Analysis on arithmetic quotients

Bill Casselman
University of British Columbia
cass@math.ubc.ca

Chapter I. The geometry of SL(2)

This chapter is about the geometry of the action of $\mathrm{SL}_2(\mathbb{R})$ on the upper half plane $\mathcal{H}$ as well as some related matters.

Contents

1. The complex projective line

The complex projective line $\mathbb{P}^1(\mathbb{C})$ is the set of all complex lines through the origin in $\mathbb{C}^2$, and is thus the quotient of $\mathbb{C}^2 - \{0\}$ by scalar $\mathbb{C}^\times$ multiplication. Let

$$(x, y) \longmapsto (\!(x, y)\!)$$

be the quotient map. The group $G_\mathbb{C} = \mathrm{GL}_2(\mathbb{C})$ acts on $\mathbb{P}^1(\mathbb{C})$ by homogeneous linear transformations. Only the scalar matrices act trivially. The stabilizer of a point in $\mathbb{P}^1(\mathbb{C})$ is the stabilizer of a line in $\mathbb{C}^2$, hence a **Borel subgroup**—a conjugate of the group $\mathbb{P}_\mathbb{C}$ of upper triangular matrices, which is the group fixing $(\!(1, 0)\!)$. The group $G_\mathbb{C}$ acts transitively on $\mathbb{P}^1(\mathbb{C})$, which may therefore be identified with $G_\mathbb{C}/P_\mathbb{C}$.

The space $\mathbb{P}^1(\mathbb{C})$ may be covered by two copies of $\mathbb{C}$, on the one hand the points $(\!(z, 1)\!)$ (where the second coordinate does not vanish) and on the other the $(\!(1, z)\!)$ (where the first doesn't). We thus have two embeddings of $\mathbb{C}$ into $\mathbb{P}^1(\mathbb{C})$:

$$z \longmapsto (\!(z, 1)\!), \qquad z \longmapsto (\!(1, z)\!)$$

whose images cover $\mathbb{P}^1(\mathbb{C})$, and on the intersection—where neither coordinate vanishes—the transformation from one coordinate to the other is $z \mapsto z^{-1}$. This agrees with the general prescription of an algebraic variety defined over $\mathbb{C}$.

The complement of the first copy of $\mathbb{C}$ is the single point $(\!(1, 0)\!)$. Since it is the limit of points $(\!(1, \varepsilon)\!) = (\!((1/\varepsilon, 1)\!)$ as $\varepsilon \to 0$ and since $1/\varepsilon \to \infty$ as $\varepsilon \to 0$ it is conventional to call it ∞. The space $\mathbb{P}^1(\mathbb{C})$ may therefore be identified with $\mathbb{C} \cup \{\infty\}$. In terms of this identification, since

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} z \\ 1 \end{bmatrix} = \begin{bmatrix} az + b \\ cz + d \end{bmatrix} = (cz + d) \begin{bmatrix} (az + b)/(cz + d) \\ 1 \end{bmatrix},$$

the group $G_\mathbb{C}$ acts by fractional linear transformations

$$g = \begin{bmatrix} a & b \\ c & d \end{bmatrix} : z \longmapsto \frac{az + b}{cz + d},$$

if we agree that $z/0 = \infty$. Such transformations are called **Möbius transformations**.

The term $j(g,z) = cz + d$ is called the **automorphy factor** and occurs frequently in this subject. Its most important property is an immediate consequence of the calculation above:

I.1.1. Proposition. *For g and h in $\mathrm{GL}_2(\mathbb{C})$ and z in $\mathbb{C}$*

$$j(gh,z) = j\big(g, h(z)\big)\, j(h,z)\,.$$

In particular, when restricted to the group fixing z it determines a character with values in $\mathbb{C}^\times$.

Another way in which it appears is this:

I.1.2. Lemma. *For any g in $\mathrm{GL}_2(\mathbb{C})$*

$$\frac{d\,g(z)}{dz} = \frac{\det(g)}{(cz+d)^2} = \frac{\det(g)}{j(g,z)^2} \qquad \left(g = \begin{bmatrix} a & b \\ c & d \end{bmatrix}\right).$$

Proof. Straightforward calculation.

The covering of $\mathbb{P}^1(\mathbb{C})$ by two copies of $\mathbb{C}$ has a simple interpretation in terms of the action of $G_\mathbb{C}$. Let $N_\mathbb{C}$ be the group of upper triangular unipotent matrices. Let $\overline{P}_\mathbb{C}$ be the subgroup of lower triangular matrices and $\overline{N} = \overline{N}_\mathbb{C}$ its unipotent radical. The group $\overline{P}_\mathbb{C}$ is said to be **opposite** to $P_\mathbb{C}$. The intersection $P_\mathbb{C} \cap \overline{P}_\mathbb{C}$ is the group of diagonal matrices, and the choice of opposite determines a splitting of the quotient P/N, which is not otherwise canonical.

We have

$$\begin{aligned} \begin{bmatrix} 1 & 0 \\ z & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} &= \begin{bmatrix} 1 \\ z \end{bmatrix} \\ \begin{bmatrix} 1 & z \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} &= \begin{bmatrix} 1 & z \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} z \\ 1 \end{bmatrix}. \end{aligned}$$

Thus one copy of $\mathbb{C}$ is an $N_\mathbb{C}$-orbit and the other is an $\overline{N}_\mathbb{C}$-orbit. Set

$$w = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Thus $wPw^{-1} = \overline{P}$. Since w swaps $(\!(1,0)\!)$ and $(\!(0,1)\!)$, $P_\mathbb{C}$ fixes $(\!(1,0)\!)$, and $\overline{P}_\mathbb{C}$ fixes $(\!(0,1)\!)$, we can see (upon inversion):

I.1.3. Proposition. (Bruhat decomposition) *The group $G_\mathbb{C}$ is the disjoint union of $P_\mathbb{C}$ and $P_\mathbb{C} w N_\mathbb{C}$ and also that of $N_\mathbb{C}\overline{P}_\mathbb{C} = P_\mathbb{C}\overline{N}_\mathbb{C}$ and $P_\mathbb{C} w = w\overline{P}_\mathbb{C}$. The two open sets $P_\mathbb{C} w N_\mathbb{C}$ and $P_\mathbb{C}\overline{N}_\mathbb{C}$ cover $G_\mathbb{C}$.*

Explicitly, suppose

$$g = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad \Delta = \det(g)\,.$$

Then g lies in $P_\mathbb{C}$ when $c = 0$ and if $c \neq 0$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & a/c \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \Delta/c & 0 \\ 0 & c \end{bmatrix}\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 1 & d/c \\ 0 & 1 \end{bmatrix}. \tag{I.1.4}$$

Furthermore, g lies in $wP_\mathbb{C}$ if and only if $d = 0$, and if $d \neq 0$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & b/d \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \Delta/d & 0 \\ 0 & d \end{bmatrix}\begin{bmatrix} 1 & 0 \\ c/d & 1 \end{bmatrix}. \tag{I.1.5}$$

2. Circles

By a 'circle' in $\mathbb{P}^1(\mathbb{C})$ I'll often mean either a circle or a straight line in $\mathbb{C}$—the straight lines are the circles that pass through ∞. One way in which this convention can be justified is through stereographic projection, which, as was known already to Ptolemy, projects true circles on the 2-sphere to circles and lines on the plane.

I.2.1. Proposition. *A Möbius transformation takes circles in $\mathbb{P}^1(\mathbb{C})$ to circles.*

Proof. The Bruhat decomposition tells us that it suffices to prove this for g diagonal, upper unipotent, or equal to w. Only the last is non-trivial, but still not difficult. However, I'll offer a more direct proof that I think makes clearer what is really going on.

CIRCLES. The equation of a circle with centre at w and radius r is

$$r^2 = (z-w)(\overline{z}-\overline{w}) = z\overline{z} - z\overline{w} - w\overline{z} + w\overline{w} = |z|^2 - 2\,\mathrm{RE}(w\overline{z}) + |w|^2 \,.$$

This may be rewritten as

$$[\,\overline{z} \quad 1\,]\begin{bmatrix} 1 & -w \\ -\overline{w} & |w|^2 - r^2 \end{bmatrix}\begin{bmatrix} z \\ 1 \end{bmatrix} = 0\,. \tag{I.2.2}$$

LINES. The equation of a line is of the form

$$\alpha x + \beta y + C = \mathrm{RE}\big((\alpha - i\beta)(x+iy)\big) + C = 0$$

where α, β, and C are all real. This may be written as

$$[\,\overline{z} \quad 1\,]\begin{bmatrix} 0 & (\alpha + i\beta)/2 \\ (\alpha - i\beta)/2 & C \end{bmatrix}\begin{bmatrix} z \\ 1 \end{bmatrix} = 0\,. \tag{I.2.3}$$

The circles and lines are therefore those curves with an equation of the form

$$A|z|^2 + 2\,\mathrm{RE}(Bz) + C = 0$$

for real numbers A, C and a complex number B, satisfying some conditions I'll formulate in a moment. Such an equation can be written in matrix form

$$[\,\overline{z} \quad 1\,]\begin{bmatrix} A & \overline{B} \\ B & C \end{bmatrix}\begin{bmatrix} z \\ 1 \end{bmatrix} = [\,\overline{z} \quad 1\,]\,H\begin{bmatrix} z \\ 1 \end{bmatrix} = 0\,.$$

The matrix

$$\begin{bmatrix} A & \overline{B} \\ B & C \end{bmatrix}$$

is Hermitian. Since the determinants of the matrices in (I.2.2) and (I.2.3) are negative, this proves one half of:

I.2.4. Lemma. *Circles and lines in $\mathbb{C} \cup \{\infty\}$ are the null cones of Hermitian matrices H with negative determinants.*

Proof. Suppose that H is a Hermitian matrix with negative determinant, say

$$H = \begin{bmatrix} A & \overline{B} \\ B & C \end{bmatrix}.$$

If $A = 0$ its null cone will be the line

$$\alpha x + \beta y + C = 0 \quad \big(\alpha = 2\,\mathrm{RE}(B),\ \beta = -2\,\mathrm{IM}(B)\big)\,.$$

If $A \neq 0$ its null cone will be a circle with center $-\overline{B}/A$ and radius $\sqrt{-\det(H)}/|A|$. If F is any function on $\mathbb{P}^1(\mathbb{C})$ and g is any 2×2 matrix, the transform ${}^gF$ of F is defined by the condition that ${}^gF(x) = F(g^{-1}x)$. If H is a Hermitian form, the matrix of its transform will therefore be

$$ {}^gH = {}^t\overline{g}^{-1} H g^{-1} \, . $$

The form of the equation of the null cone is therefore preserved. ∎

I.2.5. Proposition. *The group* $\mathrm{GL}_2(\mathbb{C})$ *acts transitively on the set of* 2×2 *Hermitian matrices with negative determinant.*

Proof. Suppose H given. We can find a unitary eigenvector matrix V in $\mathbb{U}(2)$ such that

$$ HV = VE, \quad {}^t\overline{V}HV = E $$

where E is the diagonal eigenvalue matrix, with first eigenvalue positive, second negative. But then we can find a diagonal matrix D such that

$$ \overline{D}ED = {}^t\overline{D}ED = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} . $$

In other words, all matrices H lie in the same $\mathrm{GL}_2(\mathbb{C})$-orbit of a single matrix. ∎

The **unitary group** $\mathrm{U}(H)$ of a Hermitian matrix H is the group of all g in $\mathrm{GL}_2(\mathbb{C})$ such that ${}^gH = H$, or ${}^t\overline{g}^{-1}Hg = H$. The **special unitary group** $\mathrm{SU}(H)$ is the subgroup of g in $\mathrm{U}(H)$ with $\det(g) = 1$.

If H is a 2×2 Hermitian matrix, let $h = h_H$ be the function

$$ h(z) = [\,\overline{z} \quad 1\,]\, H \begin{bmatrix} z \\ 1 \end{bmatrix} $$

on $\mathbb{C}$. It is an easy calculation to show:

I.2.6. Proposition. *For*

$$ g = \begin{bmatrix} a & b \\ c & d \end{bmatrix} $$

in $\mathrm{SU}(H)$ *we have*

$$ h(g(z)) = \frac{h(z)}{|cz + d|^2} \, . $$

I.2.7. Corollary. *The group of all* g *in* $\mathrm{SL}_2(\mathbb{C})$ *preserving either of the connected components of the complement of the null cone of* H *is* $\mathrm{SU}(H)$.

There are two special Hermitian matrices we shall be interested in.

I.2.8. Proposition. *If*

$$ H = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} $$

then its null cone is the real projective line, and $\mathrm{SU}(H) = \mathrm{SL}_2(\mathbb{R})$. *If*

$$ H = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} $$

then its null cone is the unit circle, and $\mathrm{SU}(H)$ *is the group of all*

$$ \begin{bmatrix} \alpha & \beta \\ \overline{\beta} & \overline{\alpha} \end{bmatrix} \quad (|\alpha|^2 - |\beta|^2 = 1) \, . $$

In the first case $h(z) = i(z - \overline{z}) = -2\,\mathrm{IM}(z)$, and in the second $h(z) = |z|^2 - 1$.

3. The upper half-plane

The upper half plane is

$$\mathcal{H} = \{x + iy \,|\, y > 0\} \ .$$

It is one of the connected components of the complement of the projective real line $\mathbb{P}^1(\mathbb{R}) = \mathbb{R} \cup \{\infty\}$.

The following is a special case of Proposition I.2.6, but can of course be proved directly:

I.3.1. Lemma. *If*

$$g = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is any real matrix then for z in $\mathcal{H}$

$$\mathrm{IM}\big(g(z)\big) = \frac{\Delta}{\big|cz + d\big|^2} \ \mathrm{IM}(z), \quad \Delta = \Delta(g) = ad - bc \ .$$

I.3.2. Proposition. *The group $\mathrm{SL}_2(\mathbb{R})$ takes $\mathcal{H}$ into itself.*

This action is compatible with that of $\mathrm{SL}_2(\mathbb{R})$ on its boundary $\mathbb{P}^1(\mathbb{R}) = \mathbb{R} \cup \{\infty\}$. Since

$$\frac{ai + b}{ci + d} = i$$

if and only if $a = d$, $b = -c$, the isotropy subgroup fixing i is

$$K = \mathrm{SO}_2 = \left\{ \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \,\middle|\, c^2 + s^2 = 1 \right\},$$

or in other words the group of rotations

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} .$$

The subgroup fixing ∞ is the real Borel subgroup

$$P = P_\infty = \left\{ \begin{bmatrix} * & * \\ 0 & * \end{bmatrix} \right\} .$$

The transformation of $\mathcal{H}$ corresponding to the element

$$a = \begin{bmatrix} t & 0 \\ 0 & t^{-1} \end{bmatrix}$$

is scalar multiplication $z \mapsto t^2 z$, while that corresponding to the unipotent element

$$n = \begin{bmatrix} 1 & x \\ 0 & 1 \end{bmatrix}$$

is the translation $z \mapsto z + x$. These elements all take ∞ to itself, and they generate P.

The group P acts transitively on $\mathcal{H}$, since

$$p = \begin{bmatrix} a & x \\ 0 & a^{-1} \end{bmatrix} : \quad i \longrightarrow ax + a^2 i$$

and consequently $G = NAK$. Explicitly:

I.3.3. Proposition. (Iwasawa decomposition) *We have the factorization*

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & (ac+bd)/r \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1/r & 0 \\ 0 & r \end{bmatrix} \begin{bmatrix} \gamma & -\sigma \\ \sigma & \gamma \end{bmatrix} .$$

where $r = \sqrt{c^2+d^2}$*,* $\gamma = d/r$*,* $\sigma = c/r$*.*

Proof. If $g = pk$ then to find p we solve

$$\frac{ai+b}{ci+d} = \frac{i + (ad+bc)}{c^2+d^2} = r^2 i + rx$$

for r and x, and to find k we solve

$$g^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = k^{-1}p^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} d/r \\ -c/r \end{bmatrix} = k^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \gamma & \sigma \\ -\sigma & \gamma \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \gamma \\ -\sigma \end{bmatrix}$$

for γ and σ.

For each $Y > 0$ define the region in $\mathcal{H}$

$$\mathcal{H}_Y := \left\{ z \in \mathcal{H} \,\middle|\, \mathrm{IM}(z) > Y \right\} .$$

I.3.4. Proposition. *If* $c \neq 0$ *then the image of* $\mathcal{H}_Y$ *under the element*

$$g = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is the open disc of height $1/(c^2Y)$ *just touching* $\mathbb{R}$ *at the rational point* a/c*.*

Proof. Since $g(\infty) = a/c$, according to Proposition I.2.1 the image of the horizontal line $y = Y$ is a circle. Since g takes $\mathbb{P}^1(\mathbb{R})$ to itself, it must be tangent to $\mathbb{R}$ at $g(\infty) = a/c$, say with top at $a/c + iy$. Use Lemma I.3.1 to see that the height of the inverse image of that top is $1/c^2y$, which we set equal to Y.

4. Non-Euclidean geometry

The transformations of $\mathcal{H}$ by elements of the group $G = \mathrm{SL}_2(\mathbb{R})$ are isometries in non-Euclidean geometry.

Since K is the isotropy subgroup of i, a G-invariant Riemannian metric on $\mathcal{H}$ is determined uniquely by a K-invariant metric on the tangent plane at i. How does K act on this plane? Let

$$\varepsilon(k) = c + is \text{ if } k = \begin{bmatrix} c & -s \\ s & c \end{bmatrix} .$$

According to Lemma I.1.2

$$k = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

acts on this tangent plane as multiplication by $\varepsilon(k)^{-2}$, a rotation by -2θ. The metric $ds^2 = dx^2 + dy^2$ is therefore invariant at i and determines a G-invariant metric on all of $\mathcal{H}$. Explicitly the metric at $z = g(i)$ is $(g^{-1})^* ds^2$, which is independent of the choice of g because of K-invariance at i.

I.4.1. Proposition. *The unique* G*-invariant metric on* $\mathcal{H}$ *that restricts to* $dx^2 + dy^2$ *at* i *is*

$$ds^2 = (dx^2 + dy^2)/y^2 .$$

Proof. The group P acts transitively on $\mathcal{H}$, and it is generated by A and N. It is easy to check that the given metric is invariant under each of A_∞ and N_∞. Thus the metric is also the unique G-invariant metric at any point of $\mathcal{H}$ that agrees with $dx^2 + dy^2$ at i. ▯

Lemma I.1.2 tells us that SO_2 acts on the tangent space at i by rotating in the negative direction can be checked easily—an element

$$k = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

with $\theta > 0$ very small is approximately

$$k = \begin{bmatrix} 1 & -\theta \\ \theta & 1 \end{bmatrix} \quad (\theta \sim 0)$$

which acting on $\mathbb{R} \subset \mathbb{P}^1(\mathbb{R})$ takes 0 to $-\theta$, so it does indeed rotate negatively.

The differential forms

$$\frac{dx}{y}, \quad \frac{dy}{y}$$

of degree one form an orthonormal basis of Λ^1_z at every point z of $\mathcal{H}$. They are P-invariant, and hence the 2-form

$$\frac{dx \wedge dy}{y^2}$$

is the unique G-invariant positively oriented one which is $dx \wedge dy$ at i, and determines a G-invariant measure $dx\,dy/y^2$ on $\mathcal{H}$. With this orientation de Rham's **adjoint** operator $\star$ from Λ^1_z to itself takes

$$\frac{dx}{y} \longmapsto \frac{dy}{y}, \quad \frac{dy}{y} \longmapsto -\frac{dx}{y}\ .$$

and the associated codifferential $\delta = \star^{-1} d \star$ takes

$$\begin{aligned} f_x dx + f_y dy &= y f_x \frac{dx}{y} + y f_y \frac{dy}{y} \\ &\xmapsto{\star} -y f_y \frac{dx}{y} + y f_x \frac{dx}{y} \\ &\xmapsto{d} \left(\frac{\partial f_x}{\partial x} + \frac{\partial f_y}{\partial y} \right) dx \wedge dy \\ &= y^2 \left(\frac{\partial f_x}{\partial x} + \frac{\partial f_y}{\partial y} \right) \frac{dx \wedge dy}{y^2} \\ &\xmapsto{\star^{-1}} y^2 \left(\frac{\partial f_x}{\partial x} + \frac{\partial f_y}{\partial y} \right)\ . \end{aligned}$$

This leads to:

I.4.2. Proposition. *The Laplacian differential operator on $\mathcal{H}$ is*

$$\Delta_{\mathcal{H}} = \delta d = y^2 \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)\ .$$

The metric restricted to the vertical line through i is dy/y. Thus the distance from i to iy is $\log y$. The non-Euclidean circle of radius r around i is the K-orbit of iy where $y = e^r$. In the next section we'll see that this orbit is a Euclidean circle. Since the map $z \mapsto -1/z$ is in K, it is that passing through iy and i/y with centre $i(y + 1/y)$.

If D is a non-Euclidean disc around i, z is a point of $\mathcal{H}$, and g an element of G with $g(i) = z$ then the transformed disc $g(D)$ depends only on z and not on the choice of g. Thus to see what $g(D)$ is, we may assume that g is in P, which is to say the product of a scalar multiplication and a horizontal translation. If D has top and bottom at height $e^{\pm r}$ then $y(D)$ has top and bottom at $ye^{\pm r}$. Horizontal translation doesn't change this. Hence:

I.4.3. Lemma. *The non-Euclidean disc of radius r around $z = x + iy$ has top and bottom at $ye^{\pm r}$.*

The geodesic path from i to iy is the vertical line segment between them, since according to the formula $ds^2 = (dx^2 + dy^2)/y^2$ any horizontal deviation adds length to a path. The rotations around i of the vertical line from 0 to ∞ are circular arcs through i. Because Möbius transformations are conformal, each of these must meet the real axis orthogonally in two points. Transforms of these make up all the geodesics, and all except vertical lines are arcs of circles intercepting the x-axis orthogonally. To summarize:

I.4.4. Proposition. *The geodesics in $\mathcal{H}$ are precisely (a) vertical lines or (b) circular arcs meeting the real line at right angles.*

5. The Cayley transform

It follows from Proposition I.2.5 that Möbius transformations act transitively on discs in the complex plane. In particular, the transformation

$$z \longmapsto \mathfrak{C}(z) = (z - i)/(z + i)$$

takes the upper half plane $\mathcal{H}$ to the interior of the unit disc $|z| < 1$, since it takes the real line to the unit circle and i to 0. This is the Möbius transformation corresponding to the matrix

$$\begin{bmatrix} 1 & -i \\ 1 & i \end{bmatrix} .$$

It is called the **Cayley transform**.

The action of $\mathrm{SL}_2(\mathbb{R})$ on $\mathcal{H}$ gives rise to one on the unit disc, which can be calculated conveniently through conjugation. The matrix g in $\mathrm{SL}_2(\mathbb{R})$ gives rise to the matrix

$$\begin{bmatrix} 1 & -i \\ 1 & i \end{bmatrix} g \begin{bmatrix} 1 & -i \\ 1 & i \end{bmatrix}^{-1} = \frac{1}{2} \begin{bmatrix} 1 & -i \\ 1 & i \end{bmatrix} g \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}$$

which takes $\mathbb{D}$ to itself.

The Cayley transform takes i to 0 and $-i$ to ∞, hence takes the complex group of all matrices

$$\begin{bmatrix} c & -s \\ s & c \end{bmatrix} ,$$

which fixes i and $-i$, to the group of all complex diagonal matrices, which fixes 0 and ∞. These are copies of the multiplicative group $\mathbb{C}^\times$ in $\mathrm{GL}_2(\mathbb{C})$, and are conjugate algebraic tori in it. The intersection of the first of these with $\mathrm{GL}_2(\mathbb{R})$ is a copy of $\mathbb{C}^\times$, while the intersection of the second is a product of two copies of $\mathbb{R}^\times$. But whereas in $\mathrm{GL}_2(\mathbb{C})$ these two tori are conjugate, in $\mathrm{GL}_2(\mathbb{R})$ they are very different.

More precisely, the transform takes $\mathrm{SO}(2)$ to unit diagonal matrices:

$$\begin{bmatrix} 1 & -i \\ 1 & i \end{bmatrix} \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \frac{1}{2i} \begin{bmatrix} i & i \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} c - is & 0 \\ 0 & c + is \end{bmatrix}$$

Because the orbits of the conjugate of K in the unit disc are clearly circles, Proposition I.2.1 implies:

I.5.1. Proposition. *The orbits of K in $\mathcal{H}$ are circles.*

To summarize: The unit disc $\mathbb{D}$ is symmetric with respect to rotations, and the Cayley transform is valuable in computing things on $\mathcal{H}$ that are invariant under $\mathrm{SO}(2)$. The upper half-plane is similarly valuable when one is dealing with P. The relationship expressed by the Cayley transform is extremely important even when dealing with more general Lie groups. It is used to keep track of the collection of conjugacy classes of algebraic tori—in effect, it conjugates the compact torus of $\mathrm{SL}_2(\mathbb{R})$ to the split torus, but inside $\mathrm{SL}_2(\mathbb{C})$.

One application of the Cayley transform is to answer easily a basic question about non-Euclidean geometry: *What is the non-Euclidean area of the non-Euclidean circle of radius r?*

The non-Euclidean metric on $\mathcal{H}$ is $(dx^2+dy^2)/y^2$ and the invariant measure $dx\,dy/y^2$. Along $i\mathbb{R}$ we have $ds = dy/y$, so the distance from i to iy with $y \geq 1$ is $r = \log y$, which when inverted tells us $y = e^r$. Transforming differentials by the Cayley transform, one sees that on $\mathbb{D}$ the metric becomes $4(dx^2 + dy^2)/(1 - r^2)^2$ and the invariant measure $4\,dx\,dy/(1 - r^2)^2$. Here r is the Euclidean radius. In polar coordinates, the Euclidean measure is $r\,dr\,d\theta$, so the non-Euclidean area of the Euclidean circle of radius r is

$$
\begin{aligned}
4\int_0^{2\pi}\int_0^{r}\frac{x\,dx\,d\theta}{(1-x^2)^2} &= 4\pi\int_0^{r^2}\frac{dx}{(1-x)^2}\\
&= \frac{4\pi r^2}{1-r^2}\,.
\end{aligned}
$$

The point iy maps to $r = (1 - y)/(1 + y)$ under the Cayley transform, and this leads to:

I.5.2. Proposition. *The area of the non-Euclidean circle of radius ρ is $2\pi\big(\cosh(\rho) - 1\big)$.*

Proof. Because

$$
\frac{4\pi(1-y)^2}{2y} = \frac{2\pi(y-2+1/y)}{2} = 2\pi(\cosh(r)-1)\,. \quad \blacksquare
$$

I.5.3. Corollary. *The length of the circumference of a non-Euclidean circle of radius ρ is $2\pi\sinh(\rho)$.*

6. Norms

For $v = (x, y)$ in $\mathbb{R}^2$, define $\|v\|$ to be its Euclidean norm $\sqrt{x^2 + y^2}$. For g in $G = \mathrm{SL}_2(\mathbb{R})$, define

$$
\|g\| = \sup_{\|v\|=1} \|g(v)\|\,.
$$

Thus (a) $\|g\|$ is right- and left-invariant with respect to $K = \mathrm{SO}(2)$; (b) $\|gh\| \leq \|g\| \cdot \|h\|$; and (c) if

$$
g = \begin{bmatrix} t & 0 \\ 0 & 1/t \end{bmatrix}
$$

then

$$
\|g\| = \sup_\theta \sqrt{t^2\cos^2\theta + t^{-2}\sin^2\theta} = \sup\big(|t|, 1/|t|\big)\,.
$$

There is no really canonical norm on G. I'll call two norms **strictly equivalent** if each is bounded by a positive multiple of the other. For example, the definition of norm above depends on a choice of compact subgroup K, but the norms we get from different K are strictly equivalent. Also, with a given K equivalent norms on the group of diagonal matrices determine equivalent norms on all of $\mathrm{SL}_2(\mathbb{R})$. Thus the norm $(|t| + |1/t|)/2$ on the group of diagonal matrices determines a K-invariant norm strictly equivalent to the one defined above.

I'll write $\equiv$ for strict equivalence.

But for most purposes, all that really matters is a much weaker notion of equivalence—two norms are said to be **weakly equivalent** if each is bounded by a power of the other.

It'll write $\asymp$ for weak equivalence.

Any norm on G that's bi-invariant with respect to right and left multiplication by $K = \mathrm{SO}(2)$ determines a K-invariant norm on both $\mathcal{H}$ and $\mathbb{D}$. There is a simple formula for the norm on $\mathbb{D}$, at least up to strict equivalence. First of all, since the norm is rotation-invariant, it is a function of r alone. The Cayley transform takes iy to $r = (y-1)/(y+1)$, so that

$$y = (1+r)/(1-r), \quad \frac{y+1/y}{2} = \frac{1}{1-r^2}\,.$$

Thus, up to strict equivalence, both norms on G amount to $1/(1-|z|^2)$ on $\mathbb{D}$.

The norm $\|\!|x+iy|\!\|$ of an element of $\mathcal{H}$ is the norm of its image under the Cayley transform. Explicitly:

I.6.1. Proposition. *The norm of $x+iy$ in $\mathcal{H}$ is*

$$\|\!|x+iy|\!\| = \frac{x^2+(y+1)^2}{4y}\,.$$

Proof. We calculate:

$$\begin{aligned} 1 - \left|\frac{z-i}{z+i}\right|^2 &= \frac{(z+i)(\overline{z}-i)-(z-i)(\overline{z}+i)}{x^2+(y+1)^2} \\ &= \frac{-2iz+2i\overline{z}}{x^2+(y+1)^2} \\ &= \frac{4y}{x^2+(y+1)^2}\,. \quad \blacksquare \end{aligned}$$

7. Vector fields

An action of a Lie group on a smooth manifold determines vector fields associated to its Lie algebra. A one parameter subgroup determined by X in $\mathfrak{g}$ gives rise to a flow on M taking m to $\exp(tX)m$. The vector field associated to this is the one associating to the point m the differential operator

$$\left.\frac{d}{dt}\right|_{t=0} f\big(\exp(tX)m\big)$$

The chain rule can be used to calculate this explicitly, but in pracice there is often a simpler technique. For X in $\mathfrak{g}$ we calculate X at m (in terms of coordinates)

$$\frac{(I+\varepsilon X)m - m}{\varepsilon}$$

where we assume $\varepsilon^2 = 0$. Here $I + \varepsilon X$ is to be interpreted in G.

Let's see how to apply this to the action of $\mathrm{SL}_2(\mathbb{R})$ on $\mathcal{H}$ and $\mathbb{D}$.

There are several different important elements of $\mathfrak{g}$, and even several different choices of bases. They are associated with Lie subalgebras of $\mathfrak{g}$:

$$\begin{aligned}
\mathfrak{n}_+ &= \text{ span of } \nu_- = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \\
\mathfrak{n}_- &= \text{ span of } \nu_+ = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \\
\mathfrak{a} &= \text{ span of } \alpha = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \\
\mathfrak{k} &= \text{ span of } \kappa = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} .
\end{aligned}$$

One basis is made up of $\nu_\pm$ and α, compatible with the Bruhat decomposition $\mathfrak{g} = \mathfrak{n}_+ + \mathfrak{a} + \mathfrak{n}_-$.

A second is made up of ν_+, α, and κ, and goes with the Iwasawa decomposition $\mathfrak{g} = \mathfrak{n}_+ + \mathfrak{a} + \mathfrak{k}$.

Thus for $\mathrm{SL}_2(\mathbb{R})$ on $\mathcal{H}$

$$\begin{aligned}
L_{\nu_+} &= \partial/\partial x \\
L_\alpha &= 2x(\partial/\partial x) + 2y(\partial/\partial y)
\end{aligned}$$

and L_{ν_-} can be computed as

$$\frac{z}{\varepsilon z + 1} - z = z(1 - \varepsilon z) - z = -\varepsilon z^2$$

divided by ε, or

$$L_\kappa = -z^2 .$$

How is this to be interpreted as a vector field? In general, a complex number $p + iq$ is to be interpreted as $p(\partial/\partial x) + q(\partial/\partial y)$. So we can write more explicitly

$$L_{\nu_-} = -(x^2 - y^2)\,\frac{\partial}{\partial x} + 2xy\,\frac{\partial}{\partial y} .$$

Incidentally, the Laplacian can be expressed in terms of these vector fields:

$$\Delta = \frac{L_\alpha^2}{4} - \frac{L_\alpha}{2} + \nu_-\nu_+ ,$$

as a simple calculation will verify.

8. References

1. A. Borel, **Automorphic forms on SL(2,R)**, Cambridge University Press, 1997.
2. C. L. Siegel, **Topics in complex function theory**, Volume II, Wiley, 1971.