping. The topological requirements for $\mathcal{D}$ vary from author to author. Curiously, for many of them a fundamental domain is not a domain (an open connected set). Here, we define a *fundamental domain* $\mathcal{D}$ for a congruence subgroup Γ as a domain $\mathcal{D} \subset \mathbb{H}$ such that distinct points in $\mathcal{D}$ are also distinct in $\Gamma\backslash\mathbb{H}$ (they do not belong to the same orbit) and any orbit contains at least an element in $\overline{\mathcal{D}}$, the closure of $\mathcal{D}$. We also assume in our examples that the boundary is composed by segments of geodesics, in this sense the domain is a polygon (although some sides can have infinite length).

In our context, the first property can be rephrased as $\mathcal{D} \cap \gamma\mathcal{D} = \emptyset$ for $\gamma \in \Gamma - \{\pm I\}$ and the second one as $\bigcup_{\gamma\in\Gamma} \gamma\overline{\mathcal{D}} = \mathbb{H}$.

The domain

$$\mathcal{F} = \big\{z \in \mathbb{H} : |z| > 1,\ |\Re(z)| < 1/2\big\}$$

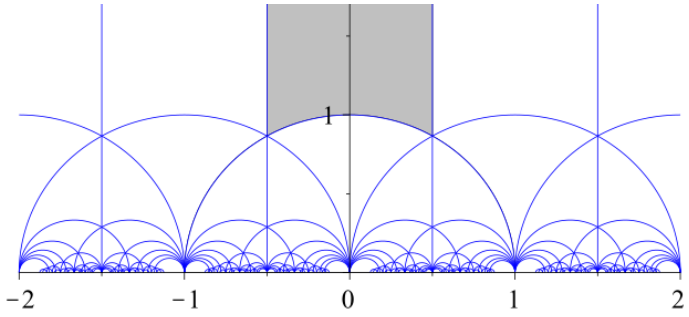
is called the *standard fundamental domain* of $\mathrm{SL}_2(\mathbb{Z})$. Fundamental domains are not unique because applying an element of the group does not affect the properties.



The standard fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$

Another fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$

The following figure, taken from [12], illustrates the mentioned tessellation for $\mathcal{D} = \mathcal{F}$:



Tiling of $\mathbb{H}$ with the images of $\mathcal{F}$ under $\mathrm{SL}_2(\mathbb{Z})$

Note that $T^{\pm 1}\mathcal{F}$ gives the translated patches at both sides of the vertical borders and $S\mathcal{F}$ the patch obtained reflecting across the curved border. Repeating this procedure with the new patches, $\mathbb{H}$ is completely covered.

Let us prove that $\mathcal{F}$ deserves its name.

Proposition 3.9. *The domain $\mathcal{F}$ is a fundamental domain for* $\mathrm{SL}_2(\mathbb{Z})$.

Proof. If $z_1, z_2 \in \mathcal{F}$ are distinct we have to prove first that $z_2 = \gamma z_1$ with $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ is impossible. By the symmetry, we can assume $\Im(z_1) \le \Im(z_2)$. If $\gamma_{21} = 0$ then $\gamma = \pm T^n$ and the contradiction is clear (recall that T is the unit translation). If $\gamma_{21} \neq 0$ the first formula in Lemma 3.1 gives $|j_\gamma(z_1)| \le 1$ that implies $\frac{1}{2}\sqrt{3}\,|\gamma_{21}| + \left|\frac{1}{2}|\gamma_{21}| - |\gamma_{22}|\right| \le 1$ because $\Im(z) > \frac{1}{2}\sqrt{3}$ and $|\Re(z)| < 1/2$ for $z \in \mathcal{F}$. The only possibility with integers is $\gamma_{21} = \pm 1$ and $\gamma_{22} = 0$. Then $\gamma z_1 = n - 1/z_1 \notin \mathcal{F}$ contradicting $z_2 = \gamma z_1$.

It remains to show that for every $z \in \mathbb{H}$ there exists $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ such that $\gamma z \in \overline{\mathcal{F}}$. Given z, chose a $\gamma_0 \in \mathrm{SL}_2(\mathbb{Z})$ minimizing $|j_\gamma(z)|$, then $\Im(\gamma_0 z)$ is maximum (by Lemma 3.1). For any $n \in \mathbb{Z}$ we have

$$\Im(T^n\gamma_0 z) = \Im(\gamma_0 z) \ge \Im(ST^n\gamma_0 z) = \frac{\Im(T^n\gamma_0 z)}{|T^n\gamma_0 z|^2}$$

which implies $|T^n\gamma_0 z| \geq 1$ and with a suitably chosen n we also get $-\frac{1}{2} \leq \Re(T^n\gamma_0 z) \leq \frac{1}{2}$. □

As an application, $\mathcal{F}$ reduces the number of quadratic forms we have to worry about.

Proposition 3.10. *Any positive definite binary primitive quadratic form is equivalent to another of the same discriminant, say $Q = Ax^2 + Bxy + Cy^2$, satisfying $|B| \leq A \leq C$.*

When this condition is fulfilled it is said that Q is a *reduced* quadratic form.

Proof. The discriminant B^2-4AC is invariant under equivalence because $\det(\gamma^t M\gamma) = \det(M)$ for $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ and M the matrix of the quadratic form. Note $\det(M) = AC - B^2/4$.

Identifying the quadratic form with its $z \in \mathbb{H}$ as in the first subsection, it is enough to note that $z = -\frac{B}{2A} + \frac{i}{2A}\sqrt{4AC - B^2}$ belongs to $\overline{\mathcal{F}}$ if and only if $|B| \leq A \leq C$. □

This does not look very impressive. Let us introduce an extra input. Fermat claimed that a prime p is the sum of a (positive) square plus the triple of a square if and only if $3 \mid p-1$. In fact, this representation is unique. Euler got a proof with great effort [2, §1.1]. Proposition 3.10 makes infinitely many representation theorems out of this.

Proposition 3.11. *Let $Q = ax^2 + 2bxy + cy^2$ be a positive definite and primitive quadratic with $ac = b^2 + 3$. Then for p prime $R_Q(p) = 2$ if $p = 3$, $R_Q(p) = 4$ if $3 \mid p - 1$ and $R_Q(p) = 0$ otherwise.*

For instance, $p = 2025 \cdot 2^{599} + 1$ is known to be prime. Clearly $3 \mid p-1$ and we can conclude that $p = 7x^2 + 32xy + 37y^2$ has four integer solutions (x, y) without computing them, which would be very hard by brute force because p has 184 digits.

Proof. By Proposition 3.10 and the invariance of the representation function under equivalence, we can assume that it Q is reduced, $2|b| \leq a \leq c$. Then $3 = ac - b^2 \geq a^2 - \lfloor a/2 \rfloor^2$ and the only possibilities are $a = 1$ and $a = 2$. The latter leads to $c = 2$ which contradicts the primitiveness. Hence $a = 1$, $c = 3$, $b = 0$ and Fermat's claim is applicable. As R_Q takes into account positive and negative solutions, the symmetry $(\pm x)^2 + 3(\pm y)^2 = x^2 + 3y^2$ translates the uniqueness into $R_Q(p) = 4$ except when $p = 3$ than requires $x = 0$. □

There is not need of repeating the proof of Proposition 3.9 for other congruence subgroups, the only important thing is to know the right cosets.

Lemma 3.12. *If $\mathrm{SL}_2(\mathbb{Z}) = \bigcup_{j=1}^{J} \Gamma\gamma_j$ is the right coset decomposition of a congruence subgroup Γ then the interior of $\bigcup_{j=1}^{J} \gamma_j\overline{\mathcal{F}}$ satisfies the properties of being a fundamental domain for Γ except for possibly the connectedness.*

Proof. Let $\mathcal{D}$ the set in the statement. The second property reduces to

$$\bigcup_{\gamma\in\Gamma} \gamma\overline{\mathcal{D}} = \bigcup_{\gamma\in\Gamma} \bigcup_{j=1}^{J} \gamma\gamma_j\overline{\mathcal{F}} = \bigcup_{\gamma\in\mathrm{SL}_2(\mathbb{Z})} \gamma\overline{\mathcal{F}} = \mathbb{H}.$$

For the first property we have to prove that $z, \gamma z \in \mathcal{D}$ with $\gamma \in \Gamma$ only happens for $\gamma = \pm I$. Let $\mathcal{U}_z$ be a neighborhood of z in $\mathcal{D}$ then $\mathcal{U}_z \cap \delta\mathcal{F} = \emptyset$ for any $\delta \in \mathrm{SL}_2(\mathbb{Z}) - \{\pm\gamma_1, \ldots, \pm\gamma_J\}$

because the sets $\gamma\mathcal{F}$ are disjoint and they determine γ except for a sign. On the other hand, $\gamma z \in \mathcal{D}$ implies $\gamma\mathcal{U}_z \cap \gamma_{j_0}\mathcal{F} \neq \emptyset$ for some γ_{j_0}, hence $\mathcal{U}_z \cap \gamma^{-1}\gamma_{j_0}\mathcal{F} \neq \emptyset$ for some j_0 and we conclude $\gamma^{-1}\gamma_{j_0} = \pm\gamma_{j_1}$ for some j_1. As the cosets are disjoint, $j_0 = j_1$ and $\gamma = \pm I$. □

With our knowledge it is unclear if Lemma 3.12 can be turned into an algorithm to obtain fundamental domains [7]. In principle, for $\Gamma(N)$ and $\Gamma_0(N)$ we could follow the proof of Proposition 1.2 to obtain representatives of the right cosets and if the result is not connected, we should translate them with elements of the subgroup to match the different pieces of the jigsaw puzzle.

Here we are only going to show the procedure in two related examples.

Proposition 3.13. *The set*

$$\mathcal{F}_2 = \left\{z \in \mathbb{H} : |\Re(z)| < 1/2,\ |z-1| > 1,\ |z-1/3| > 1/3\right\}$$

is a fundamental domain for $\Gamma_0(2)$.

Proof. We have the right coset decomposition $\mathrm{SL}_2(\mathbb{Z}) = \Gamma_0(2)\gamma_1 \sqcup \Gamma_0(2)\gamma_2 \sqcup \Gamma_0(2)\gamma_3$ with $\gamma_1 = I$, $\gamma_2 = S$, $\gamma_3 = ST$. This follows because $[\mathrm{SL}_2(\mathbb{Z}) : \Gamma_0(2)] = 3$ (Proposition 1.2) and $\gamma_j\gamma_k^{-1} \notin \Gamma_0(2)$ for $j \neq k$. After noting that $\bigcup_j \gamma_j\overline{\mathcal{F}} = \overline{\mathcal{F}}_2$ is connected, Lemma 3.12 assures that $\mathcal{F}_2$ is a fundamental domain. □

We could use that the theta group Γ_θ is a conjugate of $\Gamma_0(2)$ to obtain a fundamental domain for it, but with our previous knowledge, it is easier to apply Lemma 3.12.

Proposition 3.14. *The sets*

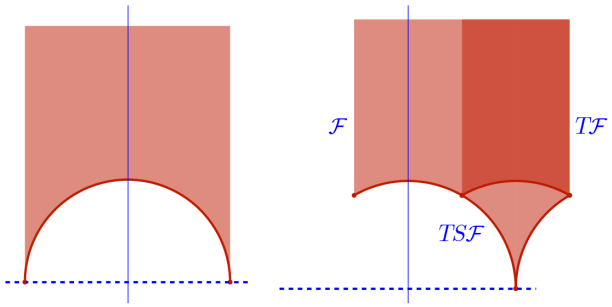
$$\mathcal{F}_\theta = \left\{z \in \mathbb{H} : |\Re(z)| < 1,\ |z| > 1\right\}$$

and

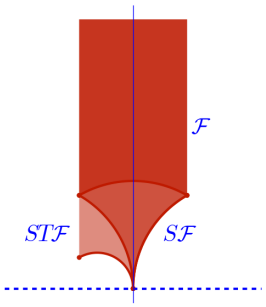
$$\mathcal{F}'_\theta = \left\{z \in \mathbb{H} : -1/2 < \Re(z) < 3/2,\ |z| > 1,\ |z-2| > 1\right\}$$

are fundamental domains for Γ_θ.

Proof. By Proposition 1.7, we can take with $\gamma_1 = I$, $\gamma_2 = T$, $\gamma_3 = TS$. Note $\gamma_1\mathcal{F} = \mathcal{F}$, $\gamma_2\mathcal{F}$ is $\mathcal{F}$ translated by 1 and $\gamma_3\mathcal{F} = T(S\mathcal{F})$ is the region $\{|z| > 1,\ |z-2| > 2,\ |z-1| < 1\}$. By Lemma 3.12 we get the fundamental domain $\mathcal{F}'_\theta$ because $\bigcup_j \gamma_j\overline{\mathcal{F}} = \overline{\mathcal{F}}'_\theta$. Applying $T^{-2} \in \Gamma_\theta$ to the zone $\Re(z) > 1$, we get $\mathcal{F}_\theta$. □



$\mathcal{F}_\theta$ and $\mathcal{F}'_\theta$: Two fundamental domains for Γ_θ

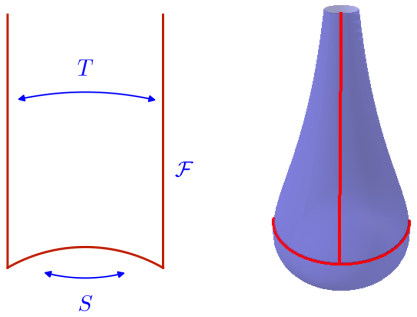


$\mathcal{F}_2$: A fundamental domain for $\Gamma_0(2)$

3.5 Modular topology

With the help of the standard fundamental domain $\mathcal{F}$ we can understand the topology and the complex structure of the Riemann surface $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ and its compactification $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}^*$. The latter is denoted in the literature $X(1)$ because the general notation is $X(N) = \Gamma_0(N)\backslash\mathbb{H}^*$. Some authors also write $Y(N) = \Gamma_0(N)\backslash\mathbb{H}$ and consequently we abbreviate $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ as $Y(1)$.

The points in the right and left vertical borders of $\mathcal{F}$ are related byt T then they are glued together in $Y(1)$. On the other hand, S glues $e^{i\theta}$ with $e^{i(\pi-\theta)}$ in the curved lower border. The result is something like an infinite drop with an asymptotic behavior at infinite if we want to preserve in part the idea of the Poincaré metric that makes points to get closer when their height grows. From the topological point of view, it is homeomorphic to a plane. The fixed points by the parabolic elements of $\mathrm{SL}_2(\mathbb{Z})$ are $\mathbb{Q}\cup\{\infty\}$ and we can think that all of them are in the orbit of a point at infinite height, $+\infty i$, in $\mathcal{F}$. Hence, there is a single point in $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}^*$ corresponding to these fixed points and adding this point with the natural topology (see the comments below) we complete our drop to get something homemorphic to a sphere.

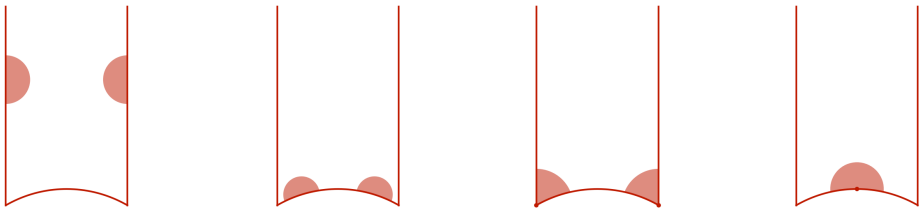


Topology of $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$

Topology of $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}^*$

The next challenge is to go beyond this topological structure adding a complex structure showing the claim in Theorem 3.6 for $\Gamma = \mathrm{SL}_2(\mathbb{Z})$.

Each element in $\mathcal{F}$ represents a unique orbit in $Y(1)$ then we can use the identity chart in $\mathcal{F}$ to give the complex structure to these orbits. The problems only can appear in the orbits corresponding to the boundary of $\mathcal{F}$.



Neighborhoods for points on the boundary of $\mathcal{F}$ with the topology of $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$

Let us give a closer view to the topology of $Y(1)$. Due to the identification of the two vertical boundaries, a small neighborhood of point on them distinct of the vertexes $e^{\pi i/3}$ and $e^{2\pi i/3}$

reappear at the other side. For the complex structure this is not a big deal. In fact the problem vanishes at those points (we can still use the identity chart) if we re-parametrize $Y(1)$ with the fundamental domain

$$\mathcal{F} \cap \{\Re(z) < 1/2 - \epsilon\} \cup \{-1/2 - \epsilon < \Re(z) \leq -1/2,\ |z+1| > 1\}.$$

The same can be said about the lower boundary if we avoid the vertexes $e^{\pi i/3}$ and $e^{2\pi i/3}$ and also i. These are the only *elliptic points* on the boundary of $\mathcal{F}$, the points fixed by elliptic elements. Notice that $e^{\pi i/3}$ and $e^{2\pi i/3}$ are in the same orbit, they are related by T, then we have to solve the problem for i and $e^{2\pi i/3}$. For i we need a map passing a half circle to a circle, multiplying the angle by 2, and for $e^{2\pi i/3}$ we need to multiply the angle by 3 (this is more clear with the alternative fundamental domain mentioned above). Natural charts are

$$\varphi(z) = i + \Big(\frac{z-i}{z+i}\Big)^2 \qquad \text{and} \qquad \varphi(z) = e^{2\pi i/3} + \Big(\frac{z - e^{2\pi i/3}}{z - e^{-2\pi i/3}}\Big)^3$$

We cannot take for instance $i+(z-i)^2$ in the first case, because the neighborhood is not exactly a semicircle, we need something collapsing the two halves of the curved border to the same segment. In the same way, the second transformation glues the geodesics $\{\Re(z) = -1/2\} \cap \mathbb{H}$ and $\{|z| = 1\} \cap \mathbb{H}$ passing through $e^{2\pi/3}$. It opens the curvilinear sector to get a circle.

Finally, a chart for $i\infty$ if given by $\varphi(z) = e^{2\pi z}$. It maps $i\infty$ to 0 and transforms its neighborhoods $\{\Im(z) > M\} \cap \overline{\mathcal{F}}$ into open circles around 0.

Once we know that $X(1) = \mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}^*$ is a Riemann surface and that it is homeomorphic to a sphere, the uniformization theorem says that it is conformally equivalent to the Riemann sphere $\widehat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. Unwrapping this bijective map $f : \mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}^* \longrightarrow \widehat{\mathbb{C}}$ on $\mathbb{H}$, it is obtained a meromorphic function $f : \mathbb{H} \longrightarrow \widehat{\mathbb{C}}$ that is well-defined in the orbits. This means

$$f\Big(\frac{az+b}{cz+d}\Big) = f(z) \qquad \text{for any} \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$$

We shall say that it is a *modular function*. This is of course, completely theoretical, we have not a formula for f. Later, we shall see how to construct this function.

A final comment is that, as mentioned before, compact Riemann surfaces are the same as (projective) algebraic curves. The Riemann sphere is the same as $P^1(\mathbb{C})$ and is the same as a straight line or a conics (a curve of *genus* zero) identifying $P^1(\mathbb{C})$ with the parameter used to parametrize them. For instance,

$$z \mapsto \big(f(z), f(z)\big) \qquad \text{and} \qquad z \mapsto \Big(\frac{1-(f(z))^2}{1+(f(z))^2}, \frac{2f(z)}{1+(f(z))^2}\Big)$$

with f as in the previous paragraph give a conformal equivalence between $X(1)$ and the algebraic curves $y = x$ and $x^2 + y^2 = 1$.

3.6 Comments on cusps

In general, for a congruence group Γ (or for a Fuchsian group with a fundamental domain of finite area), the fixed points of parabolic elements are called *cusps* and it is usual to employ small Fraktur letters to denote them. For instance, saying that $\mathfrak{a} = i\infty$ is a cusp for $\mathrm{SL}_2(\mathbb{Z})$. There is certain ambiguity in the literature about this concept. Sometimes cusps are considered as orbits, as elements of $\Gamma\backslash\mathbb{H}^*$, and some other times as single elements in $\mathbb{R}\cup\{\infty\}$ saying that $\mathfrak{a}$ and $\mathfrak{b}$ are *equivalent cusps* if they belong to the same orbit.

Cusps are very important in the theory of modular forms. Roughly speaking, they indicate where are the infinite points to look at when checking if a function is "globally meromorphic", the generalization of rational functions. The cusps correspond to points in $\mathbb{R}\cup\{\infty\}$ of the closure of $\mathcal{F}$ taken in the Riemann sphere $\mathbb{C}\cup\{\infty\}$ (by the way, naming added the point ∞ or $i\infty$ is only a question of taste). These points give a complete set of representatives of the cusps thought as orbits. Thanks to Lemma 3.12, a finite number of images of $\mathcal{F}$ gives a fundamental domain for a congruence subgroup then the number of cusps in $\Gamma\backslash\mathbb{H}^*$ is always finite and nonzero. In fact there are relatively compact formulas to count them for the special congruence subgroups introduced in the first section [3, §3.8].

Let us check some examples. Using the standard fundamental $\mathcal{F}$ we see the cusp $\mathfrak{a} = \infty$ (recall, a different notation for $\mathfrak{a} = i\infty$) and in $S\mathcal{F}$ we see the cusp $\mathfrak{a} = 0$. Both are equivalent and the typical form of the "visible cusps" like $S\mathcal{F}$ near the origin motivates the name *cusp*. Hence in $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}^*$ we have only a cusp, the orbit or 0 or ∞ that is $\mathbb{Q}\cup\{\infty\}$ as mentioned before.

Using $\mathcal{F}_2$ we see the cusps $\mathfrak{a} = \infty$ and $\mathfrak{b} = 0$ for $\Gamma_0(2)$. They are not equivalent because $(az+b)/(cz+d)$ at ∞ is a/c which cannot be 0 for c even because $\gcd(a,c) = 1$. Consequently, there are two cusps in $\Gamma_0(2)\backslash\mathbb{H}^*$.

If we use the fundamental domain $\mathcal{F}'_\theta$ for Γ_θ, we get two cusps: $\mathfrak{a} = \infty$ and $\mathfrak{b} = 1$. It is easy to see that they are not equivalent. On the other hand, using $\mathcal{F}_\theta$ we obtain a third cusp $\mathfrak{c} = -1$. Clearly $\mathfrak{b}$ and $\mathfrak{c}$ are equivalent because $T^2\mathfrak{c} = \mathfrak{b}$.

The existence of cusps for any congruence subgroup is fundamental in the applications of modular forms to number theory because there are certain expansions attached to the cusps whose coefficients have arithmetic significance.

Exercises of lecture 3

Exercise 1. Prove Lemma 3.1.

Exercise 2. Prove Lemma 3.2.

Exercise 3. For each $z\in\mathbb{H}$ consider all $\gamma\in\mathrm{SL}_2(\mathbb{R})$ such that $z = \gamma i$. Prove that it establishes a bijection $\mathbb{H}\longrightarrow\mathrm{SL}_2(\mathbb{R})/\mathrm{SO}(2)$. Recall that the *special orthogonal group* $\mathrm{SO}(2)$ is the group of matrices corresponding to rotations in $\mathbb{R}^2$ around the origin.

Exercise 4. With the help of your favorite geometry textbook, explain why $\mathbb{H}$ with the Poincaré metric (3) has -1 constant curvature.

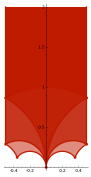
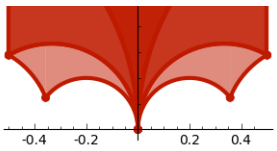
EXERCISE 5. Compute the center and the radius of the circle $|z - i| \leq \frac{1}{2}$ when using the Poincaré metric.

EXERCISE 6. For the following elements of $\mathrm{SL}_2(\mathbb{R})$

$$\begin{pmatrix} 5/4 & 3/4 \\ 3/4 & 5/4 \end{pmatrix}, \qquad \begin{pmatrix} 3/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} 1 & -2 \\ 1 & -1 \end{pmatrix}$$

decide their type, their fixed points in $\mathbb{H} \cup \mathbb{R} \cup \{\infty\}$. Compute the equations of all the Euclidean circles $C \subset \mathbb{H}$ such that C is preserved for the last one.

EXERCISE 7. Prove that the group $\Gamma_0(5)$ admits a fundamental domain of the shape

$\mathcal{D} =$  close up of the lower part = 

Compute explicitly the 5 vertexes of this polygon (excluding $i\infty$). The different color shades do not mean anything, are only for help.

EXERCISE 8. Prove that a positive definite binary primitive quadratic form is equivalent to exactly a quadratic form $Ax^2 + Bxy + Cy^2$ with A, B and C satisfying

$$-A < B \leq A < C \qquad \text{or} \qquad 0 \leq B \leq A = C.$$

EXERCISE 9. Determine explicitly the cusps in $\Gamma_\theta \backslash \mathbb{H}^*$, equivalently, the orbits of the nonequivalent cusps for $\mathcal{F}_\theta$.

EXERCISE 10. Prove that the only elliptic points in $\overline{\mathcal{F}}$ are $e^{\pi i/3}$, $e^{2\pi i/3}$ and i.

References

[1] L. V. Ahlfors. *Complex analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill Book Co., New York, third edition, 1978. An introduction to the theory of analytic functions of one complex variable.

[2] D. A. Cox. *Primes of the form $x^2 + ny^2$. Fermat, class field theory and complex multiplication*. New York etc.: John Wiley &— Sons, 1989.

[3] F. Diamond and J. Shurman. *A first course in modular forms*, volume 228 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2005.

[4] C. F. Gauss. *Disquisitiones arithmeticae*. Springer-Verlag, New York, 1986. Translated and with a preface by A. A. Clarke, Revised by W. C. Waterhouse, C. Greither and A. W. Grootendorst and with a preface by Waterhouse.

[5] E. Girondo and G. González-Diez. *Introduction to compact Riemann surfaces and dessins d'enfants*, volume 79 of *London Mathematical Society Student Texts*. Cambridge University Press, Cambridge, 2012.

[6] M. N. Huxley. Introduction to Kloostermania. In *Elementary and analytic theory of numbers (Warsaw, 1982)*, volume 17 of *Banach Center Publ.*, pages 217–306. PWN, Warsaw, 1985.

[7] R. S. Kulkarni. An arithmetic-geometric method in the study of the subgroups of the modular group. *Amer. J. Math.*, 113(6):1053–1133, 1991.

[8] M. Masdeu. Modular forms (MA4H9). `https://mdave16.github.io/notes/Modular%20Forms%20-%20Marc%20Masdeu.pdf`, 2015.

[9] D. Raboso. When the modular world becomes non-holomorphic. In *Trends in number theory*, volume 649 of *Contemp. Math.*, pages 221–244. Amer. Math. Soc., Providence, RI, 2015.

[10] R. A. Rankin. *Modular forms and functions*. Cambridge University Press, Cambridge-New York-Melbourne, 1977.

[11] G. Shimura. *Introduction to the arithmetic theory of automorphic functions*. Kanô Memorial Lectures, No. 1. Iwanami Shoten Publishers, Tokyo; Princeton University Press, Princeton, NJ, 1971. Publications of the Mathematical Society of Japan, No. 11.

[12] Wikipedia contributors. Fundamental domain — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Fundamental_domain&oldid=1214702129`, 2024. [Online; accessed 12-February-2025].

[13] Wikipedia contributors. Inversive geometry — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Inversive_geometry&oldid=1274260627`, 2025. [Online; accessed 10-February-2025].