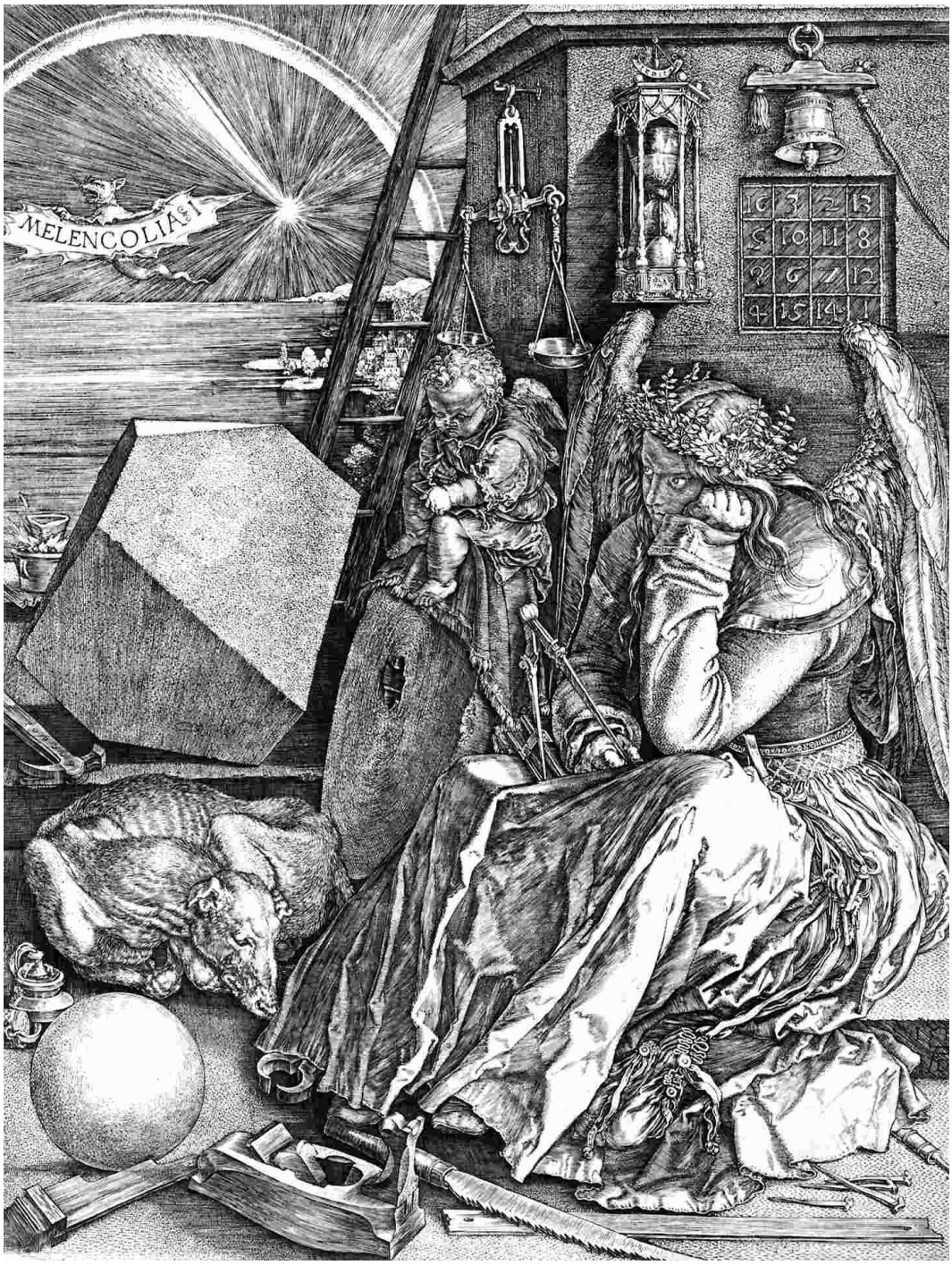
Department of Pure Mathematics and Mathematical Statistics
University of Cambridge

GEOMETRY AND GROUPS



Albrecht Dürer's engraving **Melencolia I** *(1514)*

Notes Michaelmas 2012

T. K. Carne.
t.k.carne@dpmms.cam.ac.uk

CONTENTS

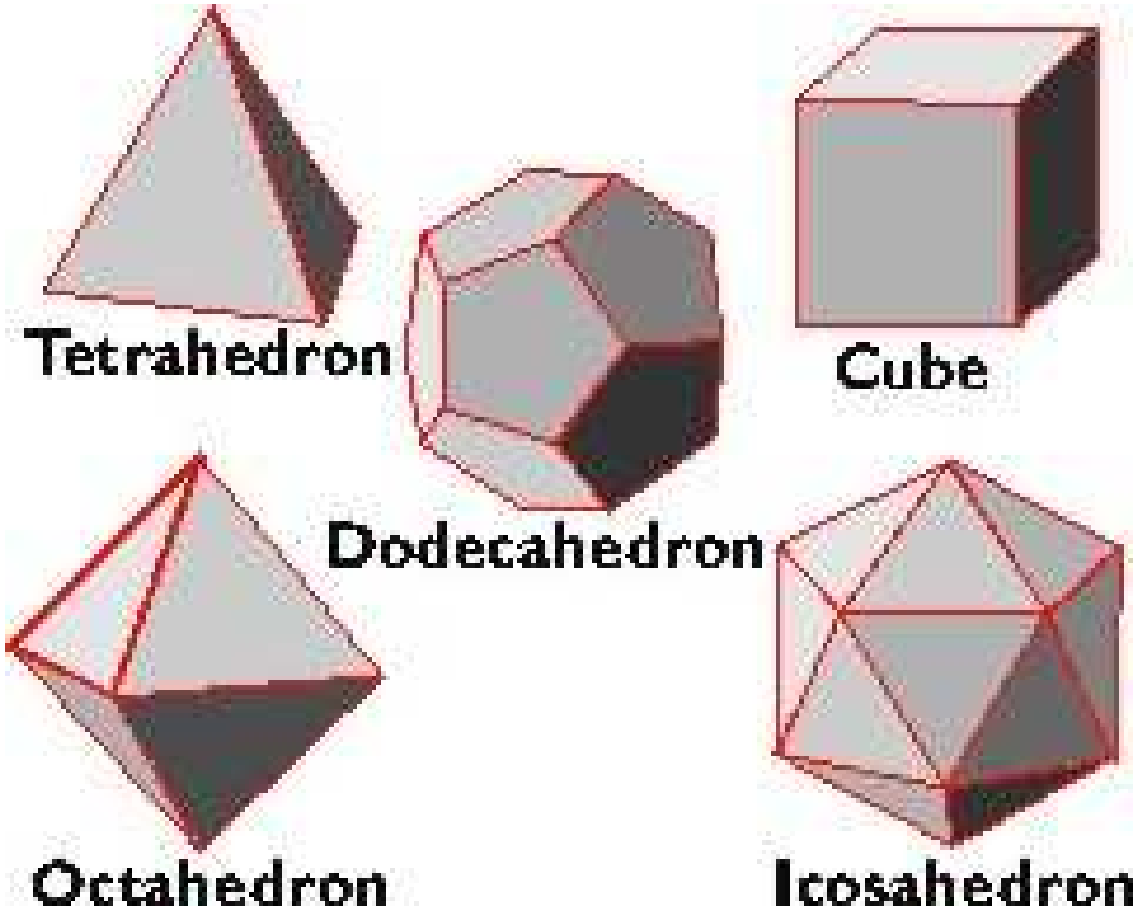
1 INTRODUCTION

1.1 Symmetry

This course will explore symmetry groups. We will look at examples of symmetry groups acting on various different geometric spaces. We wish to have a large variety of symmetries, so we will look at the simplest and most symmetric spaces, beginning with the Euclidean spaces $\mathbb{R}^2, \mathbb{R}^3$ and the spheres S^1, S^2. There is also a even more important example, the hyperbolic spaces, that we will look at later in the course. Each of these geometric spaces has a metric and we will study the *isometries* that preserve the distances between any pair of points.

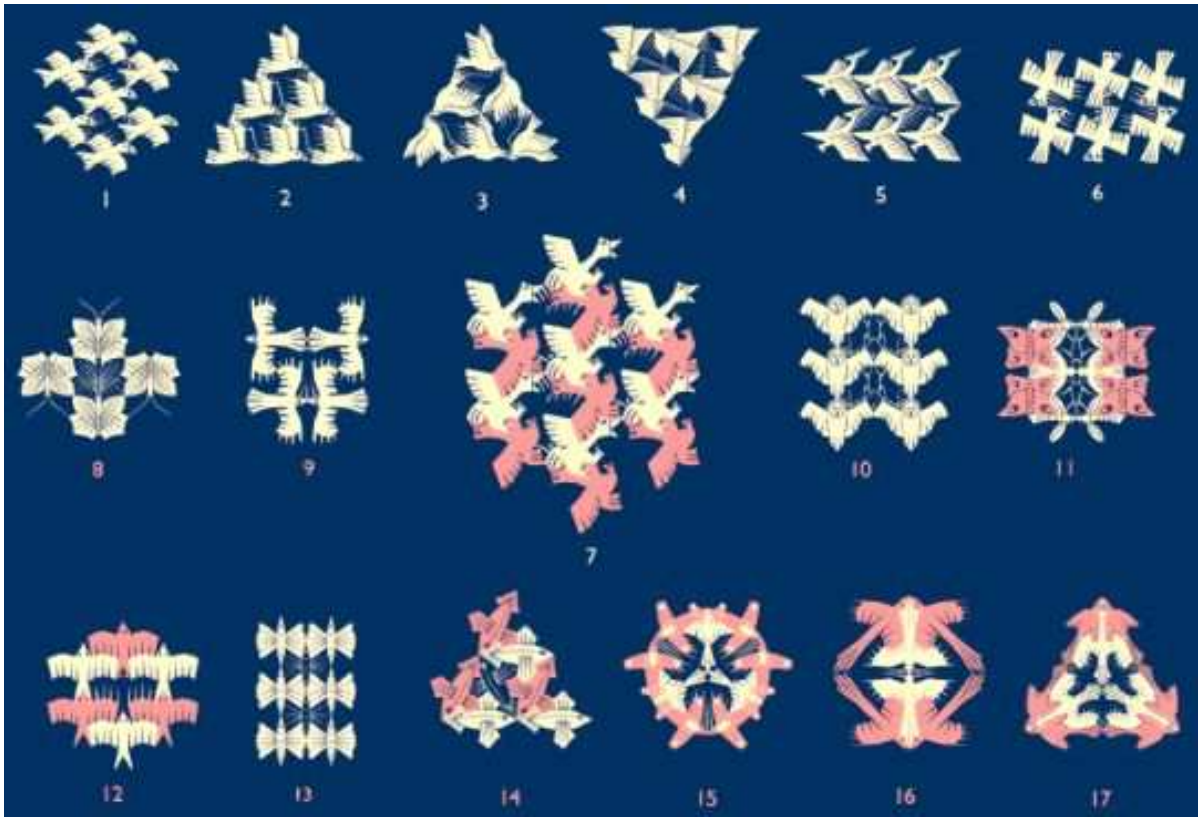
These isometries are represented by matrices, so we will usually be considering straightforward matrix groups such as $\mathrm{O}(2), \mathrm{O}(3), \mathrm{SL}(2,\mathbb{R}), \mathrm{SL}(2,\mathbb{C})$. These groups act on the geometric space and we will want to study how they do so. We will look at various patterns on the geometric spaces and the subgroups that preserve these patterns and so are symmetries of the patterns.

A first example is the "Platonic Solids":



These are the five convex regular solids. We will want to consider what makes them regular and prove that there are only five.

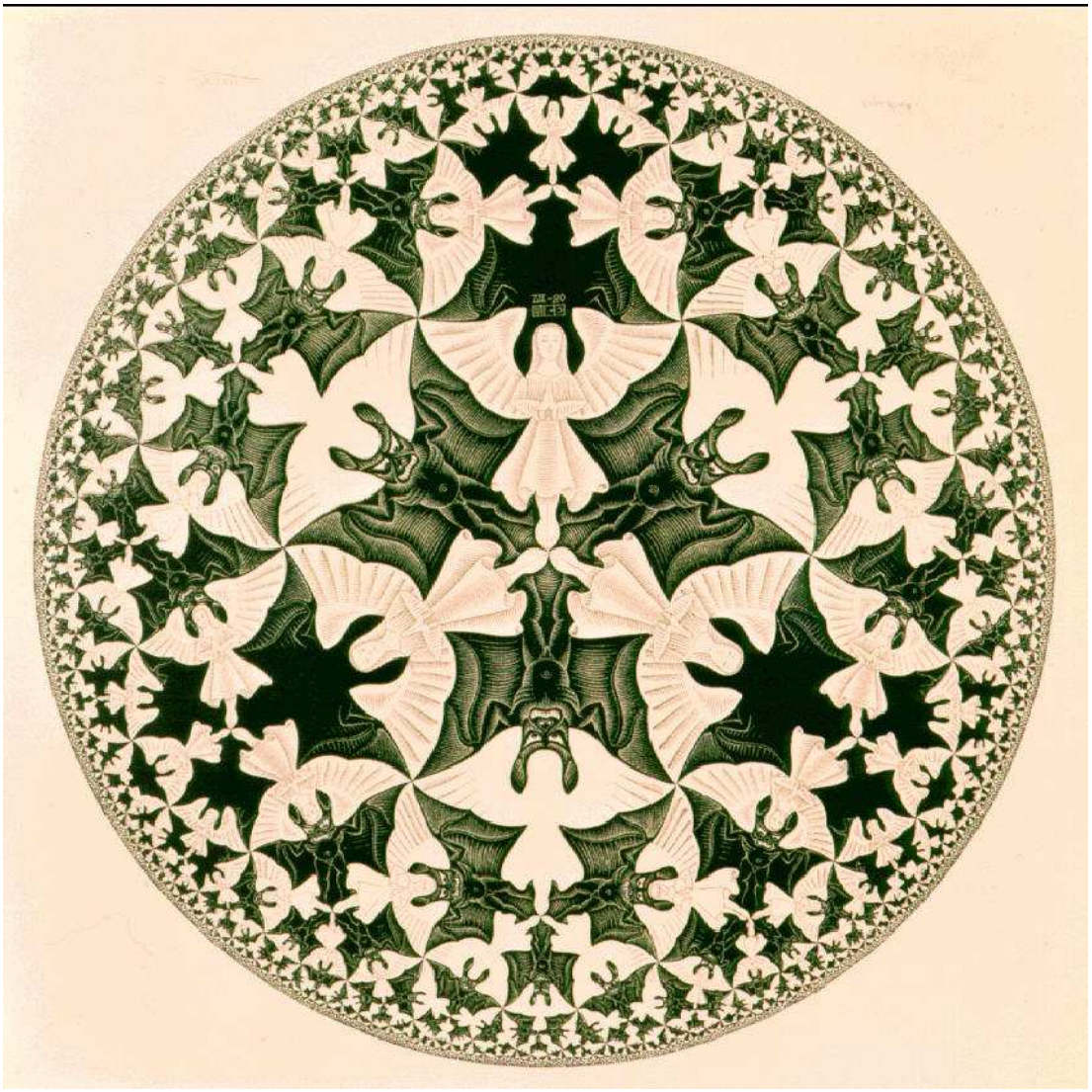
A second is the “wallpaper patterns”:



Andrew Crompton

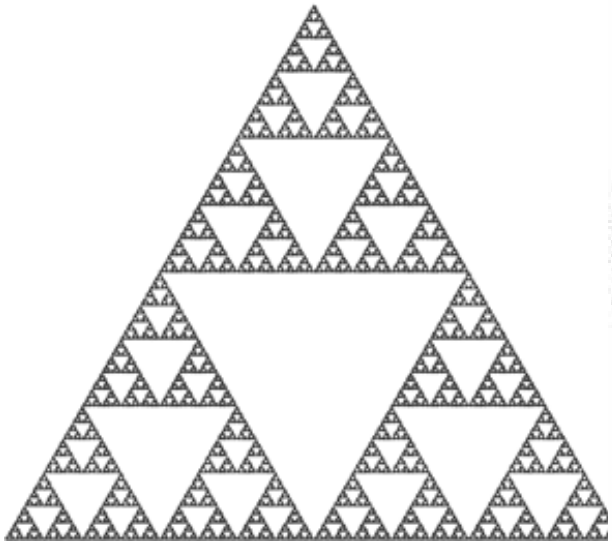
There are 17 different groups of symmetries for wallpaper patterns.

Then there are tessellations of the hyperbolic plane:

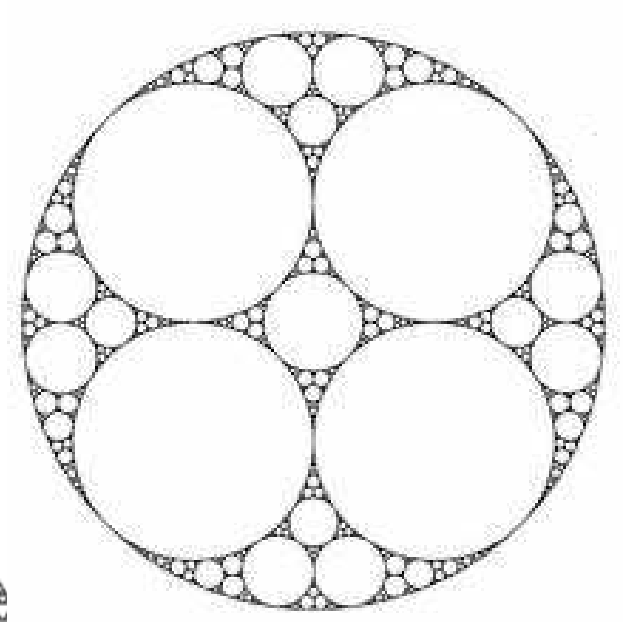


M.C. Escher, **Circle Limit IV** *(1960)*

Studying symmetry groups of the hyperbolic plane will lead us to consider limit sets which are fractal. For example:



Sierpiński Gasket



Kleinian Limit Set

and show how to calculate their dimension.

Although many of the earlier courses in the Tripos are relevant to this one, there is rather little that is required in the way of background. The most important background is from the "Vectors and Matrices" and "Groups" courses. You will need to recall the definition of a group and how it acts on a set. We will review this briefly. You also need to know about groups of matrices such as $\mathrm{GL}(N,\mathbb{R}), \mathrm{SL}(N,\mathbb{C}), \mathrm{SO}(N), \mathrm{SU}(N)$ and the group of Möbius transformations:

$$z \mapsto \frac{az+b}{cz+d} \qquad ad-bc=1$$

acting on the Riemann sphere. We will also use the notion of a metric space and compactness from the "Metric and Topological Spaces" course.

1.2 Group Actions

We will be interested in groups G that act as symmetries of a space X. So, each group element $g \in G$ gives us a symmetry $X \to X$. More formally we say that a group G *acts on a set* X if there is a map:

$$G \times X \to X \ ; \quad (g,x) \mapsto g \cdot x$$

which satisfies:

(a) $e \cdot x = x$ for the identity e of G and any point $x \in X$;

(b) $g \cdot (h \cdot x) = (gh) \cdot x$ for $g, h \in G$ and any point $x \in X$.

For example, the group $\mathrm{GL}(n,\mathbb{R})$ of invertible $n \times n$ real matrices acts on $\mathbb{R}^n$ by matrix multiplication: $(M, \boldsymbol{x}) \mapsto M\boldsymbol{x}$. The group $\mathrm{SL}(2,\mathbb{C})$ of 2×2 complex matrices with determinant 1 acts on the Riemann sphere by:

$$\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}, z \right) \mapsto \frac{az+b}{cz+d} \ .$$

Note that the set $\mathrm{Bij}(X)$ of all bijections from X to itself forms a group under composition. If the group G acts on X then the map $\theta : G \to \mathrm{Bij}(X)$ with

$$\theta(g) : X \to X \; ; \quad x \mapsto g \cdot x$$

is a group homomorphism. For condition (a) shows that $\theta(e) = I_X$ and condition (b) shows that $\theta(g) \circ \theta(h)x = \theta(gh)x$.

Exercise:

1. Show that, for any group homomorphism $\theta : G \to \mathrm{Bij}(X)$, the group G acts on X by

$$(g, x) \mapsto \theta(g)x \; .$$

The group action is *faithful* or *effective* when $\theta : G \to \mathrm{Bij}(X)$ is injective.

Suppose that the group G acts on the set X. For each $x \in X$, the *orbit* $\mathrm{Orb}(x)$ is the set $\{g \cdot x : g \in G\}$ of points that x is mapped to by G. It is a subset of the space X. The *stabilizer* $\mathrm{Stab}(x)$ is $\{g \in G : g \cdot x = x\}$. This is a subgroup of G.

Proposition 1.1 Orbit – Stabilizer theorem
If a group G acts on a set X, then the map

$$\alpha : G/\mathrm{Stab}(x) \to \mathrm{Orb}(x) \; ; \quad g\mathrm{Stab}(x) \mapsto g \cdot x$$

is a bijection. When G is a finite group, this shows that

$$|G| = |\mathrm{Stab}(x)||\mathrm{Orb}(x)|$$

for each $x \in X$.

Proof:

Note that

$$\begin{aligned} g \cdot x = h \cdot x &\Leftrightarrow (h^{-1}g) \cdot x = x \\ &\Leftrightarrow h^{-1}g \in \mathrm{Stab}(x) \\ &\Leftrightarrow g \in h\mathrm{Stab}(x) \end{aligned}$$

So α maps each coset $g\mathrm{Stab}(x)$ onto the element $g \cdot x$ in the orbit of x. Consequently, α is well defined and a bijection. □

We can also consider the quotient of X by the group action. The relation

$$x \sim y \Leftrightarrow y = g \cdot x \;\; \text{for some } g \in G$$

is an equivalence relation on X with the orbits as the equivalence classes. The *quotient* X/G is the set of equivalence classes. A *fundamental set* is a subset F of X that contains exactly one element from each orbit.

For example, let $\mathbb{Z}$ be the additive group of integers acting on the plane $\mathbb{R}^2$ by translations:

$$n \cdot \boldsymbol{x} = \boldsymbol{x} + n\boldsymbol{i} \; .$$

The stabiliser of each point $\boldsymbol{x} \in \mathbb{R}^2$ is the identity alone, while the orbit of $\boldsymbol{x}$ is the set $\{\boldsymbol{x} + n\boldsymbol{i} : n \in \mathbb{Z}\}$ of all translations of $\boldsymbol{x}$ by integer multiples of the unit vector $\boldsymbol{i}$. Let F be the strip

$$\left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 : 0 \leqslant x_1 < 1 \right\} \; .$$

Each point $\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in \mathbb{R}^2$ is equivalent to a point $\begin{pmatrix} \lfloor y_1 \rfloor \\ y_2 \end{pmatrix}$ in F, so F is a fundamental set. Two points in closed strip $\overline{F}$ are only equivalent to one another when they are points

$$\begin{pmatrix} 0 \\ y_2 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 \\ y_2 \end{pmatrix}$$

on the opposite edges. If we identify these points on the opposite edges then we obtain a cylinder. Each orbit corresponds to a unique point of this cylinder, so we can identify this cylinder with the quotient X/G.

Exercise:

2. Show that additive the group $\mathbb{Z} \times \mathbb{Z}$ acts on the plane $\mathbb{R}^2$ by

$$\begin{pmatrix} n_1 \\ n_2 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 + n_1 \\ x_2 + n_2 \end{pmatrix}$$

   and that the unit square $S = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : 0 \leqslant x_1 < 1 \text{ and } 0 \leqslant x_2 < 1 \right\}$ is a fundamental set. Hence show that we can identify the quotient $\mathbb{R}^2/\mathbb{Z} \times \mathbb{Z}$ with a torus.
   Let $\boldsymbol{u}, \boldsymbol{v}$ be two vectors in $\mathbb{R}^2$ and let P be the parallelogram:

$$\{\lambda \boldsymbol{u} + \mu \boldsymbol{v} : 0 \leqslant \lambda < 1 \text{ and } 0 \leqslant \mu < 1\} \ .$$

   Suppose that P is also a fundamental set for the action of $\mathbb{Z} \times \mathbb{Z}$ on $\mathbb{R}^2$. Show that

$$\boldsymbol{u} = \begin{pmatrix} a \\ c \end{pmatrix} \ , \quad \boldsymbol{v} = \begin{pmatrix} b \\ d \end{pmatrix}$$

   for some **integers** a, b, c, d with $ad - bc = \pm 1$.

3. Consider the two maps:

$$A : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1 \\ x_2 + 1 \end{pmatrix} \ ; \quad B : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1 + 1 \\ -x_2 \end{pmatrix}$$

   acting on the plane $\mathbb{R}^2$. Let G be the group they generate. Is G Abelian? Find the orbit of a point $\boldsymbol{x}$ under this group. Find a fundamental set and hence describe the quotient $\mathbb{R}^2/G$.

---

2 ISOMETRIES OF EUCLIDEAN SPACE

2.1 Definitions

Let M be a metric space with d as its metric. A map $T : M \to M$ is an *isometry* if it is invertible and preserves distances, so

$$d(T(x), T(y)) = d(x, y) \qquad \text{for all } x, y \in M \ .$$

The set of isometries of M form a group $\mathrm{Isom}(M)$ under composition.

In this section we want to study the isometries of the Euclidean space $\mathbb{E}^N$ and especially of the Euclidean plane. Euclidean space consists of $\mathbb{R}^N$ with the Euclidean metric, which is defined in terms of the inner product (or scalar product). The inner product of two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$ is given by

$$\boldsymbol{x} \cdot \boldsymbol{y} = x_1 y_1 + x_2 y_2 + \ldots + x_N y_N \ .$$

We can use this to define the *norm* or *length* of a vector $\boldsymbol{x}$ as

$$||\boldsymbol{x}|| = \sqrt{(\boldsymbol{x} \cdot \boldsymbol{x})} = \sqrt{(x_1 x_1 + x_2 x_2 + \ldots + x_N x_N)} \ .$$

The *Euclidean metric* is given by

$$d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}|| \ .$$

For any vector $\boldsymbol{a} \in \mathbb{R}^N$ the map $T : \boldsymbol{x} \mapsto \boldsymbol{x} + \boldsymbol{a}$ is clearly an isometry and is called the *translation by* $\boldsymbol{a}$. Let $\boldsymbol{u}$ be a unit vector. For each real number λ, the set $\pi = \{\boldsymbol{x} : \boldsymbol{x} \cdot \boldsymbol{u} = \lambda\}$ is a translation of a vector subspace of dimension $N - 1$. This is called a *hyperplane*. Any vector $\boldsymbol{x}$ can be written as $(\boldsymbol{x} \cdot \boldsymbol{u})\boldsymbol{u} + \boldsymbol{x}^\perp$ with $\boldsymbol{x}^\perp$ perpendicular to $\boldsymbol{u}$. Then reflection in π should leave $\boldsymbol{x}^\perp$ unaltered but map $t\boldsymbol{u}$ to $(2\lambda - t)\boldsymbol{u}$. Hence we define *reflection in the hyperplane* π to be

$$R : \boldsymbol{x} \mapsto \boldsymbol{x} - 2(\boldsymbol{x} \cdot \boldsymbol{u} - \lambda)\boldsymbol{u} \ .$$

It is straightforward to check that

$$||R(\boldsymbol{x}) - R(\boldsymbol{y})|| = ||(\boldsymbol{x} - \boldsymbol{y}) - 2((\boldsymbol{x} - \boldsymbol{y}) \cdot \boldsymbol{u})\boldsymbol{u}|| = ||\boldsymbol{x} - \boldsymbol{y}||$$

so the reflection is an Euclidean isometry.

We wish to describe all of the isometries of Euclidean space. It is easiest to do this by first considering those isometries that fix the origin. Our first lemma shows that any such isometry is an orthogonal linear map.

Recall that a linear map $T : \mathbb{R}^N \to \mathbb{R}^N$ is *orthogonal* when it preserves the inner product, so

$$T(\boldsymbol{x}) \cdot T(\boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y} \qquad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N \ .$$

Since the inner product is given by $\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x}^t \boldsymbol{y}$, this is equivalent to the $N \times N$-matrix M that represents T satisfying

$$\boldsymbol{x}^t M^t M \boldsymbol{y} = \boldsymbol{x}^t \boldsymbol{y} \qquad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N \ .$$

Hence, $M^t M = I$. These matrices give us the *orthogonal group*

$$\mathrm{O}(N) = \{M : M \text{ is an } N \times N \text{ real matrix with } M^t M = I\} \ .$$

Lemma 2.1 Orthogonal maps as Euclidean isometries
Every orthogonal linear map $B : \mathbb{R}^N \to \mathbb{R}^N$ is an isometry for the Euclidean metric. Conversely, if $B : \mathbb{R}^N \to \mathbb{R}^N$ is a Euclidean isometry that fixes the origin, the B is an orthogonal linear map.

Proof:

Let B be an orthogonal linear map. Then

$$d(B\boldsymbol{x}, B\boldsymbol{y})^2 = ||B\boldsymbol{x} - B\boldsymbol{y}||^2 = B(\boldsymbol{x}-\boldsymbol{y}) \cdot B(\boldsymbol{x}-\boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y}) \cdot (\boldsymbol{x}-\boldsymbol{y}) = d(\boldsymbol{x}, \boldsymbol{y})^2$$

so B is an isometry.

Now suppose that B is an isometry of $\mathbb{E}^N$ that fixes the origin. The *polarization identity:*

$$\begin{aligned} 2\boldsymbol{x} \cdot \boldsymbol{y} &= ||\boldsymbol{x}||^2 + ||\boldsymbol{y}||^2 - ||\boldsymbol{x} - \boldsymbol{y}||^2 \\ &= d(\boldsymbol{x}, \boldsymbol{0})^2 + d(\boldsymbol{y}, \boldsymbol{0})^2 - d(\boldsymbol{x}, \boldsymbol{y})^2 \end{aligned}$$

shows that

$$B\boldsymbol{x} \cdot B\boldsymbol{y} = \boldsymbol{x} \cdot \boldsymbol{y}$$

for any vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$. This shows that B preserves the inner product. We need to show that it is linear.

Let $(\boldsymbol{e}_n)$ be the standard orthonormal basis for $\mathbb{R}^N$. Because B preserves the inner product, $(B\boldsymbol{e}_n)$ is another orthonormal basis. For any vector $\boldsymbol{x}$ we have

$$\boldsymbol{x} = \sum (\boldsymbol{x} \cdot \boldsymbol{e}_n)\boldsymbol{e}_n \ .$$

Also,

$$\begin{aligned} B\boldsymbol{x} &= \sum (B\boldsymbol{x} \cdot B\boldsymbol{e}_n) B\boldsymbol{e}_n \\ &= \sum (\boldsymbol{x} \cdot \boldsymbol{e}_n) B\boldsymbol{e}_n \end{aligned}$$

because B preserves the inner product. Hence

$$B : \sum x_n \boldsymbol{e}_n \mapsto \sum x_n B\boldsymbol{e}_n$$

and so B is a linear map.

Consequently, B is an orthogonal linear map. □

Now let A be any isometry of Euclidean space. The translation T by the vector $A\boldsymbol{0}$ is an isometry so $B = T^{-1} \circ A$ will be an isometry that fixes the origin. Hence we obtain:

Proposition 2.2 Euclidean isometries are affine
If $A : \mathbb{E}^N \to \mathbb{E}^N$ is an isometry of the Euclidean space $\mathbb{E}^N$, then there is a vector $\boldsymbol{v} \in \mathbb{R}^N$ and an orthogonal matrix B with

$$A(\boldsymbol{x}) = B\boldsymbol{x} + \boldsymbol{v} \qquad \text{for } \ \boldsymbol{x} \in \mathbb{R}^N \ .$$

Conversely, this map is an isometry for any vector $\boldsymbol{v}$ and any orthogonal matrix B.

Proof:

Let A be an isometry of $\mathbb{E}^N$ and set $\boldsymbol{v} = A\boldsymbol{0}$. Then the translation T by $\boldsymbol{v}$ is an isometry and

$$B = T^{-1} \circ A : \boldsymbol{x} \mapsto A\boldsymbol{x} - A\boldsymbol{0}$$

is an isometry that fixes the origin. Hence Lemma 2.1 shows that B is an orthogonal linear map with $A\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{v}$.

Conversely, suppose that $B \in \mathrm{O}(N)$, $\boldsymbol{v} \in \mathbb{R}^N$ and $A\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{v}$. Then A is the composition of the isometries B and translation by $\boldsymbol{v}$, so it is also an isometry. □

2.2 Isometries of the Euclidean Plane

We can use the previous results to describe all of the isometries of the Euclidean plane: $\mathbb{E}^2$. First suppose that B is an isometry that fixes the origin. Then

$$B = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{O}(2) \ .$$

This means that the columns of B are unit vectors orthogonal to one another. Hence we can choose an angle $\theta \in [0, 2\pi)$ with

$$\begin{pmatrix} a \\ c \end{pmatrix} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \ .$$

The vector $\begin{pmatrix} b \\ d \end{pmatrix}$ must be a unit vector orthogonal to $\begin{pmatrix} a \\ c \end{pmatrix}$, so there are just two possibilities:

$$\begin{pmatrix} b \\ d \end{pmatrix} = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix} \qquad \text{or} \qquad \begin{pmatrix} \sin\theta \\ -\cos\theta \end{pmatrix} \ .$$

The first case gives a rotation about the origin through an angle θ (or the identity when $\theta = 0$). The second gives reflection in the line $\{(x, y) : y = (\tan\frac{1}{2}\theta)x\}$ at an angle $\frac{1}{2}\theta$ from the x-axis.

Note that the determinant of the orthogonal matrix B must be either $+1$ or -1. It is $+1$ for the identity or rotations that preserve the orientation of the plane and -1 for reflections that reverse the orientation.

If we conjugate B by a translation $T : \boldsymbol{x} \mapsto \boldsymbol{x} + \boldsymbol{p}$, then we obtain another isometry $T \circ B \circ T^{-1}$. This first translates $\boldsymbol{p}$ back to the origin, then applies B, and then translates the origin back to $\boldsymbol{p}$. When B is a rotation through an angle θ, then $T \circ B \circ T^{-1}$ is a rotation about $T\mathbf{0} = \boldsymbol{p}$ through an angle θ. When B is a reflection in a line ℓ, then $T \circ B \circ T^{-1}$ is reflection in the line $T(\ell)$.

Now consider an isometry A of the Euclidean plane that does not fix the origin. Proposition 2.2 shows that $A\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{v}$ for some $B \in \mathrm{O}(2)$ and some vector $\boldsymbol{v} \in \mathbb{R}^2$. When B is the identity, then A is a translation. When B is a rotation, we can always choose a vector $\boldsymbol{p}$ with $(I - B)\boldsymbol{p} = \boldsymbol{v}$. This means that

$$T \circ B \circ T^{-1}(\boldsymbol{x}) = B\boldsymbol{x} + (I - B)\boldsymbol{p} = B\boldsymbol{x} + \boldsymbol{v}$$

so the isometry A is a rotation about the point $\boldsymbol{p}$. When B is a reflection in a line ℓ through the origin, then we can split the vector $\boldsymbol{v}$ into a part $\boldsymbol{v}_1$ perpendicular to ℓ and a part $\boldsymbol{v}_2$ parallel to ℓ. The linear map $I - B$ maps onto the vector subspace of vectors perpendicular to ℓ, so we can choose a vector $\boldsymbol{p}$ with $(I - B)\boldsymbol{p} = \boldsymbol{v}_1$. This means that

$$T \circ B \circ T^{-1}(\boldsymbol{x}) = B\boldsymbol{x} + (I - B)\boldsymbol{p} = B\boldsymbol{x} + \boldsymbol{v}_1$$

so $A\boldsymbol{x} = T \circ B \circ T^{-1}(\boldsymbol{x}) + \boldsymbol{v}_2$. When $\boldsymbol{v}_2 = \mathbf{0}$, this shows that A is reflection in the line ℓ translated by $\boldsymbol{p}$. However, when $\boldsymbol{v}_2 \neq \mathbf{0}$, then A is a *glide reflection*, that is reflection in the line ℓ translated by $\boldsymbol{p}$ followed by a translation parallel to ℓ.

Proposition 2.3 Isometries of $\mathbb{E}^2$
An orientation preserving isometry of the Euclidean plane $\mathbb{E}^2$ is:

(a) The identity.

(b) A translation.

(c) A rotation about some point $\boldsymbol{c} \in \mathbb{E}^2$.

An orientation reversing isometry of $\mathbb{E}^2$ is:

(d) A reflection.

(e) A glide reflections, that is a reflection in a line ℓ followed by a translation parallel to ℓ.

□

A similar, but slightly more involved, argument gives the corresponding result for Euclidean space $\mathbb{E}^3$.

Proposition 2.4 Isometries of $\mathbb{E}^3$
An orientation preserving isometry of Euclidean 3-space $\mathbb{E}^3$ is:

(a) The identity.

(b) A translation.

(c) A rotation about some line ℓ.

(d) A screw rotation, that is a rotation about some line ℓ followed by a translation parallel to ℓ.

An orientation reversing isometry of $\mathbb{E}^3$ is:

(e) A reflection in some plane Π.

(f) A glide reflection, that is a reflection in a plane Π followed by a translation parallel to Π.

(g) A rotatory reflection, that is a rotation about some axis ℓ followed by reflection in a plane perpendicular to ℓ.

□

3 THE ISOMETRY GROUP OF EUCLIDEAN SPACE

3.1 Quotients of the Isometry group

Recall from Proposition 2.2 that every isometry A of $\mathbb{E}^N$ can be written as

$$A : \boldsymbol{x} \mapsto B\boldsymbol{x} + \boldsymbol{v}$$

with $B \in \mathrm{O}(N)$ and $\boldsymbol{v} \in \mathbb{R}^N$. Hence we can define a map

$$\phi : \mathrm{Isom}(\mathbb{E}^N) \to \mathrm{O}(N) \qquad \text{by} \qquad A \mapsto B \ .$$

Proposition 3.1
The map $\phi : \mathrm{Isom}(\mathbb{E}^N) \to \mathrm{O}(N)$ given above is a group homomorphism with kernel equal to the group $\mathrm{Trans}(\mathbb{E}^N)$ of all translations of $\mathbb{E}^N$.

Proof:
Let A_1, A_2 be two isometries with $A_k\boldsymbol{x} = B_k\boldsymbol{x} + \boldsymbol{v}_k$. Then

$$A_2 \circ A_1(\boldsymbol{x}) = A_2(B_1\boldsymbol{x} + \boldsymbol{v}_1) = B_2(B_1\boldsymbol{x} + \boldsymbol{v}_1) + \boldsymbol{v}_2 = (B_2B_1)\boldsymbol{x} + (B_2\boldsymbol{v}_1 + \boldsymbol{v}_2) \ .$$

Hence,

$$\phi(A_2 \circ A_1) = B_2B_1 = \phi(A_2)\phi(A_1)$$

which shows that ϕ is a group homomorphism.

The isometry $A : \boldsymbol{x} \mapsto B\boldsymbol{x} + \boldsymbol{v}$ is in the kernel of ϕ when $\phi(A) = B = I$. This means that A is a translation. □

The group $\mathrm{Trans}(\mathbb{E}^N)$ of translations is a normal subgroup of the isometry group, since it is the kernel of the homomorphism ϕ. This means that an isometry A acts on the translations by conjugation. If $A : \boldsymbol{x} \mapsto B\boldsymbol{x} + \boldsymbol{v}$ and T is the translation $T : \boldsymbol{x} \mapsto \boldsymbol{x} + \boldsymbol{t}$, then

$$A \circ T \circ A^{-1}(\boldsymbol{x}) = \boldsymbol{x} + B\boldsymbol{t} \ .$$

This action will be very important to us later when we look at crystallographic groups.

There is another important homomorphism from $\mathrm{Isom}(\mathbb{E}^N)$ that tells us whether an isometry preserves or reverses orientation. When $A : \boldsymbol{x} \mapsto B\boldsymbol{x} + \boldsymbol{v}$, we define $\varepsilon(A)$ to be $\det B$. Then

$$\varepsilon : \mathrm{Isom}(\mathbb{E}^N) \to \{-1, +1\}$$

is a group homomorphism. The kernel of ε is the group $\mathrm{Isom}^+(\mathbb{E}^N)$ of orientation preserving isometries of $\mathbb{E}^N$. This is a normal subgroup of $\mathrm{Isom}(\mathbb{E}^N)$.

Recall that, for any surjective group homomorphism $\alpha : G \to H$, the inverse images $\alpha^{-1}(h)$ are the cosets of $\ker\alpha$ in G. The number of cosets is equal to the number of elements in H and is called the *index* of $\ker\alpha$ in G. Since $\varepsilon : \mathrm{Isom}(\mathbb{E}^N) \to \{-1, +1\}$ is a group homomorphism onto $\{-1, +1\}$, the subgroup $\mathrm{Isom}^+(\mathbb{E}^N)$ has index 2 in $\mathrm{Isom}(\mathbb{E}^N)$. This means that it has just two cosets $\mathrm{Isom}^+(\mathbb{E}^N) = \varepsilon^{-1}(+1)$ and the complement $\mathrm{Isom}^-(\mathbb{E}^N) = \varepsilon^{-1}(-1)$. For any orientation reversing isometry J, this complement is equal to the coset $J\mathrm{Isom}^+(\mathbb{E}^N)$.

If G is any subgroup of $\mathrm{Isom}(\mathbb{E}^N)$, then the restriction

$$\varepsilon|_G : G \to \{-1, +1\}$$

is a group homomorphism. Let $G^+ = G \cap \mathrm{Isom}^+(\mathbb{E}^N)$ be the orientation preserving isometries in G. These form the normal subgroup $\ker\varepsilon|_G$ of G. The image of $\varepsilon|_G$ is either $\{+1\}$ or $\{-1, +1\}$. Consequently, either $G = G^+$ contains only orientation preserving isometries or else G^+ is of index 2 in G. When we study subgroups like G we often begin by looking at G^+ and then consider how we can add orientation reversing isometries.

3.2 Matrices for Isometries

The Euclidean space $\mathbb{E}^N$ is the same set as $\mathbb{R}^N$ but it has different properties. In the vector space $\mathbb{R}^N$ the origin $\mathbf{0}$ is a special point. However, in Euclidean space it is not since translations can be used to send $\mathbf{0}$ to any other point. Rather than thinking of $\mathbb{E}^N$ as equal to $\mathbb{R}^N$, it is often more convenient to identify it with a hyperplane in $\mathbb{R}^{N+1}$ that does not go through the origin. Then we can represent every isometry as an $(N+1)\times(N+1)$ matrix.

Set $\mathbb{E}^N$ to be the hyperplane $\{\boldsymbol{y}=(y_n)_{n=1}^{N+1}\in\mathbb{R}^{N+1} : y_{N+1}=1\}$in $\mathbb{R}^{N+1}$. Any vector $\boldsymbol{x}\in\mathbb{R}^N$ then corresponds to a point $\begin{pmatrix}\boldsymbol{x}\\1\end{pmatrix}\in\mathbb{E}^N$. The Euclidean metric on $\mathbb{E}^N$ is just the restriction of the usual metric on $\mathbb{R}^{N+1}$:

$$d(\boldsymbol{y},\boldsymbol{z})=||\boldsymbol{y}-\boldsymbol{z}|| \ .$$

Let A be an isometry of $\mathbb{E}^N$. Then Proposition 2.2 shows that A maps a vector $\boldsymbol{x}\in\mathbb{R}^N$ to the vector $B\boldsymbol{x}+\boldsymbol{v}$ for some $B\in\mathrm{O}(N)$ and $\boldsymbol{v}\in\mathbb{R}^N$. When we think of $\mathbb{E}^N$ as the hyperplane $\{\boldsymbol{y} : y_{N+1}=1\}$ we see that A sends the vector

$$\begin{pmatrix}\boldsymbol{x}\\1\end{pmatrix} \qquad \text{to} \qquad \begin{pmatrix}B & \boldsymbol{v}\\ \mathbf{0} & 1\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\1\end{pmatrix}=\begin{pmatrix}B\boldsymbol{x}+\boldsymbol{v}\\1\end{pmatrix} \ .$$

So A is represented by the $(N+1)\times(N+1)$ matrix

$$\begin{pmatrix}B & \boldsymbol{v}\\ \mathbf{0} & 1\end{pmatrix} \ .$$

Exercise:

4. Show that an $(N+1)\times(N+1)$ matrix maps the hyperplane $\mathbb{E}^N$ isometrically onto itself if and only if

$$M=\begin{pmatrix}B & \boldsymbol{v}\\ \mathbf{0} & 1\end{pmatrix}$$

for some matrix $B\in\mathrm{O}(N)$ and some vector $\boldsymbol{v}\in\mathbb{R}^N$.

This means that we can regard the group $\mathrm{Isom}(\mathbb{E}^N)$ as a group of matrices. This can make computations simpler and it also gives us a natural metric on the group. For the $(N+1)\times(N+1)$ matrices are a vector space of dimension $(N+1)^2$ and have a inner product given by

$$K\cdot M=\operatorname{tr}K^tM=\sum_{i=1}^{N+1}\sum_{j=1}^{N+1}k_{ij}m_{ij} \ .$$

We can then define the distance between two matrices K and M as

$$d(K,M)^2=(K-M)\cdot(K-M)=\sum_{i,j}(k_{ij}-m_{ij})^2 \ .$$

This metric behaves as you would expect. A sequence $(M(k))$ of matrices converges to a limit matrix M if and only if the sequence $(M(k)_{ij})$ of ij entries converges to M_{ij} for each pair ij.

3.3 Finite Groups of Isometries of the Plane

In Proposition 2.2 we showed that each isometry A of $\mathbb{E}^N$ is of the form $\boldsymbol{x} \mapsto B\boldsymbol{x} + \boldsymbol{v}$ with $B \in \mathrm{O}(N)$, $\boldsymbol{v} \in \mathbb{R}^N$. This means that A is affine, so

$$A\left(\sum \lambda_j \boldsymbol{x}_j\right) = \sum \lambda_j A(\boldsymbol{x}_j) \qquad \text{provided that} \qquad \sum \lambda_j = 1 \ .$$

Let G be a finite subgroup of $\mathrm{Isom}(\mathbb{E}^N)$. Choose any point $\boldsymbol{a} \in \mathbb{E}^N$. The centroid of the orbit of $\boldsymbol{a}$:

$$\boldsymbol{c} = \frac{1}{|G|} \sum_{T \in G} T(\boldsymbol{a})$$

then satisfies $A(\boldsymbol{c}) = \boldsymbol{c}$ for each $A \in G$. Therefore all of the elements of G fix the point $\boldsymbol{c}$.

Proposition 3.2 Finite subgroups of $\mathrm{Isom}(\mathbb{E}^2)$.
A finite subgroup G of $\mathrm{Isom}(\mathbb{E}^2)$ is either a cyclic group consisting of N rotations through angles $2\pi k/N$ $(k = 0, 1, 2, \ldots, N-1)$ about some point $\boldsymbol{c}$, or else a dihedral group consisting of N rotations through angles $2\pi k/N$ $(k = 0, 1, 2, \ldots, N-1)$ about some point $\boldsymbol{c}$ and N reflections in lines through $\boldsymbol{c}$.

Proof:

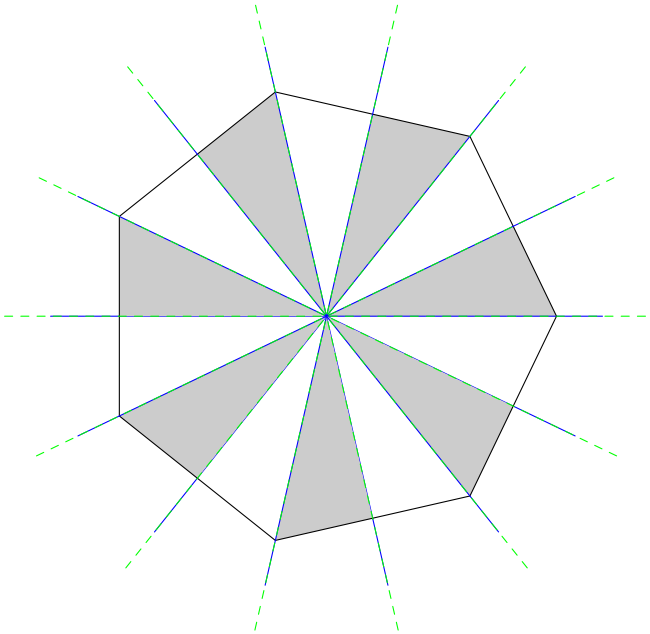
The centroid $\boldsymbol{c}$ is fixed by all of the isometries in G.

The subgroup $G^+ = G \cap \mathrm{Isom}^+(\mathbb{E}^2)$ is also a finite group, with order N say. Then each transformation $A \in G^+$ is a rotation about $\boldsymbol{c}$ with $A^N = I$. Hence A must be a rotation through an angle $2\pi k/N$ for some integer $k \in \{0, 1, 2, \ldots, N-1\}$. There are only N such rotations about $\boldsymbol{c}$ so all of them must lie in G^+. Hence G^+ must be the cyclic group of order N generated by a rotation R about $\boldsymbol{c}$ through an angle $2\pi/N$.

If G consists only of orientation preserving isometries, then $G = G^+$ is cyclic. Otherwise, there must be an orientation reversing isometry M in $G \setminus G^+$. This fixes $\boldsymbol{c}$ so it must be a reflection in a line ℓ through $\boldsymbol{c}$. The homomorphism $\varepsilon : G \to \{-1, +1\}$ maps G^+ onto $+1$ and the coset G^+M onto -1, so $|G| = 2N$.

The products $M, RM, R^2M, \ldots, R^{N-1}M$ are all distinct and are reflections in the line obtained by rotating ℓ about $\boldsymbol{c}$ through angles $\pi k/N$ for $k = 0, 1, 2, \ldots, N-1$ respectively. So we see that G is dihedral of order $2N$. □

The groups described in this proposition are the symmetries of a regular N-gon centred on $\boldsymbol{c}$, either the orientation preserving isometries for the cyclic group or all the isometries for the dihedral group. This is illustrated in the picture below. The orientation preserving isometries permute the shaded regions while the orientation reversing isometries interchange the shaded and unshaded regions.



Regular polygon of order $N = 7$.

3.4 Compositions of Reflections

We wish to find the result of composing two reflections. The simplest way to do this is to choose co-ordinates so that one of the reflections is, say, reflection in the x-axis and then simply use matrices. You should do this. We will adopt a different approach.

Let M be the reflection of $\mathbb{E}^2$ in the line ℓ. Let T be a translation perpendicular to ℓ. Then the conjugate MTM^{-1} is translation in the reverse direction, that is T^{-1}. So $MTM^{-1} = T^{-1}$ and hence $MT = T^{-1}M$. Similarly, if R is a rotation about a point on ℓ, then the conjugate MRM^{-1} is rotation about the same point in the opposite direction, so $MRM^{-1} = R^{-1}$.

Now suppose that M' is reflection in a second line ℓ'. If the two lines ℓ and ℓ' are parallel, then we can find a translation T perpendicular to both line that maps ℓ onto ℓ'. Hence $M' = TMT^{-1}$. Consequently

$$M'M = TMT^{-1}M = TMMT = T^2 \ .$$

So the composition of two reflections in parallel lines is a translation perpendicular to those lines by twice the distance between them.

If the two lines ℓ and ℓ' meet at a point P, then there is a rotation R about P that maps ℓ onto ℓ'. Hence $M' = RMR^{-1}$. Consequently

$$M'M = RMR^{-1}M = RMMR = R^2 \ .$$

So the composition of the two reflections is a rotation about the point P of intersection through twice the angle from ℓ to ℓ'

4 FINITE SYMMETRY GROUPS OF EUCLIDEAN SPACE

If P is an object in $\mathbb{E}^3$, then the isometries of $\mathbb{E}^3$ that are symmetries of P form a subgroup. By choosing P to be highly symmetric, such as one of the Platonic solids, we obtain a non-trivial, finite group of symmetries. The aim of this lecture is to prove that these give us all of the finite subgroups of $\mathrm{Isom}(\mathbb{E}^3)$. We will begin by looking only at the orientation preserving symmetries.

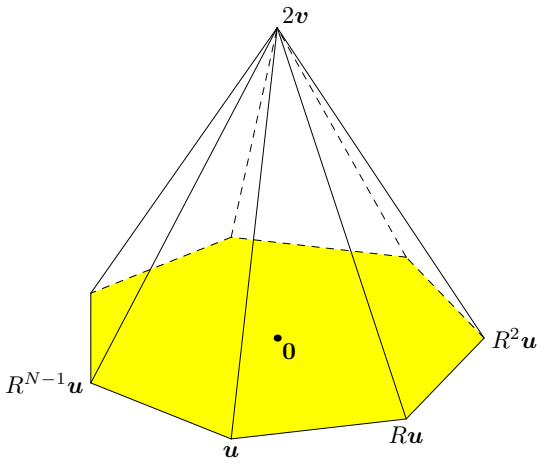
Recall from §3 that every finite subgroup G of $\mathrm{Isom}(\mathbb{E}^N)$ must fix a point $\boldsymbol{c} \in \mathbb{E}^N$. By translating, we can ensure that this point $\boldsymbol{c}$ is at the origin. Then each symmetry $A \in G$ must be in $\mathrm{O}(N)$. This means that A maps the unit sphere $S^{N-1} = \{\boldsymbol{x} \in \mathbb{R}^N : ||\boldsymbol{x}|| = 1\}$ to itself isometrically. Hence we can equally well look at the finite subgroups of $\mathrm{Isom}(S^N)$.

The only orientation preserving symmetries of $\mathbb{E}^3$ that fix the origin are the identity and rotations about an axis through the origin. Hence we need only consider which rotations are symmetries.

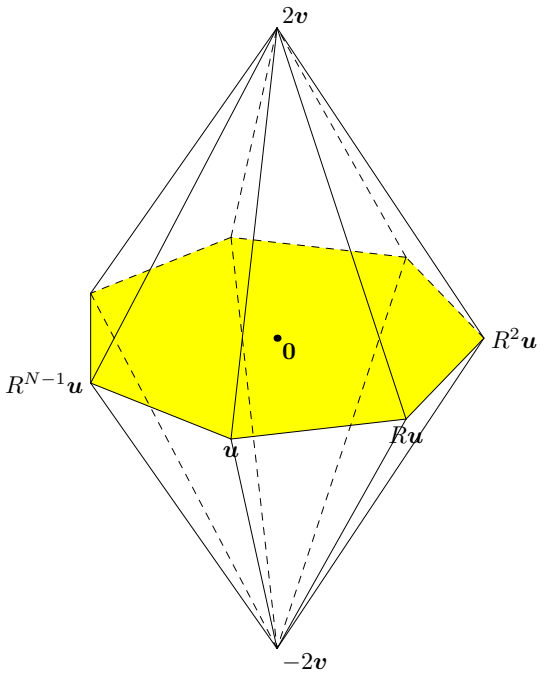
4.1 Examples of Finite Symmetry Groups

We begin by giving examples of objects with finite symmetry groups. It is usually easy to see what the symmetries are and to verify that we have them all by using the orbit – stabilizer theorem.

Let $\boldsymbol{v}$ be a unit vector in $\mathbb{R}^3$ and R the rotation about this vector through an angle $2\pi/N$ for some $N = 2, 3, 4, \ldots$. Let $\boldsymbol{u}$ be a unit vector orthogonal to $\boldsymbol{v}$. Then the points $\boldsymbol{u}, R\boldsymbol{u}, R^2\boldsymbol{u}, \ldots, R^{N-1}\boldsymbol{u}$ are the vertices of a regular N-gon P. By joining each vertex of P to $2\boldsymbol{v}$ we get the *cone* on P. By joining each vertex of P to both $2\boldsymbol{v}$ and $-2\boldsymbol{v}$ we get the *double cone* on P.



Regular cone



Regular double cone

It is clear that each of maps $I, R, R^2, \ldots, R^{N-1}$ is a symmetry of the cone or double cone on P. For the single cone, these are the only orientation preserving symmetries. For the group of symmetries acts on the vertices of the polygon P with the orbit of $\boldsymbol{u}$ being all N vertices and the stabilizer of $\boldsymbol{u}$ being only the identity. So the orbit – stabilizer theorem shows that the orientation preserving symmetry group of the cone on P has N elements and so is the cyclic group $\{I, R, R^2, \ldots, R^{N-1}\} \cong C_N$.

For the double cone on P, there are other symmetries. The rotation S about $\boldsymbol{u}$ through an angle π is one and $SR, SR^2, \ldots, SR^{N-1}$ are the others. Each of these is a rotation through an angle π. Once again, the orbit – stabilizer theorem shows that this is all of the orientation preserving symmetries.

They form a group isomorphic to the dihedral group D_{2N}. (Note that we usually think of the dihedral group as the plane symmetries of P. Then it has N reflections. The rotations through angle π act of the plane of P in the same way as the reflections but interchange the half-spaces above and below P.)

There are also orientation reversing symmetries. For the cone, there are N reflections in planes through $\boldsymbol{v}$. So the full symmetry group is isomorphic to D_{2N}. For the double cone, the reflection J in the plane of the polygon P is one symmetry. The others are $JR, JR^2, \ldots, JR^{N-1}$, which are rotatory reflections, and $JS, JSR, \ldots, JSR^{N-1}$, which are reflections in planes through $\boldsymbol{v}$. Since J commutes with all the other symmetries, we see that the full symmetry group is isomorphic to $D_{2N} \times C_2$.

We can do a similar analysis when P is one of the Platonic solids. Recall that the Platonic solids are the regular tetrahedron, octahedron, cube, dodecahedron and icosahedron. Let G be the group of symmetries of one of these, say P, centred on the origin. Then G^+ is the subgroup of orientation preserving symmetries. These groups act on the vertices. The regularity of the Platonic solids shows that there are symmetries that move any vertex to any other, so there is one orbit consisting of all the vertices. The stabilizer in G^+ consists of all the rotations that fix that vertex. These must permute the faces that meet at the vertex, so the stabilizer in G^+ is cyclic with order equal to the number of faces meeting at each vertex. The various numbers are:

	Orbit *(vertices)*	Stabilizer *(faces at each vertex)*	$\lvert G^+ \rvert$
Tetrahedron	4	3	12
Cube	8	3	24
Octahedron	6	4	24
Dodecahedron	20	3	60
Icosahedron	12	5	60

These give finite groups of orientation preserving symmetries. (We will see later that the groups are not all distinct. A Platonic solid and its dual have the same symmetries, so the groups for the cube and octahedron are the same, as are the groups for the dodecahedron and the icosahedron.) If we include the orientation reversing symmetries we obtain finite groups with twice as many elements.

We have not shown that these finite symmetry groups really exist. To do so, we would need to show that the Platonic solids exist and their symmetry groups act as claimed on the vertices. This was done in the Geometry course and various approaches to it are outlined in the example sheets.

4.2 Finite Subgroups of Isom($\mathbb{E}^3$).

Now we aim to prove that the finite groups described in the previous section are the only ones that exist. We will concentrate on the groups of orientation preserving symmetries. The groups are then cyclic of order N for $N = 1, 2, 3, \ldots$; dihedral of order $2N$; and the rotational symmetry groups of the Platonic solids, called the *tetrahedral, octahedral and icosahedral groups.*

Theorem 4.1 Finite symmetry groups in $\mathrm{Isom}^+(\mathbb{E}^3)$
Let G be a finite subgroup of $\mathrm{Isom}^+(\mathbb{E}^3)$ consisting of orientation preserving isometries. Then G is the orientation preserving symmetry group of one of the following:

(a) A cone on a regular plane polygon.
(b) A double cone on a regular plane polygon.
(c) A regular tetrahedron.
(d) A regular octahedron.
(e) A regular icosahedron.

Proof:

First translate so that G fixes the origin. Then each non-identity element of G is a rotation about an axis through $\mathbf{0}$. Let Ω be the set of unit vectors that are fixed by by some non-identity element of G. Then Ω is a finite set and G acts on it. Let $\Omega_1, \Omega_2, \ldots, \Omega_J$ be the different orbits in Ω. The orbit – stabilizer theorem shows that each vector $\boldsymbol{u} \in \Omega_j$ has a stabilizer of order $S_j = |G|/|\Omega_j|$.

Now we count the number of pairs in the set

$$X = \{(A, \boldsymbol{u}) : A \in G \setminus \{I\}, \quad \boldsymbol{u} \in S^2 \quad \text{and} \quad A\boldsymbol{u} = \boldsymbol{u}\} \ .$$

Each $A \in G \setminus \{I\}$ is a rotation and so fixes exactly two unit vectors. Therefore $|X| = 2(|G| - 1)$. Alternatively, each $\boldsymbol{u} \in \Omega$ gives rise to $|\mathrm{Stab}(\boldsymbol{u})| - 1$ pairs in X. So

$$|X| = \sum_{j=1}^{J} (S_j - 1)|\Omega_j| = \sum_{j=1}^{J} |G| - |\Omega_j| \ .$$

Dividing by $|G|$ we see that

$$2 - \frac{2}{|G|} = \sum_{j=1}^{J} 1 - \frac{1}{S_j} \ . \qquad *$$

Each stabilizer of $\boldsymbol{u} \in \Omega$ has order at least 2, so

$$1 - \frac{1}{S_j} \geqslant \frac{1}{2} \ .$$

Hence,

$$2 > 2 - \frac{2}{|G|} = \sum_{j=1}^{J} 1 - \frac{1}{S_j} \geqslant \tfrac{1}{2} J$$

and so J is $1, 2$ or 3.

Order the orbits so that $S_1 \geqslant S_2 \geqslant S_3 \geqslant 2$. Clearly there are no solutions to $(*)$ with $J = 1$. If $J = 2$ then $(*)$ gives

$$2 - \frac{2}{|G|} = 2 - \frac{1}{S_1} - \frac{1}{S_2} \leqslant 2 - \frac{2}{S_1}$$

so $S_1 \geqslant |G|$. This implies that $S_1 = |G|$ and $|\Omega_1| = 1$. Hence,

$$2 - \frac{2}{|G|} = 2 - \frac{1}{S_1} - \frac{1}{S_2} = 2 - \frac{1}{|G|} - \frac{1}{S_2}$$

and so $S_2 = |G|$ and $|\Omega_2| = 1$. Hence Ω consists of two unit vectors $\boldsymbol{v}$ and $-\boldsymbol{v}$ which are fixed by each isometry in G. The group G is then a finite isometry group of the plane orthogonal to $\boldsymbol{v}$ so it is cyclic by Proposition 3.2 . This shows that G is the symmetry group of a cone as in (a).

When $J = 3$, equation $(*)$ gives

$$\frac{1}{S_1} + \frac{1}{S_2} + \frac{1}{S_3} = 1 + \frac{2}{|G|} \ .$$

This implies that

$$\frac{3}{S_3} \geqslant 1 + \frac{2}{|G|} > 1$$

so $S_3 = 2$.

Equation $(*)$ now yields

$$\frac{2}{S_2} \geqslant \frac{1}{S_1} + \frac{1}{S_2} = \frac{1}{2} + \frac{2}{|G|} > \frac{1}{2}$$

which implies that $S_2 = 2$ or 3.

When $S_2 = 2$ we have

$$\frac{1}{S_1} = \frac{2}{|G|}$$

which gives $S_1 = N, S_2 = 2, S_3 = 2$ and $|G| = 2N$. The orbit Ω_1 has has just two points. Let $\boldsymbol{v}$ be one of them. The stabilizer $\mathrm{Stab}(\boldsymbol{v})$ is a finite group of N rotations about $\boldsymbol{v}$, so it is cyclic generated by a rotation R. Choose $\boldsymbol{u}$ as one of the points in Ω_3. Then the others are $R\boldsymbol{u}, R^2\boldsymbol{u}, \ldots, R^{N-1}\boldsymbol{u}$. The stabilizer of $\boldsymbol{u}$ has order 2, so it contains a rotation S of order 2. This maps $\boldsymbol{v}$ to $-\boldsymbol{v}$. It is now apparent that each element of G is a symmetry of a double cone as in (b).

When $S_2 = 3$ we have

$$\frac{1}{S_1} = \frac{2}{|G|} + \frac{1}{6} > \frac{1}{6}$$

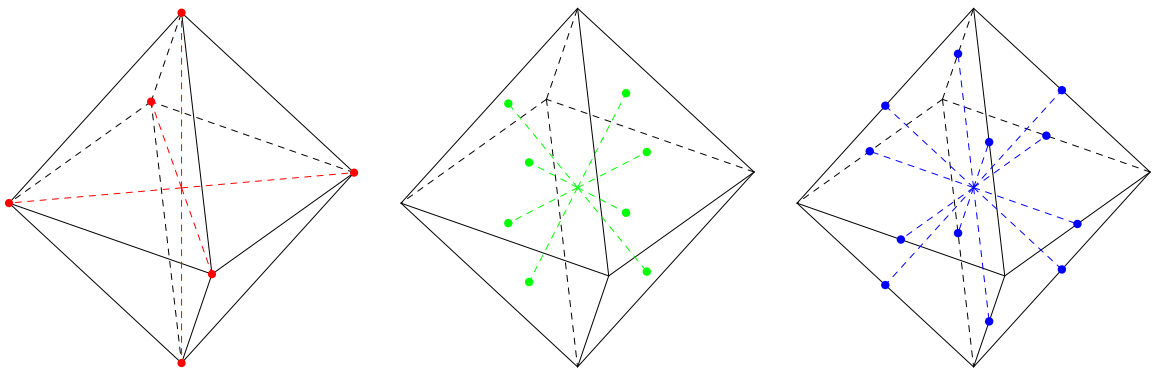
so $S_1 = 3, 4$ or 5. The possibilities are:

S_1	$\lvert\Omega_1\rvert$	S_2	$\lvert\Omega_2\rvert$	S_3	$\lvert\Omega_3\rvert$	$\lvert G\rvert$
3	4	3	4	2	6	12
4	6	3	8	2	12	24
5	12	3	20	2	30	60

We need to show that these correspond to the symmetry groups of the tetrahedron, octahedron and icosahedron respectively.

We will consider the middle case as an example. Here Ω_1 has 6 points. The stabiliser of each is a cyclic group of order $S_1 = 4$. Choose one point $\boldsymbol{v} \in \Omega_1$. The stabiliser of $\boldsymbol{v}$ is a cyclic group of order 4; let R be a generator. Now $-\boldsymbol{v}$ is also fixed by R and has the same stabiliser. So it must be in Ω_1. There remain 4 other points in Ω_1 and these must be $\boldsymbol{w}, R\boldsymbol{w}, R^2\boldsymbol{w}, R^3\boldsymbol{w}$ all lying in the plane through $\mathbf{0}$ orthogonal to $\boldsymbol{v}$. Hence the points of Ω_1 are the 6 vertices of a regular octahedron.

Note that the points of Ω_2 are the midpoints of the faces of this octahedron and the points of Ω_3 are the midpoints of the edges. The points of Ω_2 are the vertices of a cube. This is the *dual* of the octahedron. The polyhedron and its dual have the same symmetry group.



Ω_1 Ω_2 Ω_3

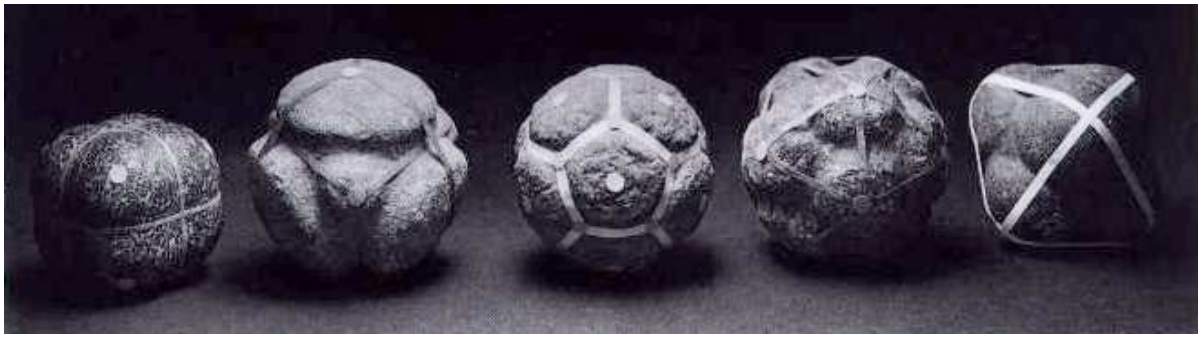
In a similar way, the first row in the table above gives us a regular tetrahedron. Ω_1 is the set of vertices; Ω_2 the centres of the faces; Ω_3 the midpoints of edges. The tetrahedron is dual to another tetrahedron.

The final row gives a icosahedron. Ω_1 is the set of vertices; Ω_2 the centres of the faces; Ω_3 the midpoints of edges. The dual is a dodecahedron. □

5 THE PLATONIC SOLIDS

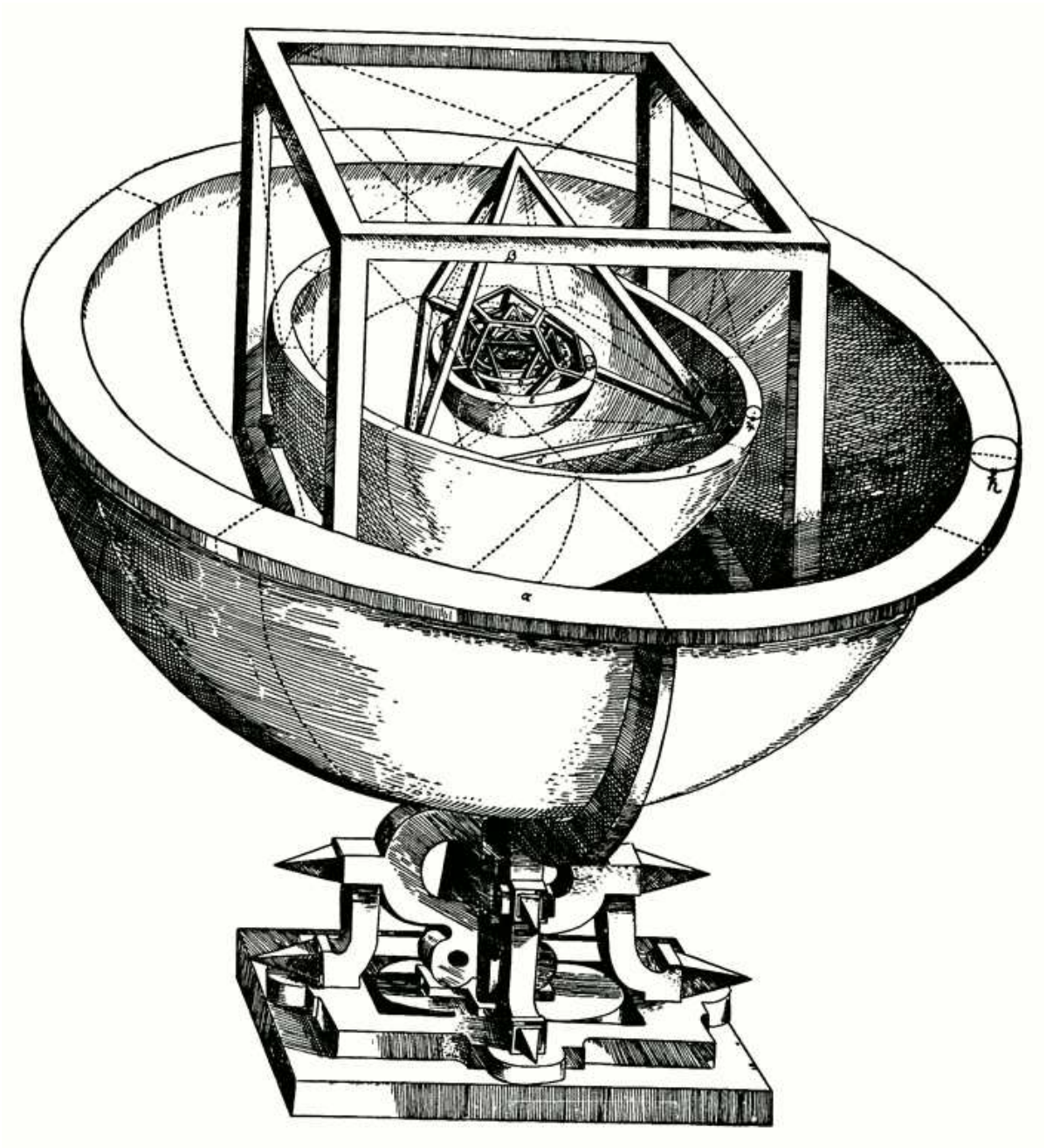
5.1 History

The Platonic solids have been known and studied for a very long time. Stones carved into polyhedral shapes date from about 2000BC in Scotland.
(See *http://www.georgehart.com/virtual-polyhedra/neolithic.html*.)



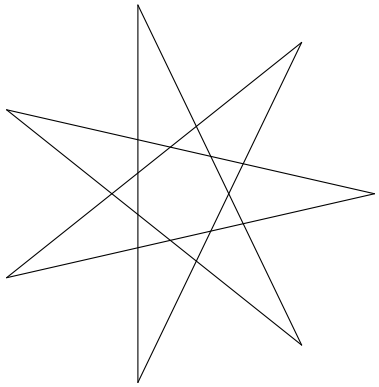
The Pythagoreans were aware of at least some of the solids and endowed them with a mystical significance. Theaetetus (c. 417BC - 369BC) was the first to prove that there were only five convex regular polyhedra. Plato refers to the solids in the *Timaeus* c. 360BC and follows the Pythagoreans in giving them mystical significance. Four of them represented the four elements: the tetrahedron for fire, the cube for earth, the octahedron for air and the icosahedron for water. This association was justified on the grounds that the icosahedron is the smoothest of the polyhedra while the tetrahedron is the sharpest. The dodecahedron represented the entire universe with the twelve faces showing the twelve signs of the zodiac. Euclid devoted the 13th and last book of his *Elements* to the Platonic solids.

In the 18th Century Kepler pursued Plato's mystical interest but also tried to use the Platonic solids to describe the known universe. In the *Mysterium Cosmographicum (1596)* he suggested that the radii of the orbits of the five known planets could be found by inscribing the solids one inside another.



5.2 Regularity

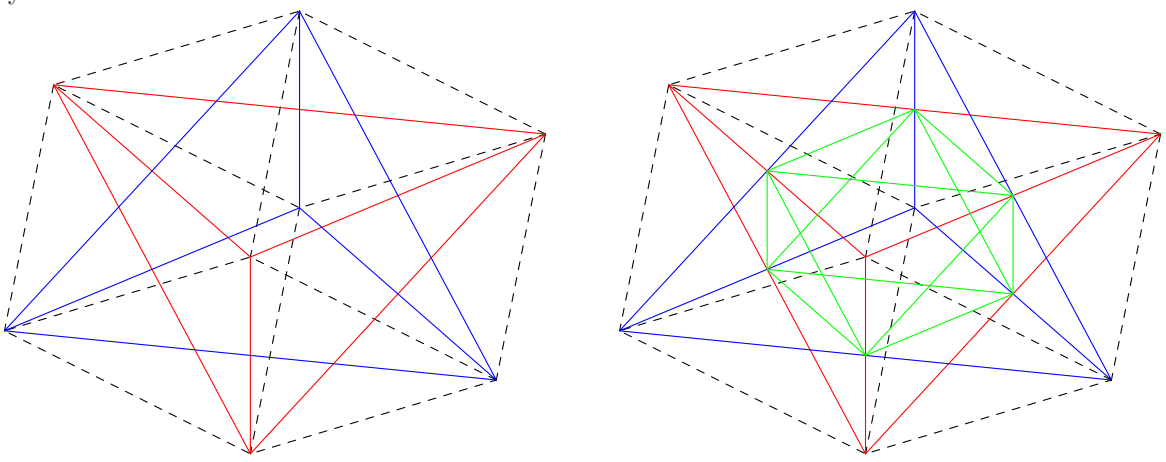
The Platonic solids are the only regular, convex polyhedra. We need to consider what regularity means. For a polygon, regularity means that each edge looks the same and each vertex looks the same. This means that there are symmetries of the polygon that map each edge to any other and each vertex to any other. Note that there are non-convex polygons that also have this property. For example:



We will consider polyhedra that have a finite number of vertices, edges and faces. For the polyhedron to be regular we want to there to be symmetries that map any vertex, edge or face to any other. A *flag* for the polyhedron P is a triple (v, e, F) consisting of a vertex v, an edge e and a face F such that v is one end of e, and e is one of the boundary edges of F. We will say that the polyhedron is *regular* if, for any two flags (v, e, F) and (v', e', F') there is a symmetry of P that maps one flag to the other.

This condition certainly implies that the group of symmetries of P acts on the set V of vertices and V is a single orbit. Once we know where a symmetry sends each vertex, we can determine where each edge and face goes and hence find the symmetry completely. This means that the symmetry group G of P is a finite group of isometries. Hence it is one of the groups found in §4. Furthermore, the group G will contain symmetries that fix a vertex v and maps any one edge at v to any other. Hence, each vertex has a non-trivial stabilizer. This means that the vertices of P must be one of the orbits $\Omega_1, \Omega_2, \Omega_3$ found in the proof of Theorem 4.1. Indeed, if we are not in the trivial situation where only two edges meet at each vertex, then the vertices can not be Ω_3 which has stabilizers of order 2. Hence the set of vertices is one of the orbits we considered in §4.

It is not, however, necessary that we join up the vertices in the expected way. See, for example, the non-convex regular polygon above. Similar possibilities arise for polyhedra. For example, consider the vertices of a cube with side length 1. We can join each vertex to the 3 other vertices at a distance $\sqrt{2}$ rather than those at unit distance. This gives two tetrahedra which together form a non-convex regular polyhedron.



Note that the intersection of these two tetrahedra is the octahedron dual to the original cube.

5.3 Convex Regular Polyhedra

Suppose that P is a convex polyhedron that is regular. The regularity certainly implies that each face is a regular polygon and an equal number of these faces meet at each vertex. Say the faces are p-gons and q meet at each vertex. Here p and q are at least 3. The pair $\{p, q\}$ is called the *Schläfi symbol* for the polyhedron.

Choose one vertex $\boldsymbol{v}$. The q vertices adjacent to $\boldsymbol{v}$ are the vertices of a regular q-gon. The angle between two edges that meet at a vertex is $\pi - 2\pi/p$. At each vertex q of these meet, so

$$q\left(\pi - \frac{2\pi}{p}\right) < 2\pi \ .$$

This simplifies to

$$(p-2)(q-2) < 4$$

so the only possible solutions are $\{3,3\}, \{3,4\}, \{4,3\}, \{3,5\}, \{5,3\}$. These give the five Platonic solids.

5.4 The Symmetry Groups

In Theorem 4.1 showed which groups could arise as the finite subgroups of $\mathrm{Isom}(\mathbb{E}^3)$. However, we have not identified which groups these are. By using some of the non-convex polyhedra with these symmetry groups, we can easily do so.

Consider first the tetrahedral group $\mathrm{Sym}(T)$. This permutes the four vertices of a tetrahedron T, and so we obtain a group homomorphism $\theta : \mathrm{Sym}(T) \to S_4$ into the symmetric group S_4 on the vertices. If a symmetry fixes all four vertices, then it is the identity. Hence θ is injective. We already know, from the orbit – stabilizer theorem, that $\mathrm{Sym}(T)$ has 24 elements, so θ must be an isomorphism. Thus the full symmetry group of the regular tetrahedron is isomorphic to S_4. By looking at the individual symmetries we see that the orientation preserving symmetries correspond to the even permutations of the vertices. So the group $\mathrm{Sym}^+(T)$ is isomorphic to the alternating group A_4.

Now consider the symmetry group of a cube centred on the origin. This is the same as the symmetry group of the dual octahedron which has vertices at the centres of each face of the cube. There are two tetrahedra embedded in this cube with vertices at the vertices of the cube, denote these by T^+ and T^-. The isometry $J\boldsymbol{x} \mapsto -\boldsymbol{x}$ interchanges T^+ and T^-. It also commutes with every other symmetry of the cube. A symmetry S of the cube either maps each tetrahedron onto itself or else interchanges the tetrahedra. In the latter case, $J \circ S$ is another symmetry that maps each tetrahedron onto itself. Hence we see that the homomorphism

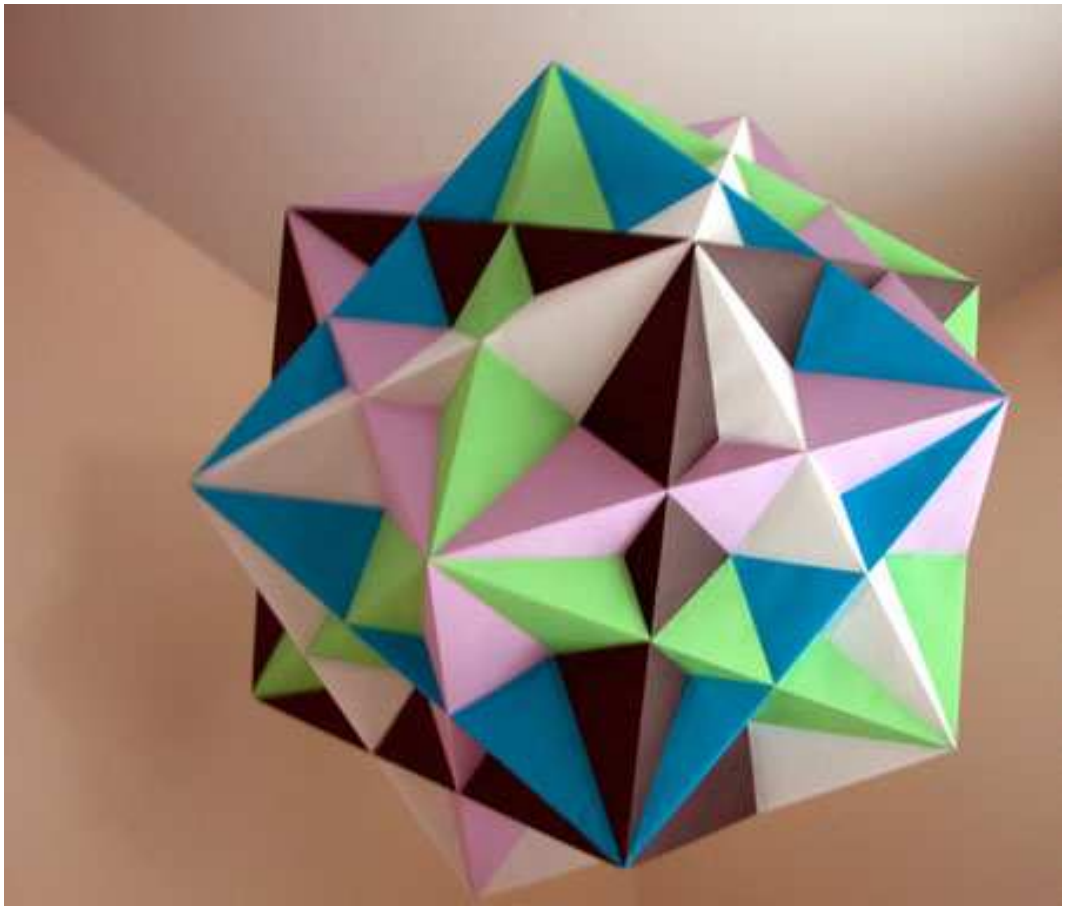
$$\theta : \mathrm{Sym}(C) \to \mathrm{Sym}(T^+) \times \{1, -1\} \ ; \quad S \mapsto \begin{cases} (S, 1) & \text{when } S(T^+) = T^+; \\ (J \circ S, -1) & \text{when } S(T^+) = T^-. \end{cases}$$

is injective. Counting elements shows that the full symmetry group $\mathrm{Sym}(C)$ is isomorphic to $S_4 \times C_2$. The subgroup $\mathrm{Sym}^+(C)$ is isomorphic to S_4. (We can also prove this by considering how a symmetry of the cube acts on the four long diagonals joining a vertex of V to its antipodal vertex.)

Finally, consider the symmetry group of the dodecahedron D, or the dual icosahedron. There are five cubes embedded inside the dodecahedron (or equivalently five octahedra within which an icosahedron is embedded). A symmetry of the dodecahedron permutes these 5 embedded cubes, so we get a group homomorphism $\theta : \mathrm{Sym}(D) \to S_5$ into the symmetric group on the cubes. By looking at each rotational symmetry of the dodecahedron we can check that θ actually maps into the alternating group A_5. Suppose that the symmetry T is in the kernel of this homomorphism. Then T maps each cube to itself. A vertex $\boldsymbol{v}$ of the dodecahedron is a vertex of exactly two of the cubes and these two cubes have only the vertices $\boldsymbol{v}$ and $J(\boldsymbol{v}) = -\boldsymbol{v}$ in common. Hence, T must map each vertex either to itself or to the antipodal vertex $J(\boldsymbol{v})$. Since T is an isometry, we must have $T = I$ or $T = J$. This shows that the map

$$\mathrm{Sym}(D) \to A_5 \times C_2 \ ; \quad T \mapsto (\theta(T), \varepsilon(T))$$

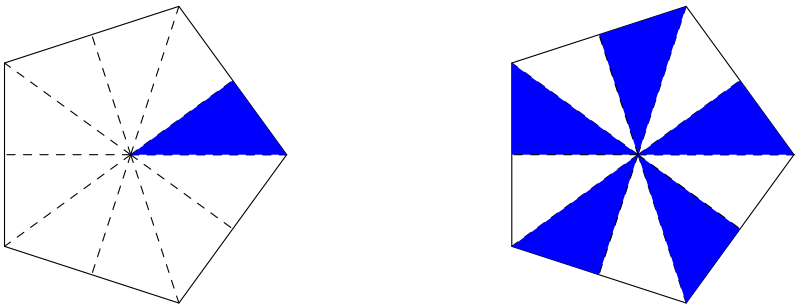
is an injective group homomorphism. The orbit – stabilizer theorem shows that there are an equal number 120 of elements in $\mathrm{Sym}(D)$ as in $A_5 \times C_2$, so this is an isomorphism. The orientation preserving symmetries of the dodecahedron form a group isomorphic to A_5.



Five cubes embedded in a regular dodecahedron

5.5 Fundamental sets

The symmetry group G of a Platonic solid P acts on the faces of P so we can look for a fundamental set. Since G permutes the faces, we can restrict our attention to one face F and look for a fundamental set of $\mathrm{Stab}(F)$. This stabilizer is the symmetry group of the polygonal face. It is easy to see that a fundamental set for this dihedral group is a closed triangle as shown in the diagram below.



The corners of this triangle are a vertex of P, the midpoint of an edge of P and the centre of the face F. The triangle thus corresponds to a flag for P. For each flag we obtain a copy of the fundamental triangle and these tessellate the surface of the polyhedron.

We can also look for a fundamental set for the symmetry group acting on all of $\mathbb{E}^3$. For this, we can take the cone on the triangle found above with its vertex at the centroid of P.

6 LATTICES

Let M be a metric space with metric d. A point $x \in M$ is *isolated* if there is some neighbourhood V of x that contains no point of M except x. This means that there is a $\delta > 0$ with

$$d(x, y) > \delta \qquad \text{for all } y \in M \setminus \{x\} \ .$$

For example, each point of $\mathbb{Z}$ is isolated. The metric space M is *discrete* if each point of M is isolated. Note that the particular metric on M is not important only the topology.

Let G be a matrix group. Then G has a natural (Euclidean) metric. We say that G is a *discrete group* if it is discrete for this metric. If the identity I is isolated in G, then every other point $T \in G$ is also isolated. For the multiplication

$$G \to G \ ; \quad A \mapsto TA$$

is continuous and has a continuous inverse $A \mapsto T^{-1}A$. Hence, to check that a group is discrete we need only check that I is isolated.

Exercise:

-5. Show that G is a discrete matrix group if and only if there is no sequence of non-identity elements elements $g_n \in G$ that converge to the identity.

For example, the group $\mathrm{SL}(2, \mathbb{Z})$ of 2×2 matrices $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with $a, b, c, d \in \mathbb{Z}$ and $ad - bc = 1$ is discrete, because any matrix $M \neq I$ satisfies

$$d(M, I)^2 = (a-1)^2 + b^2 + c^2 + (d-1)^2 \geqslant 1 \ .$$

Every finite group is certainly discrete. Discrete groups can be infinite but they can not be too large.

First we will look at the discrete groups of translations. These groups correspond to additive subgroups of $\mathbb{R}^N$. Let G be a subgroup of $\mathrm{Trans}(\mathbb{E}^N)$, then the set

$$\Lambda = \{T(\mathbf{0}) : T \in G\}$$

is an additive subgroup of $\mathbb{R}^N$. Conversely, if Λ is any additive subgroup of $\mathbb{R}^N$, then

$$G = \{\boldsymbol{x} \mapsto \boldsymbol{x} + \boldsymbol{v} : \boldsymbol{v} \in \Lambda\}$$

is a subgroup of $\mathrm{Trans}(\mathbb{E}^N)$. Note that $d(T_1, T_2) = ||\boldsymbol{v}_1 - \boldsymbol{v}_2||$, so G is a discrete group if and only if Λ is a discrete subset of $\mathbb{R}^N$. We call a discrete additive subgroup of $\mathbb{R}^N$ a *lattice* in $\mathbb{R}^N$.

For example, the set $\mathbb{Z}\boldsymbol{w}_1$ for any non-zero vector $\boldsymbol{w}_1$ is a lattice in $\mathbb{R}^2$. Similarly, $\mathbb{Z}\boldsymbol{w}_1 + \mathbb{Z}\boldsymbol{w}_2$ is a lattice in $\mathbb{R}^2$ for any two linearly independent vectors $\boldsymbol{w}_1, \boldsymbol{w}_2$. We will show that these are the only lattices in $\mathbb{R}^2$.

Proposition 6.1 Lattices in $\mathbb{R}$
Each lattice in $\mathbb{R}$ is of the form $\mathbb{Z}\boldsymbol{w}$ for some $\boldsymbol{w} \in \mathbb{R}$.

Proof:

Since Λ is discrete, there is a $\delta > 0$ with $|\lambda - 0| > \delta$ for each $\lambda \in \Lambda \setminus \{0\}$. Hence, $|\lambda_1 - \lambda_2| > \delta$ for each pair of distinct points $\lambda_1, \lambda_2 \in \Lambda$. Consequently, there can be no more than a finite number of points of Λ in any ball $B(0, r)$. This implies that either $\Lambda = \{0\}$ or else there is a point $a \in \Lambda \setminus \{0\}$ closest to 0.

In the first case we have $\Lambda = \mathbb{Z}\boldsymbol{w}$ for $\boldsymbol{w} = 0$. In the second case we will see that $\Lambda = \mathbb{Z}a$. We certainly have $a \in \Lambda$, so $\mathbb{Z}a \subset \Lambda$. Suppose that $b \in \Lambda$. Then b is a scalar multiple of a, say $b = ta$ with $t \in \mathbb{R}$. Now $t = k + t'$ with $k \in \mathbb{Z}$ and $0 \leqslant t' < 1$. So $b' = b - ka$ is also in Λ and has

$$|b'| = t'|a| < |a| \ .$$

The choice of a tells us that b' must be 0, so $b = ka \in \mathbb{Z}a$ as required. □

We can extend the argument used above to find the lattices in $\mathbb{R}^2$ and, indeed, in $\mathbb{R}^N$ for any N.

Proposition 6.2 Lattices in $\mathbb{R}^2$
Each lattice in $\mathbb{R}^2$ is either $\{\mathbf{0}\}$, or $\mathbb{Z}\boldsymbol{w}_1$, or $\mathbb{Z}\boldsymbol{w}_1 + \mathbb{Z}\boldsymbol{w}_2$ for a pair of linearly independent vectors $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^2$.

Proof:

Let Λ be a lattice in $\mathbb{R}^2$ with $\Lambda \neq \{\mathbf{0}\}$. As in the previous proposition, we can find a $\delta > 0$ with $d(\lambda_1, \lambda_2) > \delta$ for each pair of distinct points $\lambda_1, \lambda_2 \in \Lambda$. Hence, only a finite number of points of Λ can lie within a ball $B(\mathbf{0}, r)$. Choose a vector $\boldsymbol{w}_1 \in \Lambda \setminus \{\mathbf{0}\}$ with $d(\boldsymbol{w}_1, 0)$ minimal. The argument used in the proof of the previous proposition shows that a scalar multiple $t\boldsymbol{w}_1$ is in Λ if and only if $t \in \mathbb{Z}$.

If there are no elements of $\Lambda \setminus \mathbb{Z}\boldsymbol{w}_1$, then we are finished. Otherwise, there are vectors $\boldsymbol{v} \in \Lambda \setminus \mathbb{Z}\boldsymbol{w}_1$. Each such vector can be written as

$$\boldsymbol{v} = \boldsymbol{v}^\perp + t\boldsymbol{w}_1 \qquad \text{with} \qquad \boldsymbol{v}^\perp \ \text{ orthogonal to } \ \boldsymbol{w}_1 \ \text{ and } \ t \in \mathbb{R} \ .$$

Observe that

$$d(\boldsymbol{v}, \mathbb{R}\boldsymbol{w}_1) = ||\boldsymbol{v}^\perp|| \qquad \text{and} \qquad d(\boldsymbol{v}, \mathbf{0}) = \sqrt{||\boldsymbol{v}^\perp||^2 + t^2||\boldsymbol{w}_1||^2} \ .$$

Since only a finite number of points of Λ lie within any ball about the origin, we can choose a vector $\boldsymbol{v} \in \Lambda \setminus \mathbb{Z}\boldsymbol{w}_1$ with $||\boldsymbol{v}^\perp|| = d(\boldsymbol{v}, \mathbb{R}\boldsymbol{w}_1)$ minimal. Call this vector $\boldsymbol{w}_2$.

The vectors $\boldsymbol{w}_1, \boldsymbol{w}_2$ are certainly linearly independent, so any vector $\boldsymbol{v} \in \Lambda$ can be written as a linear combination $\boldsymbol{v} = t_1\boldsymbol{w}_1 + t_2\boldsymbol{w}_2$. The real numbers t_j can be written as $t_j = k_j + t'_j$ with $k_j \in \mathbb{Z}$ and $0 \leqslant t'_j < 1$. Then

$$\boldsymbol{v}' = \boldsymbol{v} - (k_1\boldsymbol{w}_1 + k_2\boldsymbol{w}_2) = t'_1\boldsymbol{w}_1 + t'_2\boldsymbol{w}_2 \in \Lambda$$

and has

$$d(\boldsymbol{v}', \mathbb{R}\boldsymbol{w}_1) = d(t'_2\boldsymbol{w}_2, \mathbb{R}\boldsymbol{w}_1) = t'_2 d(\boldsymbol{w}_2, \mathbb{R}\boldsymbol{w}_1) < d(\boldsymbol{w}_2, \mathbb{R}\boldsymbol{w}_1) \ .$$

The choice of $\boldsymbol{w}_2$ ensures that $\boldsymbol{v}'$ must be in $\mathbb{R}\boldsymbol{w}_1$ and in Λ. We showed above that such a vector must be an integer multiple of $\boldsymbol{w}_1$, so $\boldsymbol{v} \in \mathbb{Z}\boldsymbol{w}_1 + \mathbb{Z}\boldsymbol{w}_2$. □

We say that the lattice Λ has *rank* $0, 1$ *or* 2 according as $\Lambda = \{\mathbf{0}\}, \mathbb{Z}\boldsymbol{w}_1$ or $\mathbb{Z}\boldsymbol{w}_1 + \mathbb{Z}\boldsymbol{w}_2$.

The vectors $\boldsymbol{w}_1, \boldsymbol{w}_2$ in this proposition are not unique. For an example, consider the hexagonal lattice $\mathbb{Z}\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \mathbb{Z}\begin{pmatrix} \frac{1}{2} \\ \frac{1}{2}\sqrt{3} \end{pmatrix}$.

The proof also shows that the parallelogram

$$\{t_1\boldsymbol{w}_1 + t_2\boldsymbol{w}_2 : 0 \leqslant t_1, t_2 < 1\}$$

is a fundamental set for the lattice Λ acting on $\mathbb{E}^2$. The quotient $\mathbb{E}^2/\Lambda$ is obtained by identifying the parallel sides of this parallelogram to give a torus.

7 EUCLIDEAN CRYSTALLOGRAPHIC GROUPS

A *(2-dimensional Euclidean) crystallographic group* is a discrete subgroup of $\mathrm{Isom}(\mathbb{E}^2)$. Let G be such a group. The homomorphism $\phi : \mathrm{Isom}(\mathbb{E}^2) \to \mathrm{O}(2)$ defined in §3 has all the translations $\mathrm{Trans}(\mathbb{E}^2)$ as its kernel. So, when we restrict it to G we get a group homomorphism

$$\phi : G \to \mathrm{O}(2)$$

with kernel equal to $G \cap \mathrm{Trans}(\mathbb{E}^2)$. This is a discrete group of translations, so gives a lattice Λ. We call the rank of Λ the rank of the original crystallographic group G. The image $\phi(G)$ is called the *point group* $\overline{G}$ of G. It is a subgroup of the orthogonal group $\mathrm{O}(2)$. We think of the group G as being made by combining the lattice of translations with the point group.

Lemma 7.1 The point group acts on the lattice
Each isometry in the point group $\overline{G}$ of a discrete subgroup G of $\mathrm{Isom}(\mathbb{E}^2)$ maps the lattice of G onto itself.

Proof:
Let $A : \boldsymbol{x} \mapsto B\boldsymbol{x} + \boldsymbol{v}$ be an isometry in G. Then $\phi(A) = B$ is in the point group $\overline{G}$. If $\boldsymbol{w}$ is in the lattice Λ for G, then the translation $T : \boldsymbol{x} \mapsto \boldsymbol{x} + \boldsymbol{w}$ is in G. Hence, the composite:

$$A \circ T \circ A^{-1} : \boldsymbol{x} \mapsto B(B^{-1}(\boldsymbol{x} - \boldsymbol{v}) + \boldsymbol{w}) + \boldsymbol{v} = \boldsymbol{x} + B\boldsymbol{w}$$

is in G. This shows that $B\boldsymbol{w} \in \Lambda$. □

We are now in a position to describe all of the 2-dimensional Euclidean crystallographic groups. However, the entire program is rather tedious so we will only explain the main themes and illustrate the results.

Let G be the crystallographic group, Λ its lattice, and $\overline{G}$ its point group.

7.1 Rank 0 : Finite Groups

If G is of rank 0, then $\Lambda = \{\mathbf{0}\}$. This means that G can contain no translations and no glide reflections (since the square of a glide reflection is a translation). Hence G contains only the identity, rotations and reflections. Let us first consider the orientation preserving subgroup G^+ of G. Suppose that G contains two rotations R_1, R_2. Then their commutator $R_1R_2R_1^{-1}R_2^{-1}$ is a translation since $\phi(R_1R_2R_1^{-1}R_2^{-1}) = \phi(R_1)\phi(R_2)\phi(R_1)^{-1}\phi(R_2)^{-1}$ and $\mathrm{SO}(2)$ is commutative. Since this translation is by a vector in the lattice $\Lambda = \{\mathbf{0}\}$, it must be the identity and so R_1 and R_2 commute. This implies that they are rotations about the same centre. Since G^+ is a discrete group of rotations, it must be a finite cyclic group. Since G^+ is a normal subgroup of G with index 1 or 2, G must also be finite. Proposition 3.2 shows that G must then be a cyclic or dihedral group.

7.2 Rank 1: Frieze Patterns

If G is of rank 1, then $\Lambda = \mathbb{Z}\boldsymbol{w}$ for some non-zero vector $\boldsymbol{w} \in \mathbb{R}^2$. We will denote the translation $\boldsymbol{x} \to \boldsymbol{x} + \boldsymbol{w}$ by T. Each orthogonal map B in the point group $\overline{G}$ maps Λ onto itself. So B must be the identity, a rotation R through angle π about the origin, a reflection M in the line $\mathbb{R}\boldsymbol{w}$, or a reflection N in the line through the origin orthogonal to $\boldsymbol{w}$. These four maps form a group D_4. Hence $\overline{G}$ must be a subgroup of $\{I, R, M, N\}$. The possible subgroups are:

$$\{I\} \qquad \{I, R\} \qquad \{I, M\} \qquad \{I, N\} \qquad \{I, R, M, N\}$$

For each of these we can work out what the possibilities are for G.

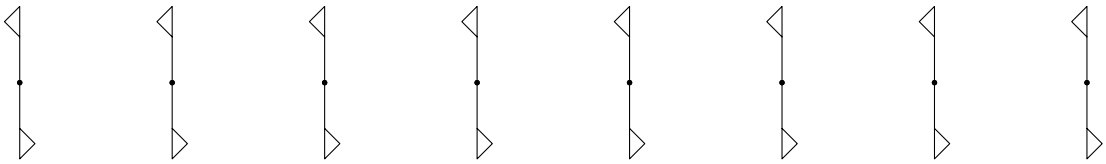
(a) $\overline{G} = \{I\}$.

Then $G = \Lambda$ and consists entirely of translations and is cyclic of infinite order. This G is the symmetry group of a pattern such as:

We call such patterns whose symmetry group is a rank 1 crystallographic group *frieze patterns*.

(b) $\overline{G} = \{I, R\}$.

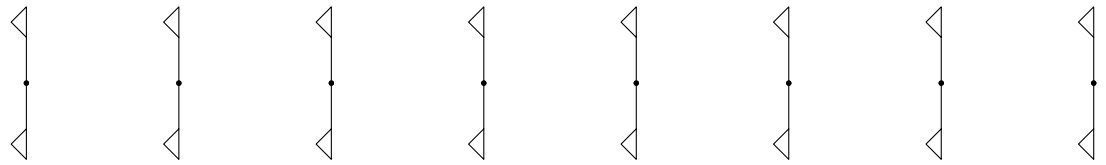
Then G must contain a rotation A with $\phi(A) = R$. So A is a rotation about some centre $\boldsymbol{c}$ through an angle π. Choose co-ordinates in the plane so that this point $\boldsymbol{c}$ is the origin. Then $A = R$. Note that $RTR^{-1} = T^{-1}$. Hence the group G consists of the translations T^k and the rotations T^kR for $k \in \mathbb{Z}$. The rotation T^kR is through angle π with centre $\frac{1}{2}k\boldsymbol{w}$. The group G is an infinite dihedral group D_∞. Such a G is the symmetry group of a frieze pattern such as:



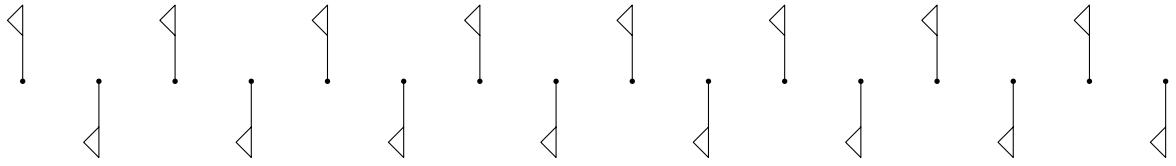
(c) $\overline{G} = \{I, M\}$.

G must contain at least one isometry A with $\phi(A) = M$. This means that A is either a reflection in a line parallel to $\boldsymbol{w}$ or a glide reflection parallel to $\boldsymbol{w}$. In either case, the other isometries of G that ϕ maps to M are T^kA. Choose co-ordinates so that the origin is on the mirror.

In the first case G contains the translations T^k, the reflection A, and the glide reflections T^kA. Note that $ATA = T$, so $G \cong C_\infty \times C_2$.

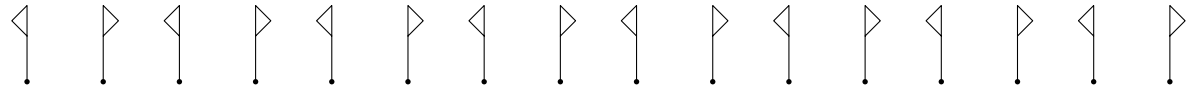


In the second case, A^2 is a translation, so $A^2 = T^r$ for some $r \in \mathbb{Z} \setminus \{0\}$. Then A is reflection followed by translation parallel to the mirror by $\frac{1}{2}r\boldsymbol{w}$. If r is even, say $r = 2k$, then $T^{-k}A$ is a reflection and we are in the previous case. If $r = 2k + 1$, then $C = T^{-k}A$ is reflection followed by translation by $\frac{1}{2}\boldsymbol{w}$. This generates the cyclic group G and gives the pattern:



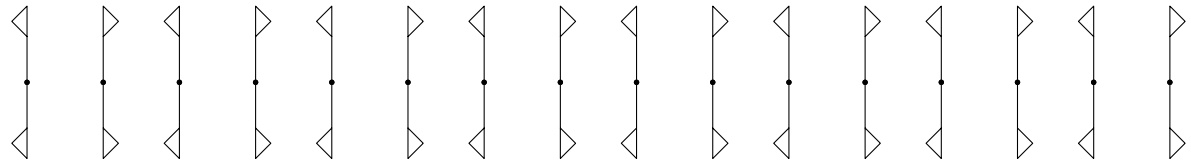
(d) $\overline{G} = \{I, N\}$.

G must contain at least one isometry A with $\phi(A) = N$. This means that A is either a reflection in a line orthogonal to $\boldsymbol{w}$ or a glide reflection orthogonal to $\boldsymbol{w}$. In the second case A^2 would be a translation orthogonal to $\boldsymbol{w}$, which is impossible. So A is a reflection in a line ℓ orthogonal to $\boldsymbol{w}$. Choose co-ordinates so that the origin is on the mirror. Note that $ATA = T^{-1}$ so the group G is an infinite dihedral group.

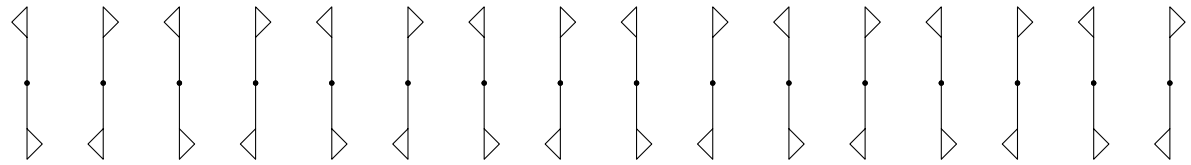


(e) $\overline{G} = \{I, R, M, N\}$.

As in (b) we can choose co-ordinates so that $R \in G$. There must be an isometry $A \in G$ that maps to M under ϕ. There are two cases, as in (b). If we can choose A to be a reflection in a line parallel to $\boldsymbol{w}$, then $ATA^{-1} = T$ and $ARA^{-1} = R$ so $G \cong D_\infty \times C_2$:



If we can choose A to be reflection followed by translation by $\frac{1}{2}\boldsymbol{w}$ then A generates an infinite cyclic group and $RAR^{-1} = A^{-1}$, so G is an infinite dihedral group.



Exercise:

6. Find fundamental sets for each of the frieze groups.

7.3 Rank 2: Wallpaper patterns

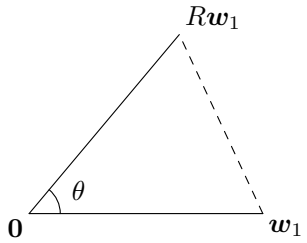
If G is of rank 2, then $\Lambda = \mathbb{Z}\boldsymbol{w}_1 + \mathbb{Z}\boldsymbol{w}_2$ for two linearly independent vectors $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^2$. Each orthogonal map B in the point group $\overline{G}$ maps Λ onto itself. The rotation about the origin through an angle π always maps Λ onto itself. For most lattices there are no other symmetries fixing the origin as the following result shows.

Lemma 7.2 The crystallographic restriction
A rotation in the point group $\overline{G}$ of a crystallographic group G must be of order $2, 3, 4$ *or* 6.

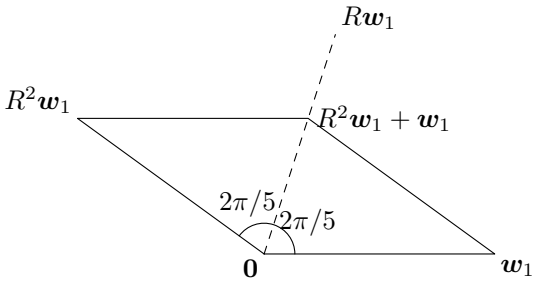
Proof:

Let $\boldsymbol{w}_1$ be an element of $\Lambda \setminus \{\mathbf{0}\}$ with $||\boldsymbol{w}_1|| = r$ minimal. Since Λ is discrete, the set $S = \Lambda \cap \{\boldsymbol{v} : ||\boldsymbol{v}|| = r\}$ is finite and the point group $\overline{G}$ must map this isometrically to itself. This certainly implies that any rotation $R \in \overline{G}$ is of finite order.

Choose R as the rotation in $\overline{G}$ through the smallest angle, say θ. The vector $R\boldsymbol{w}_1 - \boldsymbol{w}_1$ is also in Λ so it must have length at least $||\boldsymbol{w}_1||$ unless it is 0. This means that $\theta \geqslant \pi/3$. Consequently R has order $2, 3, 4, 5$ or 6.



Suppose that R had order 5. Then, $\boldsymbol{w}_1 + R^2\boldsymbol{w}_1 \in \Lambda$ would be at a distance less than r from the origin, which is forbidden.



□

Corollary 7.3 Point groups
The point group of a 2-dimensional Euclidean crystallographic group is either cyclic C_1, C_2, C_3, C_4, C_6 or dihedral $D_2, D_4, D_6, D_8, D_{12}$.

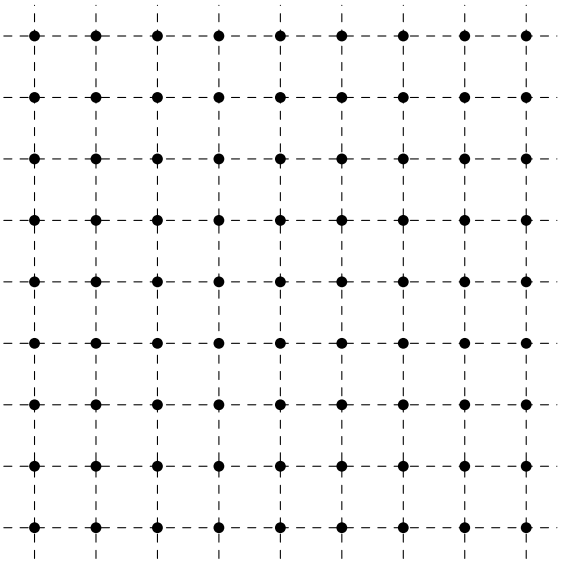
Proof:

Take $\boldsymbol{w}_1$ as in the previous proof and let $R \in \overline{G}$ be the rotation through the smallest angle θ.

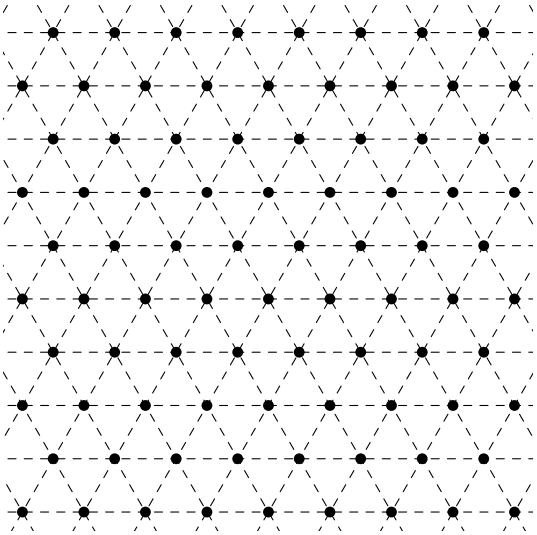
If $\theta = \pi/2$, then $\boldsymbol{w}_1$ and $\boldsymbol{w}_2 = R\boldsymbol{w}_1$ generate the lattice. This is the *square lattice*.

If $\theta = 2\pi/3$ or $\theta = \pi/3$, then $\boldsymbol{w}_1$ and $\boldsymbol{w}_2 = R\boldsymbol{w}_1$ generate the lattice. This is the *hexagonal lattice*.

□

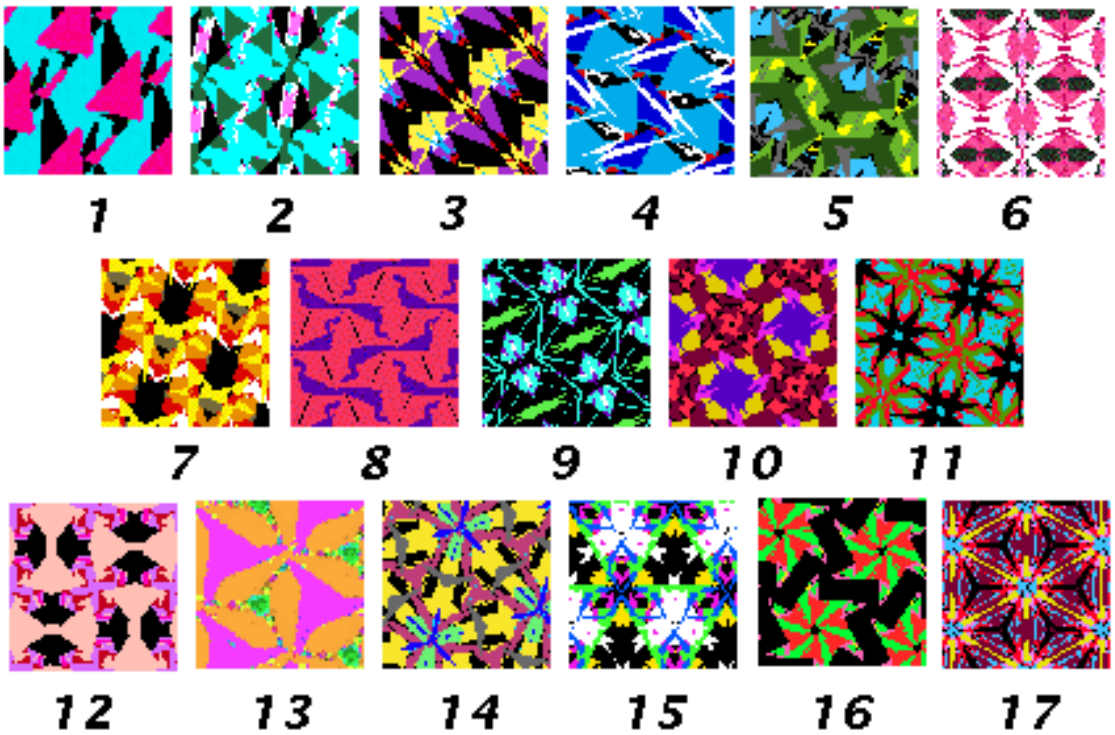


Square lattice



Hexagonal lattice

We can now proceed as in the previous part to find all the possible crystallographic groups with rank 2. These are the *wallpaper groups*. There are 17 of them, as illustrated below.



http://www.clarku.edu/∼djoyce/wallpaper/

8 MÖBIUS TRANSFORMATIONS

8.1 The Riemann Sphere

It is useful to add an extra element ∞ to the complex plane to form the extended complex plane $\mathbb{C} \cup \{\infty\}$. It seems that the point at infinity is very different from the other, finite points but Riemann showed that this is not really the case. He did this by representing all of the points of the extended complex plane by points of the unit sphere S^2 in $\mathbb{R}^3$. This sphere is called the *Riemann sphere*.

Let $\mathbb{P}$ be the unit sphere

$$\mathbb{P} = \{(z,t) \in \mathbb{C} \times \mathbb{R} : |z|^2 + t^2 = 1\}$$

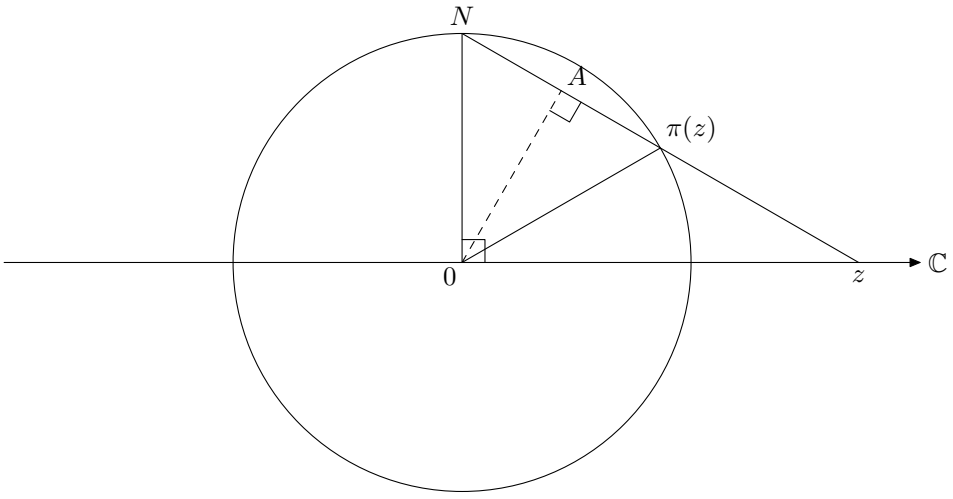
in the three dimensional real vector space $\mathbb{C} \times \mathbb{R}$. The "North pole" of this sphere will be denoted by $N = (0,1)$. Stereographic projection maps points of the complex plane to points of the Riemann sphere $\mathbb{P}$ and *vice versa*. Let $z \in \mathbb{C}$. Then the straight line through N and $(z,0)$ crosses the sphere at N and another point $(w,t) \in \mathbb{P}$. We write $\pi(z) = (w,t)$ and define $\pi(\infty) = N$. Then π gives us a map $\pi : \mathbb{C} \cup \{\infty\} \to \mathbb{P}$. This map is invertible, for if (t,w) is any point of $\mathbb{P}$ except N, then the straight line through N and (w,t) will cross the plane $\{(z,s) : s = 0\}$ at a single point $(z,0)$ with $\pi(z) = (w,t)$.

It is easy to give a formula for stereographic projection. The points on the line from $z \in \mathbb{C}$ to N are $\{s(z,0) + (1-s)(0,1) : s \in \mathbb{R}\}$. This line crosses the sphere $\mathbb{P}$ when $s = 0$, giving the North pole, and when $s = \frac{2}{1-|z|^2}$, giving

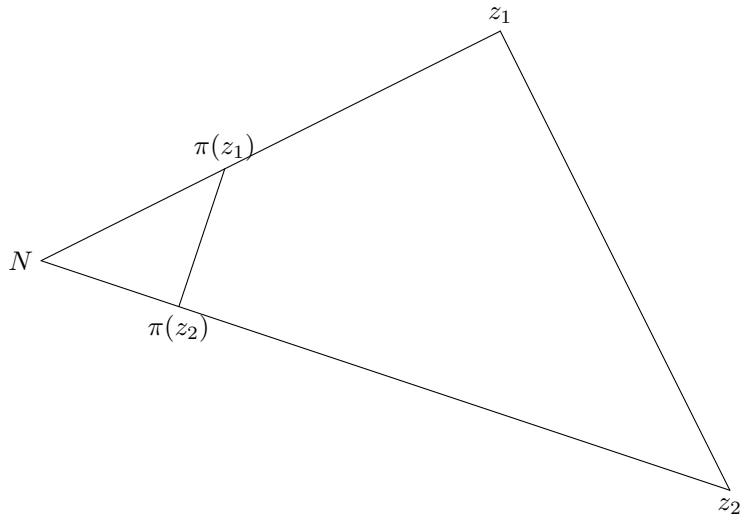
$$\pi(z) = \left(\frac{2z}{1+|z|^2}, \frac{-1+|z|^2}{1+|z|^2} \right) .$$

Note from the diagram below that the triangles $\triangle z0N$, $\triangle 0AN$ and $\triangle 0A\pi(z)$ are all similar. Pythagoras' theorem shows that $d(N,z) = \sqrt{1+|z|^2}$. Hence,

$$d(N,A) = d(A,\pi(z)) = \frac{1}{\sqrt{1+|z|^2}} \qquad \text{and} \qquad d(N,\pi(z)) = \frac{2}{\sqrt{1+|z|^2}} .$$



Now let us consider two points $z_1, z_2 \in \mathbb{C}$. The *chordal distance* $\kappa(z_1, z_2)$ is the Euclidean distance between the stereographic projections $\pi(z_1)$ and $\pi(z_2)$. In the diagram below, we show the triangle with vertices N, z_1 and z_2.



We know that

$$d(N, z_j) = \sqrt{1 + |z_j|^2} \qquad \text{and} \qquad d(N, \pi(z_j)) = \frac{2}{\sqrt{1 + |z_j|^2}} \ .$$

So the triangles $\triangle N z_1 z_2$ and $\triangle N \pi(z_2)\pi(z_1)$ are similar with scale factor $\frac{2}{\sqrt{1+|z_1|^2}\sqrt{1+|z_2|^2}}$. Consequently,

$$\kappa(z_1, z_2) = d(\pi(z_1), \pi(z_2)) = \frac{2|z_1 - z_2|}{\sqrt{1 + |z_1|^2}\sqrt{1 + |z_2|^2}} \ .$$

When one of the points, say z_2, is ∞ then we interpret this as

$$\kappa(z_1, \infty) = \frac{2}{\sqrt{1 + |z_1|^2}} \ .$$

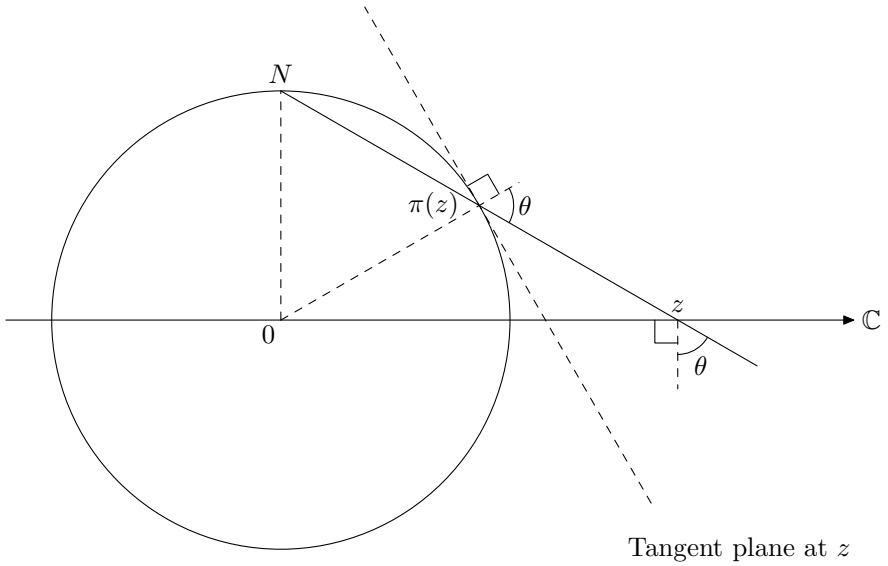
Exercise:

7. Prove the formula for the chordal distance between two points $z_1, z_2 \in \mathbb{C} \cup \{\infty\}$ algebraically by using the formula for stereographic projection.

-8. Prove that the chordal metric is a metric on the Riemann sphere.

9. Two points $z, z' \in \mathbb{C} \cup \{\infty\}$ are *antipodal* if their stereographic projections satisfy $\pi(z') = -\pi(z)$, so they are at the opposite ends of a diameter of the Riemann sphere. Show that z, z' are antipodal if and only if
$$z' = -1/\overline{z} \ .$$

A similar argument to that used above to find the chordal metric shows that stereographic projection is conformal: it preserves the angle between curves. For, in the diagram below, the straight line from N to z crosses the tangent plane at $\pi(z)$ and the complex plane $\mathbb{C}$ at the same angle θ. Hence projection with centre N from the tangent plane to $\mathbb{C}$ preserves angles. Consequently, π also preserves the angle between two curves that meet at z.



It is convenient to define *circles* in $\mathbb{C} \cup \{\infty\}$ to mean both straight lines and circles. The following result explains why this is so.

Proposition 8.1 Stereographic projection preserves circles.
A curve Γ in $\mathbb{C} \cup \{\infty\}$ is a circle or a straight line if, and only if, the stereographic projection $\pi(\Gamma)$ is a circle on the Riemann sphere.

Proof:

We can write any circle or straight line in $\mathbb{C} \cup \{\infty\}$ as

$$a_o|z|^2 + \overline{a}z + a\overline{z} + a_\infty = 0 \tag{1}$$

where $a_0, a_\infty \in \mathbb{R}$ and $a \in \mathbb{C}$. The stereographic projection $\pi(z)$ is

$$(w, t) = \left(\frac{2z}{1+|z|^2}, \frac{-1+|z|^2}{1+|z|^2} \right) .$$

So (1) is equivalent to

$$(a_0 + a_\infty) + \overline{a}w + a\overline{w} + (a_0 - a_\infty)t = 0 . \tag{2}$$

This is the intersection with the Riemann sphere of a plane, so it is a circle on the sphere.

Note that the plane intersects the sphere if, and only if, $|a|^2 - a_0 a_\infty \geqslant 0$. This same condition ensures that (1) does describe a circle or straight line rather than the empty set. □

8.2 Möbius Transformations

Let a, b, c, d be complex numbers with $ad - bc \neq 0$. Then we can define a map $T : \mathbb{C} \cup \{\infty\} \to \mathbb{C} \cup \{\infty\}$ by

$$T : z \mapsto \frac{az+b}{cz+d} .$$

Note that $T(\infty) = a/c$ and $T(-d/c) = \infty$. These maps are called *Möbius transformations* and form a group Möb under composition of maps. The map

$$\phi : \mathrm{GL}(2, \mathbb{C}) \to \text{Möb} ; \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto T$$

is a group homomorphism.

A matrix $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is in the kernel of this homomorphism when

$$\frac{az+b}{cz+d} = z \qquad \text{for all } z \in \mathbb{C} \cup \{\infty\} \ .$$

This occurs if and only if $a = d$ and $b = c = 0$, so $\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \lambda I$ for some scalar $\lambda \in \mathbb{C} \setminus \{0\}$. This shows that a Möbius transformation is unaltered when we multiply each of the coefficients a, b, c, d by a non-zero scalar λ. Usually we choose the scalar λ so that the determinant $ad - bc$ is 1. Then the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is in $\mathrm{SL}(2, \mathbb{C})$. Now

$$\phi : \mathrm{SL}(2, \mathbb{C}) \to \text{Möb} \ ; \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto T$$

is a group homomorphism whose kernel consists of the two matrices I and $-I$. Consequently, the Möbius group is the quotient $\mathrm{SL}(2, \mathbb{C})/\{I, -I\}$. We denote this quotient by $\mathrm{PSL}(2, \mathbb{C})$ and call it the *projective special linear group*.

Recall that Möbius transformations map circles to circles.

Proposition 8.2 Möbius transformations map circles to circles
A Möbius transformation maps any circle on the Riemann sphere to a circle on the Riemann sphere.

Proof:

Let Γ be the circle $\{z : p_0|z|^2 + p\overline{z} + \overline{p}z + p_\infty = 0\}$. Let T be a Möbius transformation with inverse

$$S : z \mapsto \frac{az+b}{cz+d} \ .$$

Then

$$T(\Gamma) = \{z : S(z) \in \Gamma\} = \left\{ z : p_0|az+b|^2 + p\overline{(az+b)(cz+d)} + \overline{p}(az+b)(cz+d) + p_\infty|cz+d|^2 = 0 \right\} \ .$$

Expanding this gives an expression of the form:

$$\left\{ z : q_0|z|^2 + q\overline{z} + \overline{q}z + q_\infty = 0 \right\}$$

which is clearly another circle. □

Proposition 8.3
For any triples of distinct points in the Riemann sphere, (a_0, a_1, a_∞) and (b_0, b_1, b_∞) , there is a unique Möbius transformation T with $T(a_0) = b_0$, $T(a_1) = b_1$ and $T(a_\infty) = b_\infty$.

Proof:

The Möbius transformation

$$S_a : z \mapsto \left(\frac{a_1 - a_\infty}{a_1 - a_0} \right) \left(\frac{z - a_0}{z - a_\infty} \right)$$

has $S_a(a_0) = 0$, $S_a(a_1) = 1$ and $S_a(a_\infty) = \infty$. Hence, $T = S_b^{-1} \circ S_a$ has the required properties. If T' is another Möbius transformation with the same properties, then $S_b \circ T' \circ S_a^{-1}$ fixes 0, 1 and ∞ so it must be the identity. □

It follows from this that we can find a Möbius transformation that maps any circle in $\mathbb{P}$ onto any other circle. For we choose three points on the first and find a Möbius transformation that maps them to three points on the second.

Exercise:

10. Let Γ_1, Γ_2 be two disjoint circles on the Riemann sphere. Show that there is a Möbius transformation that maps them to two circles in $\mathbb{C}$ centred on 0.

Proposition 8.4 Isometries of the Riemann sphere.
A Möbius transformation is an isometry of the Riemann sphere (for the chordal metric) if, and only if, it is represented by a matrix

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SU}(2) \ .$$

$M \in \mathrm{SU}(2)$ means that the matrix M preserves the complex inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{C}^2$ and has determinant 1. This means that $M^* M = I$ or, equivalently, that $d = \overline{a}$, $c = -\overline{b}$ and $|a|^2 + |b|^2 = 1$.

Proof:

Let $J : \mathbb{P} \to \mathbb{P}$ be the map $z \mapsto -1/\overline{z}$ that maps a point to the antipodal point. Then $J(z)$ is the unique point of $\mathbb{P}$ that is at a distance 2 from z. Hence any Möbius isometry T must satisfy $J \circ T \circ J = T$. Let T be represented by the matrix $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with $\det M = 1$. We then have

$$J \circ T \circ J : z \mapsto \frac{-\overline{d} z + \overline{c}}{\overline{b} z - \overline{a}} \ .$$

This shows that $J \circ T \circ J$ is itself a Möbius transformation represented by the matrix $\begin{pmatrix} -\overline{d} & \overline{c} \\ \overline{b} & -\overline{a} \end{pmatrix}$.

Therefore, if T is an isometry, then the two matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $\begin{pmatrix} -\overline{d} & \overline{c} \\ \overline{b} & -\overline{a} \end{pmatrix}$ both represent T. This implies that

$$\begin{pmatrix} -\overline{d} & \overline{c} \\ \overline{b} & -\overline{a} \end{pmatrix} = \pm \begin{pmatrix} a & b \\ c & d \end{pmatrix} \ .$$

The + sign requires that $d = -\overline{a}$ and $c = \overline{b}$ so $1 = ad - bc = -|a|^2 - |b|^2$, which is impossible. Hence we must have the $-$ sign and this shows that M is unitary.

Conversely, suppose that the Möbius transformation T is represented by the matrix $M \in \mathrm{SU}(2)$. Each $z \in \mathbb{P}$ can be written as $z = z_1/z_2$ with $\boldsymbol{z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \in \mathbb{C}^2$. Similarly, write $w = w_1/w_2$ with $\boldsymbol{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$. Write $\boldsymbol{w}^* = \begin{pmatrix} w_1^* \\ w_2^* \end{pmatrix} = \begin{pmatrix} -\overline{w_2} \\ \overline{w_1} \end{pmatrix}$, so that $Jw = w_1^*/w_2^* = -\overline{w_2}/\overline{w_1}$. Then we have

$$\kappa(Jw, z) = \frac{2|Jw - z|}{\sqrt{1 + |Jw|^2}\sqrt{1 + |z|^2}} = \frac{2|-\overline{w_2} z_2 - \overline{w_1} z_1|}{\sqrt{|w_1|^2 + |w_2|^2}\sqrt{|z_1|^2 + |z_2|^2}} = \frac{2|\langle \boldsymbol{w}, \boldsymbol{z} \rangle|}{||\boldsymbol{w}||\,||\boldsymbol{z}||} \ .$$

Since $J \circ T = T \circ J$ and $\langle M\boldsymbol{a}, M\boldsymbol{b} \rangle = \langle \boldsymbol{a}, \boldsymbol{b} \rangle$ we see that

$$\kappa(T(Jw), T(z)) = \kappa(J(Tw), T(z)) = \frac{2|\langle M(\boldsymbol{w}), M(\boldsymbol{z}) \rangle|}{||M(\boldsymbol{w})||\,||M(\boldsymbol{z})||} = \frac{2|\langle \boldsymbol{w}, \boldsymbol{z} \rangle|}{||\boldsymbol{w}||\,||\boldsymbol{z}||} = \kappa(Jw, z)$$

and so T is an isometry. □

9 VISUALISING MÖBIUS TRANSFORMATIONS

9.1 Fixed Points

Let T be the Möbius transformation represented by the matrix $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with $ad - bc = 1$. The vector $\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$ is an eigenvector of M precisely when $T(z_1/z_2) = z_1/z_2$. This means that z_1/z_2 is a fixed point of T. We know, from Linear Algebra, that the matrix M is conjugate to $\begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix}$ for some $\lambda \neq 0$, or to $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. This gives corresponding results for the Möbius transformations. Let us prove this directly using the transformations.

Theorem 9.1 Fixed points of Möbius transformations
A non-identity Möbius transformation has either 1 *or* 2 *fixed points in* $\mathbb{P}$. *If it has* 1, *then it is conjugate in Möb to* $P : z \mapsto z + 1$. *If it has* 2, *then it is conjugate in Möb to* $M_k : z \mapsto kz$ *for some* $k \neq 0, 1$.

Proof:

Suppose that the Möbius transformation T has two fixed points z_0 and z_∞. Choose a Möbius transformation A with $A(z_0) = 0$ and $A(z_\infty) = \infty$. Then the conjugate $A \circ T \circ A^{-1}$ fixes 0 and ∞. This implies that $A \circ T \circ A^{-1}$ is represented by a matrix $\begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix}$ for some $\lambda \neq -1, 0, +1$. Consequently, $A \circ T \circ A^{-1} = M_{\lambda^2}$.

Suppose that T has 1 fixed point z_0 only. Choose A with $A(z_0) = \infty$. Then $A \circ T \circ A^{-1}$ fixes ∞ alone. This means that it is $z \mapsto z + b$ for some $b \neq 0$. By replacing A by $b^{-1}A$ we get $A \circ T \circ A^{-1} = P$. □

Exercise:

-11. When are two of the Möbius transformations M_k and P conjugate?

12. Find all of the Möbius transformations that commute with M_k for a fixed k. Hence describe the group
$$Z(T) = \{A \in \text{Möb} : A \circ T = T \circ A\}$$
for an arbitrary Möbius transformation T. Describe the set $\{A(z_o) : A \in Z(T)\}$ for z_o a point in $\mathbb{P}$.

A non-identity Möbius transformation is said to be:

parabolic if it is conjugate to P;

elliptic if it is conjugate to M_k for $|k| = 1$ $(k \neq 1)$;

hyperbolic if it is conjugate to M_k for $k \in \mathbb{R}^+$ $(k \neq 0, +1)$;

loxodromic if it is conjugate to M_k for $k \in \mathbb{C}$ with $|k| \neq 1$ and $k \notin \mathbb{R}^+$.

So a Möbius transformation $T : z \mapsto \frac{az+b}{cz+d}$, with $ad - bc = 1$, is

the identity if $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is conjugate to $\pm \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

parabolic if $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is conjugate to $\pm \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$.

elliptic if $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is conjugate to $\begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix}$ for some λ with $|\lambda| = 1$.

hyperbolic if $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is conjugate to $\begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix}$ for some λ with $\lambda \in \mathbb{R}$ and $\lambda \neq -1, 0, +1$.

loxodromic if $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is conjugate to $\begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix}$ for some λ with $\lambda \notin \mathbb{R}$ and $|\lambda| \neq 1$.

Theorem 9.1 shows that every non-identity transformation falls into one of these classes. It is simple to tell which by considering the trace.

Corollary 9.2 Trace determines conjugacy class of a Möbius transformation
Let T be a non-identity Möbius transformation represented by a matrix $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with determinant 1. Then T is

parabolic	$\Leftrightarrow$	$\operatorname{tr} M = \pm 2$;
elliptic	$\Leftrightarrow$	$-2 < \operatorname{tr} M < 2$;
hyperbolic	$\Leftrightarrow$	$\operatorname{tr} M < -2$ *or* $\operatorname{tr} M > 2$;
loxodromic	$\Leftrightarrow$	$\operatorname{tr} M \notin \mathbb{R}$.

Proof:

We know that T is conjugate to M_k or to P. This means that the matrix M is conjugate to $\begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix}$ for λ a square root of k, or to $\pm \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. Now we simply note that

$$\operatorname{tr} \begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix} = \lambda + \lambda^{-1} \qquad \text{and} \qquad \operatorname{tr} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = 2 \ .$$

□

Exercise:

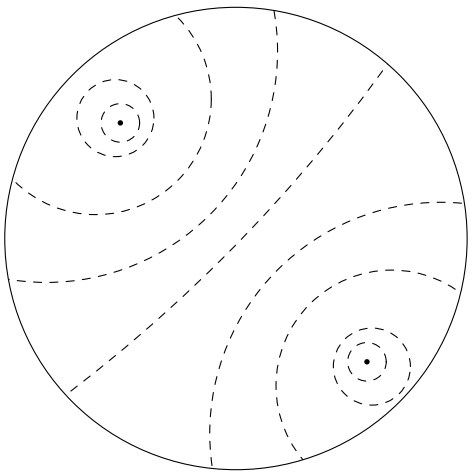
-13. If M is a 2×2 matrix with determinant 1, show that the characteristic equation for M is

$$t^2 - (\operatorname{tr} M)t + 1 = 0 \ .$$

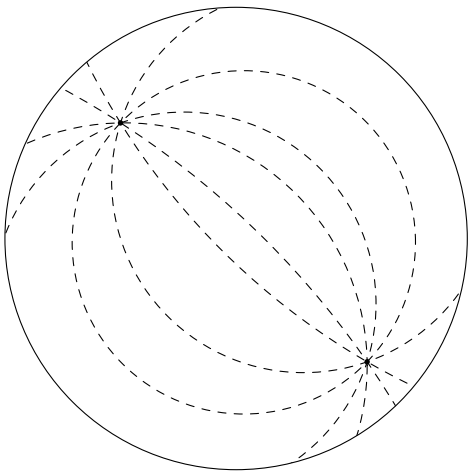
Deduce that the trace determines the eigenvalues of M.

14. Suppose that the Möbius transformation T is represented by the matrix M but that $\det M \neq 1$. Show that T is parabolic if and only if $(\operatorname{tr} M)^2 = 4 \det M$. Establish similar conditions for T to be elliptic, hyperbolic or loxodromic.

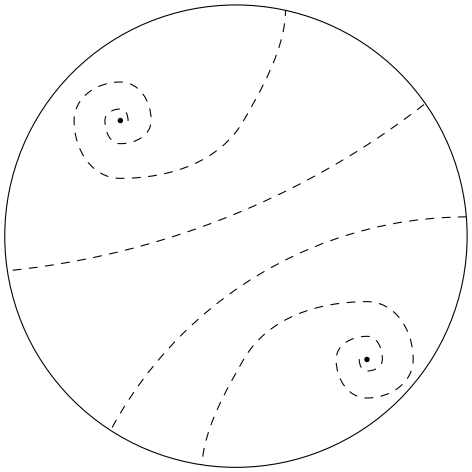
It is now fairly simple to visualise how Möbius transformations act on the Riemann sphere. An elliptic transformation is conjugate to $z \mapsto e^{i\theta} z$. This rotates the sphere fixing 0 and ∞. Each point is moved along a circle. A hyperbolic transformation is conjugate to $z \mapsto kz$ for $k > 1$. This moves points along arcs of circles from one fixed point towards the other. A loxodromic transformation is conjugate to $z \mapsto kz$ for $k \notin \mathbb{R}$. This moves points along logarithmic spirals away from one fixed point and towards the other. Finally, a parabolic transformation is conjugate to $z \mapsto z + 1$. This maps points along a circle through the single fixed point. The points are mapped away on one side and towards on the other. The pictures below illustrate this.



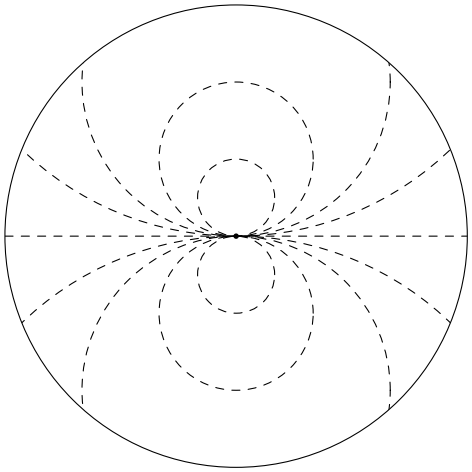
Elliptic



Hyperbolic



Loxodromic



Parabolic

9.2 Inversion

Let Γ be a circle on the Riemann sphere. Two distinct points z, z' are *inverse points for* Γ if every circle orthogonal to Γ through z also passes through z'. Also, when $z \in \Gamma$, we say that z and z itself are inverse points. We will prove later that, for every point z there is a unique point z' so that z, z' are inverse points for Γ.

For example, two points are inverse for the real axis $\mathbb{R} \cup \{\infty\}$ when they are complex conjugates of one another.

Lemma 9.3 Möbius transformations preserve inverse points
Let T be a Möbius transformation and Γ a circle in $\mathbb{P}$. If z and z' are inverse points for Γ then $T(z)$, $T(z')$ are inverse points for $T(\Gamma)$.

Proof:

We know that the Möbius transformation T preserves angles and maps circles to circles. Any circle through z orthogonal to Γ is therefore mapped to a circle through $T(z)$ orthogonal to $T(\Gamma)$. The original circle passes through z' so the the image passes through $T(z')$ as required. □

Proposition 9.4 Inversion
For each circle Γ in the Riemann sphere and each point $z \in \mathbb{P}$ there is a unique point $J(z)$ with $z, J(z)$ inverse points for Γ.

The map J is called *inversion in* Γ. It is an involution: $J^2 = I$ and reverses orientation, so it is certainly not a Möbius transformation. J fixes every point of the circle Γ.

Proof:

It is clear that $C(z) = \overline{z}$ is the unique point with $z, C(z)$ inverse points for $\mathbb{R} \cup \{\infty\}$.

For any circle Γ we can find a Möbius transformation T that maps $\mathbb{R} \cup \{\infty\}$ onto Γ. Now the lemma shows that $J(z) = T(C(T^{-1}(z)))$ is the unique point with $z, J(z)$ inverse points for Γ. □

.

Example: Inversion in the unit circle.

The Möbius transformation

$$T : z \mapsto -i\left(\frac{z+i}{z-i}\right)$$

maps the unit circle $\mathbb{T}$ onto the real axis. (It is a rotation of the Riemann sphere about an axis through 1 and -1.) Hence, inversion in $\mathbb{T}$ is given by

$$J(z) = T^{-1}(\overline{T(z)}) = \frac{1}{\overline{z}} \ .$$

More generally, when $\Gamma = \{z \in \mathbb{C} : |z - c| = r\}$, then two points z, z' are inverse points for Γ when they lie on the same half-line from c to ∞ and $|z - c|\,|z' - c| = r^2$. Hence

$$J(z) = c + \frac{r^2}{\overline{z} - \overline{c}} \ .$$

Exercise:

15. Show that inversion maps any circle to another circle. Show that inversion preserves the magnitude of angles but reverses their orientation.

Inversions in circles are analogous to reflections. Our ultimate aim is to find a metric on the ball $B^3 = \{\boldsymbol{x} \in \mathbb{R}^3 : ||\boldsymbol{x}|| < 1\}$, so that the Möbius transformations form the orientation preserving isometry group of B^3 for this metric. The inversions will then be reflections in hyperbolic planes. Before achieving this goal we will study the simpler case of those Möbius transformations that map a disc onto itself.

Proposition 9.5 The composition of two inversions is Möbius.
The composition of an even number of inversions is a Möbius transformation.

Proof:

Let C be the particular inversion $z \mapsto \overline{z}$. If Γ is a circle, we can find a Möbius transformation T that maps $\mathbb{R} \cup \{\infty\}$ onto Γ. Then inversion in Γ is given by $J = T \circ C \circ T^{-1}$.

If $T(z) = \dfrac{az+b}{cz+d}$, then $C \circ T \circ C(z) = \dfrac{\overline{a}z + \overline{b}}{\overline{c}z + \overline{d}}$. This shows that $C \circ T \circ C$ is a Möbius transformation. Hence,

$$C \circ J = C \circ T \circ C \circ T^{-1} = (C \circ T \circ C) \circ T^{-1}$$

is a Möbius transformation, as is its inverse $J \circ C$.

Now, if J_1, J_2 are two inversions we can write

$$J_2 \circ J_1 = (J_2 \circ C) \circ (C \circ J_1)$$

to show that $J_2 \circ J_1$ is a Möbius transformation. □

It is quite easy to identify the Möbius transformations that we obtain by composing two inversions. Suppose that J_1, J_2 are inversions in the circles Γ_1, Γ_2 respectively. If Γ_1 and Γ_2 cross at two points w_1, w_2, then we can conjugate by a Möbius transformation that sends these points to 0 and ∞ respectively. This transforms Γ_1 and Γ_2 to two straight lines through 0 (and ∞). The inversions become reflections in these lines. So $J_2 \circ J_1$ is the elliptic transformation rotating about an axis through 0 and ∞ through twice the angle between the lines. Clearly we can obtain any elliptic transformation in this way.

Similarly, if Γ_1 and Γ_2 touch at a single point w, then we can transform w to ∞ and see that $J_2 \circ J_1$ is a parabolic transformation fixing ∞.

Finally, if Γ_1 and Γ_2 are disjoint, then we can conjugate them to get two concentric circles $\{z : |z| = 1\}$ and $\{z : |z| = R\}$. Then

$$J_1(z) = \frac{1}{\overline{z}} \; ; \quad J_2(z) = \frac{R^2}{\overline{z}} \; ; \quad \text{so} \quad J_2 \circ J_1(z) = R^2 z \; .$$

Hence $J_2 \circ J_1$ is a hyperbolic transformation.

It is clear that we can obtain any elliptic, parabolic or hyperbolic transformation in this way. There are no other ways that two circles can intersect, so loxodromic transformations can not be the composition of two inversions.

Exercise:

-16. Show that any loxodromic transformation is the composite of 4 inversions.

---

10 THE HYPERBOLIC PLANE, I

10.1 Möbius Transformations of the unit disc

Inversion in the unit circle is $J : z \mapsto 1/\overline{z}$. This fixes points of the unit circle $\mathbb{T}$ and interchanges the unit disc $\mathbb{D}$ and the complementary disc $\mathbb{D}' = \{z \in \mathbb{P} : |z| > 1\}$.

Proposition 10.1 Möbius transformations of the unit disc
A Möbius transformation T maps the unit disc onto itself if, and only if, it is of the form:

$$z \mapsto \frac{az+b}{\overline{b}z+\overline{a}} \qquad \textit{with} \;\; |a|^2 - |b|^2 = 1 \; .$$

Proof:

Let T be the Möbius transformation $T : z \mapsto \frac{az+b}{cz+d}$ with $ad - bc = 1$. If T maps the unit disc $\mathbb{D}$ onto itself, then it must also map the unit circle onto itself. Hence, it must map any pair of inverse points for $\mathbb{T}$ to another pair of inverse points. This implies that $J \circ T \circ J = T$. Now

$$J(T(J(z))) = \frac{\overline{d}z + \overline{c}}{\overline{b}z + \overline{a}}$$

so the matrices $\begin{pmatrix} \overline{d} & \overline{c} \\ \overline{b} & \overline{a} \end{pmatrix}$ and $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ both represent the same transformation. This means that

$$\begin{pmatrix} \overline{d} & \overline{c} \\ \overline{b} & \overline{a} \end{pmatrix} = \pm \begin{pmatrix} a & b \\ c & d \end{pmatrix} \; .$$

If we have the + sign, then the matrix is

$$\begin{pmatrix} a & b \\ \overline{b} & \overline{a} \end{pmatrix} \qquad \text{with} \;\; |a|^2 - |b|^2 = 1$$

as required. If we have the − sign, then the matrix is

$$\begin{pmatrix} a & b \\ -\overline{b} & -\overline{a} \end{pmatrix}$$

with $-|a|^2 + |b|^2 = 1$. So $|T(0)| = |-b/\overline{a}| > 1$, which is impossible if T maps $\mathbb{D}$ onto itself.

Conversely, suppose that

$$T : z \mapsto \frac{az+b}{\overline{b}z+\overline{a}} \qquad \text{with} \;\; |a|^2 - |b|^2 = 1 \; .$$

Then $J \circ T \circ J = T$. Any point $z \in \mathbb{T}$ satisfies $J(T(z)) = T(J(z)) = T(z)$ so T maps the unit circle onto itself. Consequently, T must map the unit disc onto either $\mathbb{D}$ or the complementary disc $\mathbb{D}'$. However, $|T(0)| = |b/\overline{a}| < 1$ so it must map onto $\mathbb{D}$. □

The Möbius transformations that map $\mathbb{D}$ onto itself form a subgroup

$$\text{Möb}(\mathbb{D}) = \left\{ z \mapsto \frac{az+b}{\overline{b}z+\overline{a}} : |a|^2 - |b|^2 = 1 \right\}$$

of the Möbius group. This is the *Möbius group of the disc* $\mathbb{D}$.

Example: For each ω with $|\omega| = 1$, the map $z \mapsto \omega z$ is clearly in Möb($\mathbb{D}$). Also, for each $z_o \in \mathbb{D}$ the map

$$z \mapsto \frac{z + z_o}{1 + \overline{z_o} z}$$

is represented by the matrix

$$\begin{pmatrix} \frac{1}{\sqrt{1-|z_o|^2}} & \frac{z_o}{\sqrt{1-|z_o|^2}} \\ \frac{\overline{z_o}}{\sqrt{1-|z_o|^2}} & \frac{1}{\sqrt{1-|z_o|^2}} \end{pmatrix}$$

and so is in Möb($\mathbb{D}$).

Exercise:

17. Show directly that the map

$$z \mapsto \frac{\omega z + z_o}{1 + \overline{z_o} \omega z}$$

maps the unit disc $\mathbb{D}$ onto itself when $|\omega| = 1$ and $z_o \in \mathbb{D}$. Show conversely, that every transformation in Möb($\mathbb{D}$) is of this form.

.

For any other disc Δ in the Riemann sphere, we can do the same argument as above to find the group Möb(Δ) of Möbius transformations that map Δ onto itself. This is a subgroup of the full Möbius group conjugate to Möb($\mathbb{D}$). For, we can find a transformation T with $T(\mathbb{D}) = \Delta$. Then

$$A \in \text{Möb}(\mathbb{D}) \text{ if, and only if, } T \circ A \circ T^{-1} \in \text{Möb}(\Delta) \ .$$

A particularly important example is when Δ is the upper half-plane: $\mathbb{R}^2_+ = \{x + iy \in \mathbb{C} : y > 0\}$. Inversion in the boundary of this is complex conjugation $J : z \mapsto \overline{z}$. So a Möbius transformation $T : z \mapsto \frac{az+b}{cz+d}$ with $ad - bc = 1$ maps $\mathbb{R}^2_+$ onto itself when $J \circ T \circ J = T$ and $T(i) \in \mathbb{R}^2_+$. Now

$$J \circ T \circ J : z \mapsto \frac{\overline{a} z + \overline{b}}{\overline{c} z + \overline{d}}$$

so we need

$$\begin{pmatrix} \overline{a} & \overline{b} \\ \overline{c} & \overline{d} \end{pmatrix} = \pm \begin{pmatrix} a & b \\ c & d \end{pmatrix} \ .$$

The + sign gives transformations $z \mapsto \frac{az+b}{cz+d}$ with $a, b, c, d \in \mathbb{R}$ and $ad - bc = 1$. These do map the upper half-plane onto itself. The $-$ sign gives $z \mapsto \frac{az+b}{cz+d}$ with $a, b, c, d \in i\mathbb{R}$ and $ad - bc = 1$. These map the upper half-plane onto the lower half-plane. This shows that every Möbius transformation mapping the upper half-plane onto itself is represented by a matrix in $\text{SL}(2, \mathbb{R})$. So $\text{Möb}(\mathbb{R}^2_+) \cong \text{SL}(2, \mathbb{R})/\{I, -I\}$.

10.2 The Hyperbolic Metric on $\mathbb{D}$

We wish to define a new metric, the hyperbolic metric, on the unit disc $\mathbb{D}$ for which each of the Möbius transformations in Möb($\mathbb{D}$) will be an isometry.

We begin by defining this metric for an infinitesimal displacement dz. This should have length $ds = \lambda(z)|dz|$ for some density function $\lambda : \mathbb{D} \to (0, \infty)$. More formally, this means that a smooth curve $\gamma : [a, b] \to \mathbb{D}$ should have length

$$L(\gamma) = \int_a^b \lambda(\gamma(t)) \, |\gamma'(t)| \, dt \ .$$

If such a metric is to have each transformation in Möb($\mathbb{D}$) as an isometry, then the function λ is almost completely determined. For

$$T : z \mapsto \frac{z + z_o}{1 + \overline{z_o} z}$$

is a Möbius transformation mapping $\mathbb{D}$ onto $\mathbb{D}$ for each $z_o \in \mathbb{D}$ and so must be an isometry. An infinitesimal displacement dz from 0 is mapped by T to the $T'(0)dz$ at z_o. So we must have

$$\lambda(0)|dz| = \lambda(z_o)|T'(0)dz| \ .$$

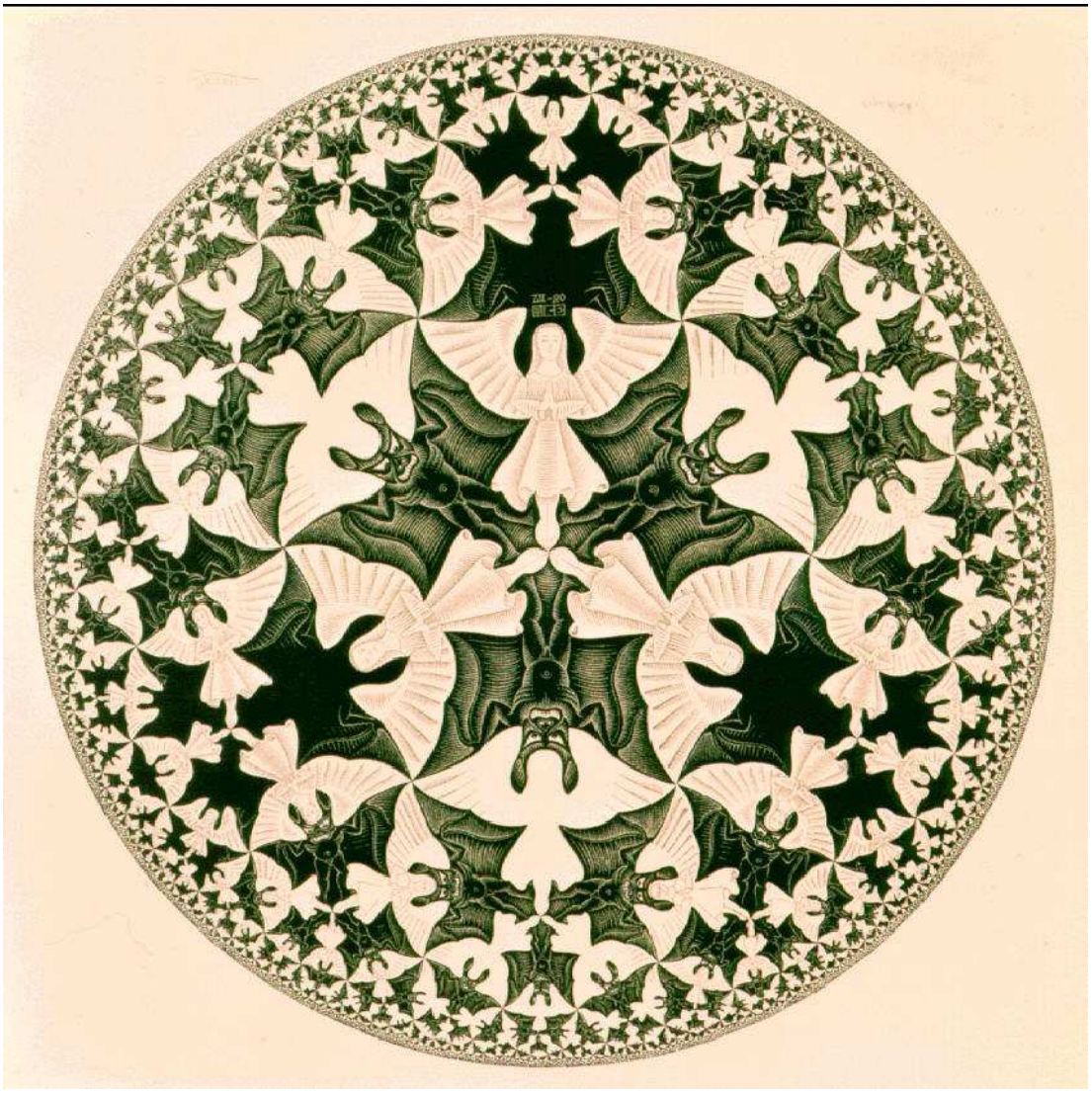
This means that

$$\lambda(0) = \lambda(z_o)|T'(0)| = \lambda(z_o)(1 - |z_o|^2) \ .$$

So the metric must be given by $ds = \frac{K}{1-|z|^2}|dz|$ at the point $z \in \mathbb{D}$. It is usual to set the constant K to 2. So we define the *hyperbolic density on* $\mathbb{D}$ to be

$$\lambda(z) = \frac{2}{1 - |z|^2} \qquad \text{for} \ \ z \in \mathbb{D} \ .$$

The hyperbolic density increases as we approach the boundary of the disc. This means that a constant Euclidean displacement increases in hyperbolic length as we approach the boundary. Similarly, a displacement of constant hyperbolic length has decreasing Euclidean length as we approach the boundary.



M.C. Escher, **Circle Limit IV** *(1960)*

In Escher's picture above, all of the angels are of the same hyperbolic size even though, to our Euclidean eyes, they appear to get smaller as we approach the boundary.

Having defined the hyperbolic density that gives the length of infinitesimal displacements, it is simple to define the hyperbolic metric. The *hyperbolic distance* $\rho(z_0, z_1)$ is the infimum of the lengths

$$L(\gamma) = \int_a^b \lambda(\gamma(t))\, |\gamma'(t)|\, dt = \int_a^b \frac{2|\gamma'(t)|}{1-|\gamma(t)|^2}\, dt$$

over all smooth curves $\gamma : [a, b] \to \mathbb{D}$ that have $\gamma(a) = z_0$ and $\gamma(b) = z_1$. We will see that there is a path γ from z_0 to z_1 that has shortest hyperbolic length. This is the *hyperbolic geodesic* from z_0 to z_1.

Lemma 10.2 Hyperbolic geodesics from the origin.
The hyperbolic geodesic from 0 *to* $z \in \mathbb{D}$ *is a radial line and it has hyperbolic length* $\log\left(\dfrac{1+|z|}{1-|z|}\right)$.

Proof:

It is clear that rotation about 0 preserves the hyperbolic length of any curve, so we may assume that $z = R \in [0, 1)$.

Use polar co-ordinates to write a curve $\gamma : [a, b] \to \mathbb{D}$ as $\gamma(t) = r(t)e^{i\theta(t)}$. We will assume that $\gamma(a) = 0$ and $\gamma(b) = z = R \geqslant 0$. Then the hyperbolic length of γ is more than the hyperbolic length of its radial part, for

$$L(\gamma) = \int_a^b \frac{2|\gamma'(t)|}{1 - |\gamma(t)|^2}\, dt = \int_a^b \frac{2|r'(t) + i\theta'(t)r(t)|}{1 - r(t)^2}\, dt \geqslant \int_a^b \frac{2|r'(t)|}{1 - r(t)^2}\, dt = L(r) \ .$$

Moreover, $|r'(t) + i\theta'(t)r(t)| = \sqrt{r'(t)^2 + r(t)^2\theta'(t)^2}$ so we have equality only when $\theta'(t)$ is identically 0. This shows that the radial path from 0 to z is the shortest path from 0 to z.

The length of this radial path is:

$$\int_0^R \frac{2}{1 - r^2}\, dr = \log\left(\frac{1 + R}{1 - R}\right)$$

so we see that

$$\rho(0, z) = \log\left(\frac{1 + |z|}{1 - |z|}\right) \ .$$

□

Theorem 10.3 Hyperbolic metric on the unit disc.
The expression

$$\rho(z_0, z_1) = \inf\{L(\gamma) : \gamma \text{ is a smooth curve in } \mathbb{D} \text{ from } z_0 \text{ to } z_1\}$$

gives a metric on $\mathbb{D}$ for which each Möbius transformation that maps $\mathbb{D}$ onto itself is an isometry. A geodesic for this metric is the arc of a circle orthogonal to the unit circle. This is the hyperbolic metric on $\mathbb{D}$.

Proof:
Proposition 10.1 shows that any Möbius transformation from $\mathbb{D}$ onto $\mathbb{D}$ must be of the form

$$T : z \mapsto \frac{az + b}{\overline{b}z + \overline{a}}$$

with $|a|^2 - |b|^2 = 1$. For this we have

$$T'(z) = \frac{1}{(\overline{b}z + \overline{a})^2} \ .$$

A straightforward calculation gives:

$$\frac{1}{1 - |T(z)|^2} = \frac{|\overline{b}z + \overline{a}|^2}{|\overline{b}z + \overline{a}|^2 - |az + b|^2} = \frac{|\overline{b}z + \overline{a}|^2}{1 - |z|^2} \ .$$

So we see that

$$\lambda(T(z))|T'(z)| = \frac{2}{1 - |T(z)|^2}|T'(z)| = \frac{2}{1 - |z|^2} = \lambda(z) \ .$$

This shows that $\rho(T(z_0), T(z_1)) = \rho(z_0, z_1)$.

The previous lemma certainly shows that $\rho(0, z) > 0$ whenever $z \neq 0$. Since we can always find a $T \in \text{Möb}(\mathbb{D})$ with $T(z_0) = 0$, it follows that $\rho(z_0, z_1) > 0$ whenever $z_0 \neq z_1$. The symmetry and triangle inequality for ρ are obvious, so ρ is a metric.

The lemma shows that the geodesic from $0 = T(z_0)$ to $T(z_1)$ is a radial line segment. This is a segment of a straight line through 0 orthogonal to the unit circle. The Möbius transformation T maps $\mathbb{T}$ onto itself and preserves angles, so the geodesic from z_0 to z_1 is an arc of the circle through z_0 and z_1 that is orthogonal to $\mathbb{T}$. □

11 THE HYPERBOLIC PLANE, II

11.1 The Hyperbolic Metric on a Half Plane.

Rather than working with the unit disc, we could work with any other disc Δ. We know that there is a Möbius transformation T that maps Δ onto $\mathbb{D}$ and we define the hyperbolic metric on Δ so that this is an isometry from Δ with this metric to $\mathbb{D}$ with the hyperbolic metric. This definition does not depend on which transformation T we choose. For, if S is another Möbius transformation that maps Δ onto $\mathbb{D}$, then $R = S \circ T^{-1}$ will be a Möbius transformation that maps $\mathbb{D}$ onto itself. This means that R is an isometry for the hyperbolic metric on $\mathbb{D}$. Hence T and S give the same metric on Δ.

An important example is when Δ is the upper half-plane $\mathbb{R}^2_+$. For this we can take T to be the map

$$T : z \mapsto i\left(\frac{z-i}{z+i}\right) .$$

(This is rotation of the Riemann sphere about an axis through ± 1.) The hyperbolic metric on $\mathbb{R}^2_+$ will then be $ds = \mu(z)|dz|$ for some density function μ. For T to be an isometry we must have

$$\mu(z)|dz| = \lambda(Tz)|T'(z)||dz| .$$

So

$$\mu(z) = \frac{2|T'(z)|}{1-|T(z)|^2} = \frac{2\left|\frac{-2}{(z+i)^2}\right|}{1-\left|\frac{z-i}{z+i}\right|^2} = \frac{1}{\mathrm{Im}(z)} .$$

The hyperbolic geodesics on the upper half-plane are half circles orthogonal to $\mathbb{R}$, together with half-lines parallel to the imaginary axis.

11.2 Inversions

Let γ be a hyperbolic geodesic for $\mathbb{D}$. Then γ is an arc of a circle Γ that is orthogonal to $\mathbb{T}$. Therefore inversion J in Γ maps $\mathbb{D}$ onto itself. We will call this *inversion in* γ.

Proposition 11.1 Inversions preserve the hyperbolic metric
Inversion in a hyperbolic geodesic preserves the hyperbolic metric.

Proof:
It is clear that the inversion $C : z \mapsto \overline{z}$ preserves the hyperbolic metric. Suppose that J is inversion in a geodesic γ. Then Proposition 9.5 shows that $C \circ J$ is a Möbius transformation. Both C and J maps the unit disc $\mathbb{D}$ onto itself, so $C \circ J$ will also do so. Hence, it must be a hyperbolic isometry. Therefore, $J = C \circ (C \circ J)$ preserves hyperbolic lengths. □

We can now consider the Möbius transformations that we get by composing two inversions in hyperbolic geodesics. The two geodesics may either cross at a point $w \in \mathbb{D}$, or meet at a point $w \in \mathbb{T}$, or not meet at all in the closure of $\mathbb{D}$. As at the end of §8 we see that the composition of these two inversions is elliptic, parabolic or hyperbolic in these three cases.

Let T be a Möbius transformation that maps $\mathbb{D}$ onto itself. Then Proposition 10.1 shows that $T : z \mapsto \dfrac{az+b}{\overline{b}z+\overline{a}}$ for some $a, b \in \mathbb{C}$ with $|a|^2 - |b|^2 = 1$. This fixes the points

$$\frac{i\mathrm{Im}(a) \pm \sqrt{\mathrm{Re}(a)^2 - 1}}{\overline{b}} .$$

The trace of the matrix is $2\mathrm{Re}(a)$, so Corollary 9.2 shows that T can not be loxodromic. If T is elliptic, then $-1 < \mathrm{Re}(a) < 1$, one fixed point w is in $\mathbb{D}$ and the other is $1/\overline{w}$. If T is parabolic, then $\mathrm{Re}(a) = \pm 1$ and there is a single fixed point on $\mathbb{T}$. If T is hyperbolic, then $\mathrm{Re}(a) > 1$ or $\mathrm{Re}(a) < -1$ and there are two fixed points both on $\mathbb{T}$.

12 FUCHSIAN GROUPS

We can give any disc $\Delta \subset \mathbb{P}$ the hyperbolic metric and take it as a model for the hyperbolic plane. The orientation preserving isometries for this metric are the Möbius transformations that map Δ onto itself. A subgroup G of Möb(Δ) is a *Fuchsian group* if it is discrete. In this section we will look at some examples of Fuchsian groups and see that there is a very great variety.

We will study Fuchsian groups by looking at the orbits of points in the hyperbolic plane. These orbits must be discrete, so their limit points will lie on the unit circle. These limit points form the *limit set* of the group. We will also try to find fundamental sets for the group acting on Δ. The images of the fundamental set under the group G cover all of the hyperbolic plane and form a *tessellation* of the plane. Often exhibiting this tessellation will be the simplest way to prove that the group is discrete.

12.1 Single generator Fuchsian groups

Proposition 12.1
A non-identity Möbius transformation T that maps a disc Δ onto itself is one of the following:

(a) **Elliptic** *with two fixed points, one in Δ and one in the complementary disc;*

(b) **Hyperbolic** *with two fixed points, both on the boundary $\partial\Delta$ of Δ in $\mathbb{P}$;*

(c) **Parabolic** *with one fixed point, which lies on $\partial\Delta$.*

Proof:

We know, from Theorem 9.1 , that T has either one or two fixed points in $\mathbb{P}$.

Suppose that T has two fixed points. Then T is conjugate to $M_k : z \mapsto kz$ for some $k \neq 0, 1$. When $|k| = 1$, the only discs M_k maps onto themselves are $\{z \in \mathbb{P} : |z| < r\}$ and $\{z \in \mathbb{P} : |z| > r\}$. Hence we are in case (a). When $k > 0$ ($k \neq 1$), the only discs mapped onto themselves are the half-planes $\{z \in \mathbb{P} : \mathrm{Re}(e^{i\theta})z) > 0\}$. Hence we are in case (b). (All other values for k give loxodromic transformations and these map no disc onto itself.)

Suppose that T has only one fixed point. Then T is conjugate to $P : z \mapsto z + 1$. The only discs mapped onto themselves by P are the half-planes $\{z \in \mathbb{P} : \mathrm{Im}(z) > c\}$. Hence we are in case (c). □

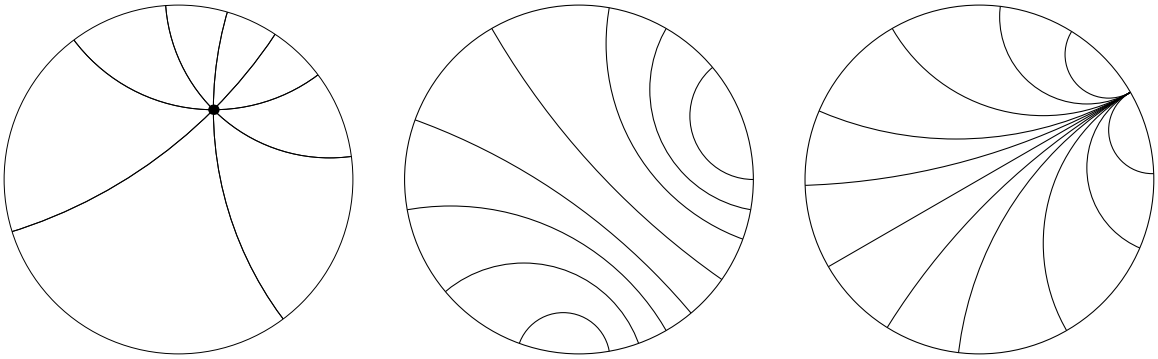
By using the conjugates M_k for $|k| = 1$, M_k for $k > 1$ and P, it is simple to see when the group $G = \langle T \rangle$ generated by a transformation T is discrete.

Elliptic:
By conjugating we get $M_k : z \mapsto kz$ for some k with $|k| = 1$. This is a rotation about 0 so the group it generates is discrete when it is of finite order. The orbits in this case are finite and a fundamental set is a sector of the unit disc. Consequently, the group generated by the elliptic transformation T is discrete when T is of finite order and a fundamental set is a sector from the fixed point bounded by two half geodesics. The tessellation by images of this fundamental set are shown in the first diagram below.

Hyperbolic:
By conjugating we get $M_k : z \mapsto kz$ for some $k > 0$ ($k \neq 1$). This generates an infinite cyclic group which is always discrete with the set $\{z : 1 \leqslant |z| < k\}$ as a fundamental set. Hence the group generated by the hyperbolic transformation T is discrete; the orbits are doubly infinite sequences $(T^n z_o)_{n=-\infty}^{n=\infty}$ that converge to the two fixed points of T; and a fundamental set is the region between two suitable disjoint geodesics.

Parabolic:
By conjugating we get $P : z \mapsto z + 1$. This generates an infinite cyclic group which is always discrete

with the set $\{z : 0 \leqslant \mathrm{Re}(z) < 1\}$ as a fundamental set. Hence the group generated by the parabolic transformation T is discrete; the orbits are doubly infinite sequences $(T^n z_o)_{n=-\infty}^{n=\infty}$ that converge to the single fixed point of T; and a fundamental set is the region between two suitable geodesics that both end at the fixed point.



Elliptic Hyperbolic Parabolic

In the case of a hyperbolic transformation T we call the hyperbolic geodesic joining the two fixed points the *axis of* T. It is clear that T maps its axis onto itself.

Exercise:

18. Show that a hyperbolic transformation T moves any point on its axis by a fixed hyperbolic distance along the axis. This distance is called the translation length of T. Show how to calculate the translation length from the trace of a matrix that represents T.

It is also useful to use Proposition 9.5 and consider the Möbius transformation as a composition of two inversions in hyperbolic geodesics.

12.2 Triangle Groups

Let U be a "triangular" region bounded by three disjoint geodesics $\gamma_1, \gamma_2, \gamma_3$. Let J_k be inversion in the geodesic γ_k. Then the Möbius transformations: $A = J_2 \circ J_1$, $B = J_3 \circ J_2$ generate a group G. Note that

$$J_1 \circ J_2 = A^{-1} \ ; \ \ J_2 \circ J_3 = B^{-1} \ ; \ \ J_3 \circ J_1 = B \circ A \ ; \ \ J_1 \circ J_3 = A^{-1} \circ B^{-1}$$

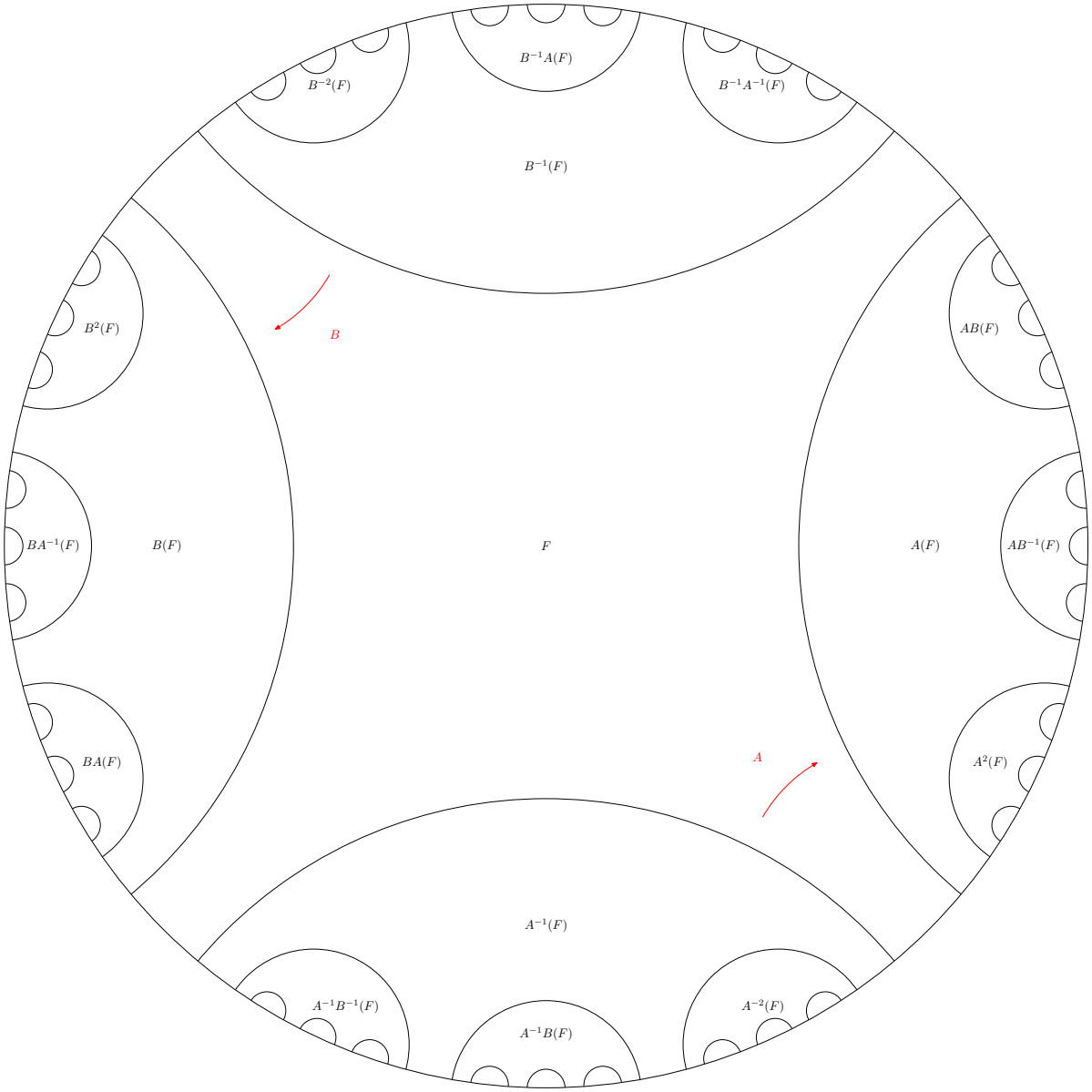
so any product of an even number of the inversions is an element of G. The set F that is the union of the closure of U in $\mathbb{D}$ and $J_2(U)$ is a fundamental set for G. (The closure of U is a fundamental set for the larger group generated by J_1, J_2 and J_3.) The quotient $\mathbb{D}/G$ is homeomorphic to a sphere with three holes: a "pair of pants".

G is a free group generated by A and B. For suppose that we can write $g \in G$ as a product

$$g = A^{k(1)} B^{k(2)} A^{k(3)} \ldots B^{k(N)} \qquad \text{with each } k(j) \neq 0 \ .$$

If $k(1) > 0$ (so the product for g begins with an A) then the copy $g(F)$ of the fundamental set is separated from F by the geodesic γ_1. Similarly, if g begins with A^{-1} (or B or B^{-1}, then $g(F)$ is separated from F by $J_2(\gamma_1)$ (or γ_3 or $J_2(\gamma_3)$). Hence we can only have $g = I$ when the product for g is trivial and does not begin with any of A, A^{-1}, B, B^{-1}.

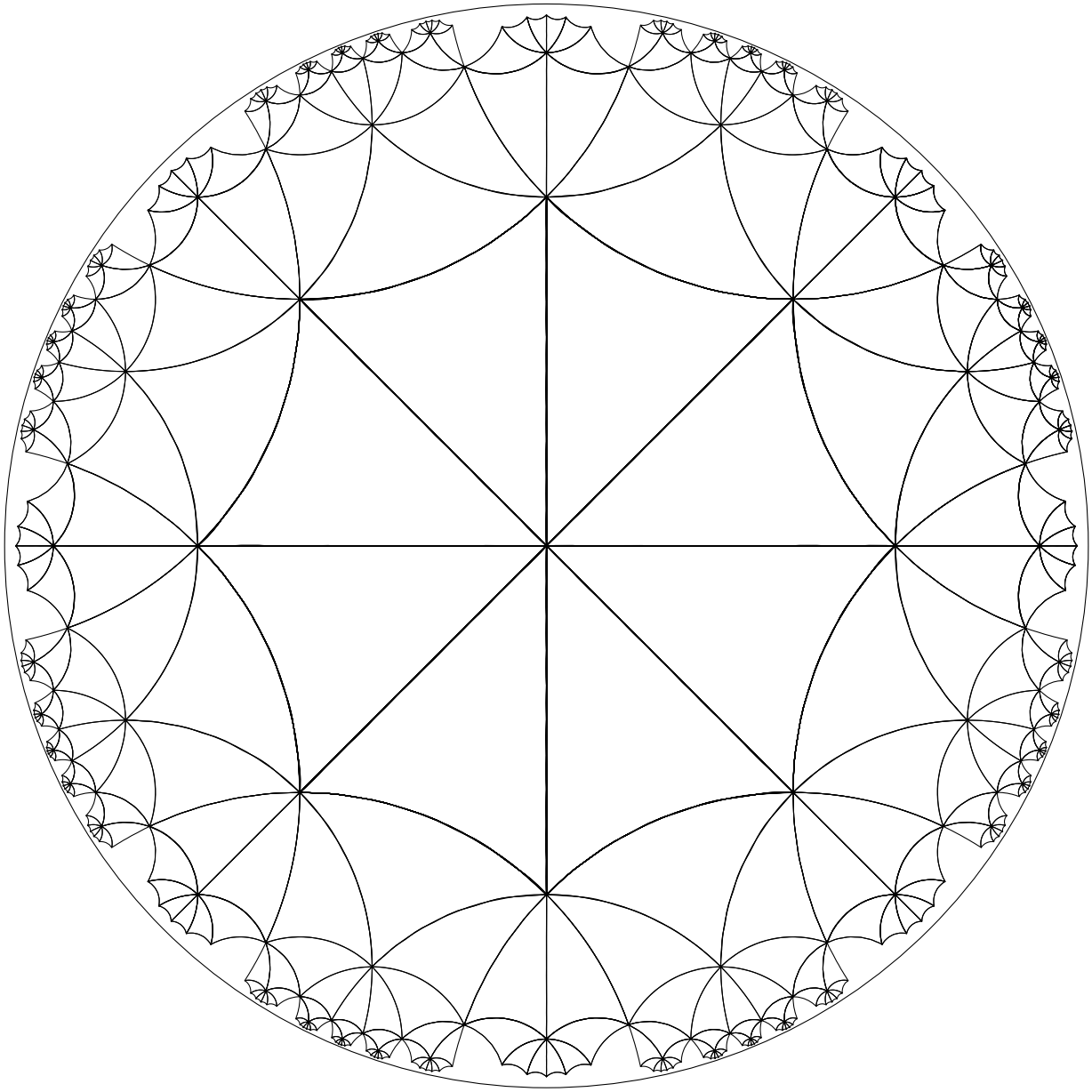
Any orbit of G contains exactly one point in each copy $g(F)$ of the fundamental set. The fixed points of each non-identity element of G clearly give points in the limit set. The limit set is like a Cantor set: It is a subset of $\partial\mathbb{D}$ but omits points in the open intervals where F meets $\partial\mathbb{D}$. Similarly, it omits points where $A(F), B(F), AB(F)$, etc meet $\partial\mathbb{D}$.



The group G is certainly discrete. For suppose that (g_n) is a sequence of non-identity transformations in G that converge to I. Then $g_n(z_o) \to z_o$ as $n \to \infty$. When we choose z_o to be an interior point of the fundamental set F we see that this is impossible.

We can also consider the groups generated by inversions in the sides of a triangle where the sides cross. Then the Möbius transformations A, B are elliptic rather than hyperbolic. To get a Fuchsian group these elliptic elements must be of finite order, so the angles of the triangle must be π/k for some integers k. The sum of the angles of a hyperbolic triangle is strictly less than π, so there are many triangles where these conditions are satisfied. For example we may construct a triangle with all angles $\pi/4$ and obtain a tessellation as shown below.

The limit set in these cases is all of the unit circle.



13 THE MODULAR GROUP

In many ways the most interesting triangle group is when the triangle has all of its angles 0. The three vertices all lie on the boundary $\partial\mathbb{D}$ and the compositions of two inversions is parabolic. This group is important in may branches of Mathematics. Rather than studying it directly, we will look at a closely related group first: the Modular group.

The *Modular group* M is the group of Möbius transformations of the form

$$z \mapsto \frac{az+b}{cz+d}$$

with $a, b, c, d \in \mathbb{Z}$ and $ad - bc = 1$. Hence M is the quotient $\mathrm{PSL}(2,\mathbb{Z})$ of the matrix group $\mathrm{SL}(2,\mathbb{Z})$ by $\{I, -I\}$. The modular group acts as a group of hyperbolic isometries on the upper half-plane $\mathbb{R}^2_+$. This group is certainly discrete because the integers are discrete. We wish to find a fundamental set for the modular group and the corresponding tessellation.

Consider the orbit Ω of a point $w \in \mathbb{R}^2_+$. For a transformation $T : z \mapsto \frac{az+b}{cz+d}$ in the modular group we have

$$\mathrm{Im}(T(w)) = \mathrm{Im}\left(\frac{(aw+b)(c\overline{w}+d)}{|cw+d|^2}\right) = \left(\frac{1}{|cw+d|^2}\right)\mathrm{Im}(w)\ .$$

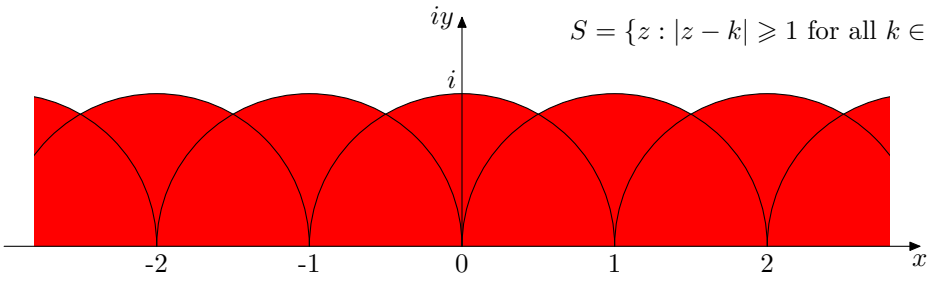
Hence,

$$\mathrm{Im}(T(w)) > \mathrm{Im}(w) \qquad \Leftrightarrow \qquad 1 > |cw+d|\ .$$

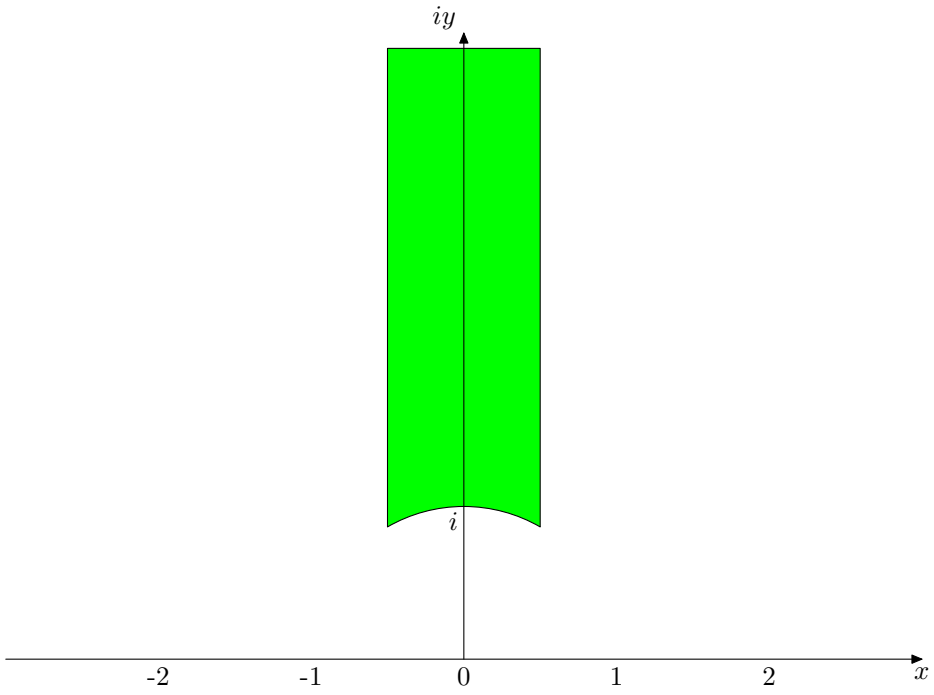
The points $cw + d$ all lie in the lattice $\mathbb{Z}w + \mathbb{Z}$, so only a finite number can lie within the unit disc. Hence there are only a finite number of pairs (c, d) with $\mathrm{Im}(T(w)) > \mathrm{Im}(w)$. In particular, there are only finitely many values of $\mathrm{Im}(z)$ for $z \in \Omega$ which are greater than $\mathrm{Im}(w)$. Consequently, there is a point $w_o \in \Omega$ where $\mathrm{Im}(w_o)$ is maximal.

There are always infinitely many points in Ω where the imaginary part is greatest. For the matrix $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ is in $\mathrm{SL}(2,\mathbb{Z})$ and so $z \mapsto z+1$ is in the modular group. Consequently, $w_o + k$ is in Ω for each $k \in \mathbb{Z}$.

For each pair of coprime integers c, d we can find integers a, b with $ad - bc = 1$, so the matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is in the modular group. Note that the point w_o must satisfy $|cw_o + d| \geqslant 1$ for each such pair (c, d). This is the same as $\left|w_o + \frac{d}{c}\right| \geqslant \frac{1}{c}$ so we see that w_o must lie outside the shaded region below.



This means that, for any point $w \in \mathbb{R}^2_+$, we can find a point w_o in the orbit of w that lies in the unshaded region or on its boundary. Since $z \mapsto z+1$ is in the modular group, we can choose w_o to lie in the region $\{z = x + iy \in \mathbb{R}^2_+ : -\frac{1}{2} < x \leqslant \frac{1}{2}\ ,\ \ |z| \geqslant 1\}$ shaded below.



This set: $\{z = x + iy \in \mathbb{R}^2_+ : -\frac{1}{2} < x < \frac{1}{2}\ ,\ \ |z| > 1\}$, together with part of the boundary is a fundamental set for the modular group.

Proposition 13.1 Fundamental set for the modular group.
For every point z in $\mathbb{R}^2_+$, there is a transformation T in the modular group with $T(z)$ lying in the set

$$F = \{z = x + iy \in \mathbb{R}^2_+ : -\tfrac{1}{2} \leqslant x \leqslant \tfrac{1}{2}\ ,\ \ |z| \geqslant 1\}\ .$$

Moreover, two points in F are in the same orbit for the modular group if and only if they are either

$$\begin{array}{rll} -\frac{1}{2} + iy \text{ and } \frac{1}{2} + iy & \text{with } \ y \geqslant \frac{1}{2}\sqrt{3} & \text{or} \\ ie^{i\theta} \text{ and } ie^{-i\theta} & \text{with } 0 \leqslant \theta \leqslant \frac{1}{3}\pi & \end{array}$$

Proof:

Let $A : z \mapsto z + 1$ and $B : z \mapsto -1/z$. Both of these are in the modular group.

If z lies in the set

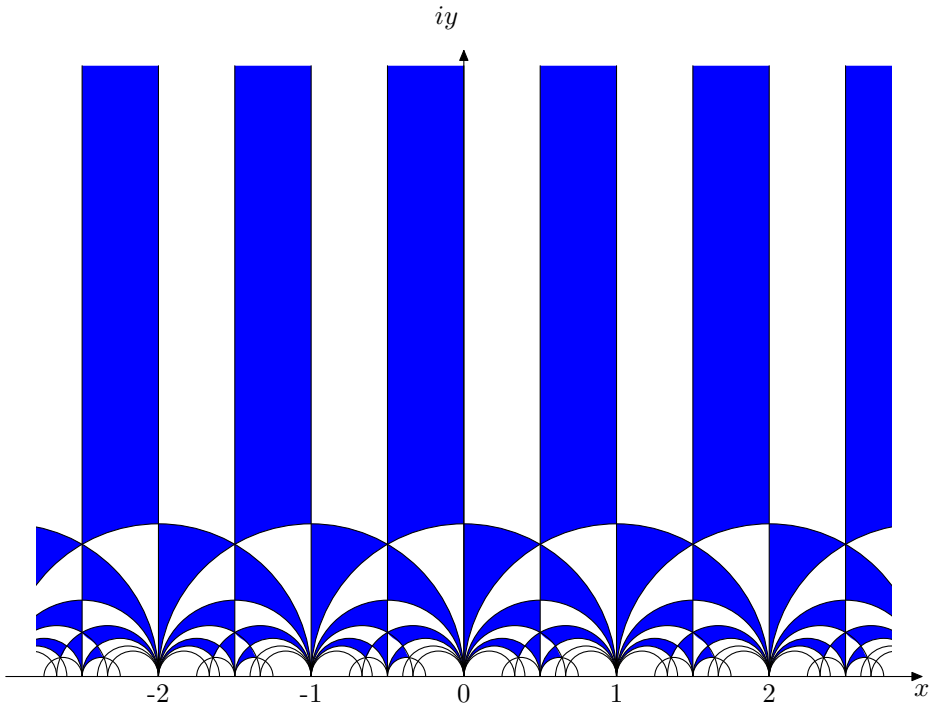
$$S = \{w : |w - k| \geqslant 1 \text{ for all } k \in \mathbb{Z}\}$$

then there is an integer k with $A^k(z) = z + k \in F$. If z lies outside this set, then we have shown that there is an element z' in the orbit of z that lies within S.

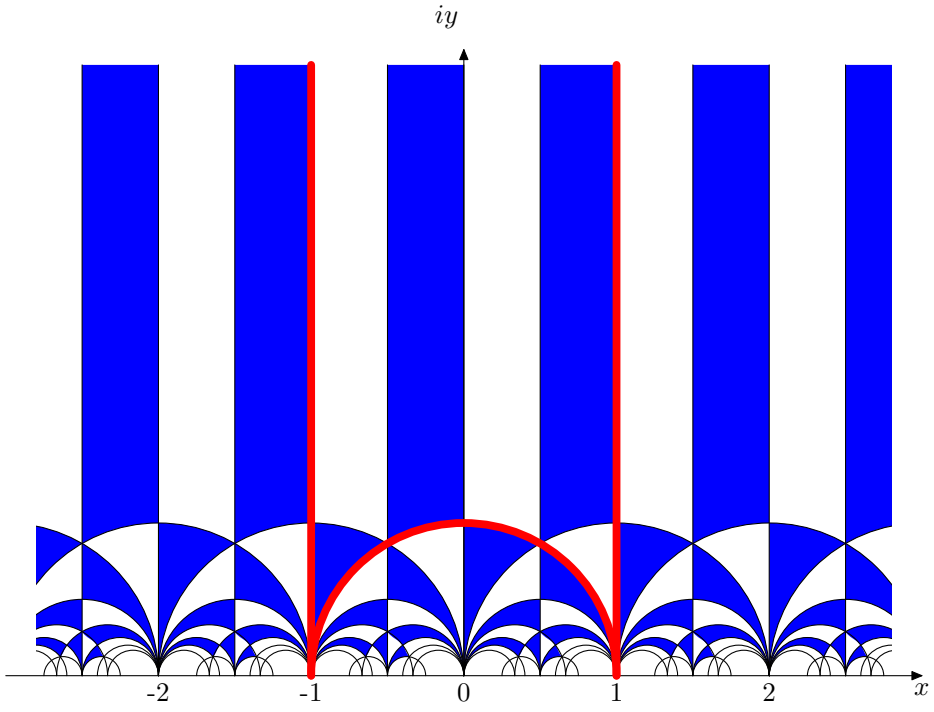
Suppose that z and z' are in the same orbit and both lie within F. Then both have maximal imaginary part for that orbit, so their imaginary parts are equal. Since they are in the same orbit, we must have $z' = \dfrac{az + b}{cz + d}$ for $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2, \mathbb{Z})$. Since the imaginary parts of z and z' are equal, we must have $|cz + d| = 1$. Hence, either $c = 0$ and $d = \pm 1$ or $c = \pm 1$ and $d = 0$. In the first case, $z' = z + k = A^k(z)$ for some integer k and we are in the first case of the proposition. In the second case, $z' = k - \dfrac{1}{z} = A^k B(z)$ and we are in the second case. □

It is useful to divide the fundamental set for the modular group into two. In the diagram below, the set is divided into two triangles, one shaded and the other unshaded. The modular group permutes

the shaded triangles and also permutes the unshaded ones. The inversions in the three sides of one of these triangles gives a larger group with the modular group as an index two subgroup.



Finally, the diagram below shows that there we can join together 6 of these triangles to form an *ideal* triangle which has all three vertices on the boundary of $\mathbb{R}^2_+$. Inversions in the sides of this ideal triangle give a subgroup of the group generated by inversions in sides of the smaller triangles.



14 HYPERBOLIC 3-SPACE

In this section we will define a hyperbolic metric on the unit ball in $\mathbb{R}^3$ so that the orientation preserving isometries are the full group of Möbius transformations. To do this, we first define the metric, then consider inversions, and finally show that every Möbius transformation is the composition of inversions.

14.1 The Hyperbolic Metric

Let B^3 be the unit ball $\{\boldsymbol{x} \in \mathbb{R}^3 : ||\boldsymbol{x}|| < 1\}$ in Euclidean 3-space. The *hyperbolic density* on B^3 is

$$\lambda(\boldsymbol{x}) = \frac{2}{1 - ||\boldsymbol{x}||^2} \ .$$

The hyperbolic length of a smooth curve $\gamma : [a, b] \to B^3$ is then

$$L(\gamma) = \int_a^b \lambda(\gamma(t)) \, ||\gamma'(t)|| \, dt = \int_a^b \frac{2||\gamma'(t)||}{1 - ||\gamma(t)||^2} \, dt \ .$$

The hyperbolic metric ρ on B^3 is defined by

$$\rho(\boldsymbol{x}_o, \boldsymbol{x}_1) = \inf \left\{ L(\gamma) : \gamma \text{ is a smooth curve in } B^3 \text{ from } \boldsymbol{x}_o \text{ to } \boldsymbol{x}_1 \right\} \ .$$

A curve that attains this infimum is a *hyperbolic geodesic from* $\boldsymbol{x}_o$ *to* $\boldsymbol{x}_1$. The arguments used for the hyperbolic metric on the unit disc (Lemma 10.2 and Theorem 10.3) show that:

Proposition 14.1 Hyperbolic metric on B^3
The hyperbolic metric ρ is a metric on the unit ball B^3. Moreover, the hyperbolic geodesic from the origin $\mathbf{0}$ *to any point $\boldsymbol{x} \in B^3$ is a radial path with hyperbolic length* $\log\left(\dfrac{1 + ||\boldsymbol{x}||}{1 - ||\boldsymbol{x}||}\right)$.

□

The disc $B^3 \cap \{\boldsymbol{x} : x_3 = 0\} = \{(x_1, x_2, 0) : |x_1|^2 + |x_2|^2 < 1\}$ can be identified with the unit disc $\mathbb{D}$ in $\mathbb{C}$ by letting $(x_1, x_2, 0)$ correspond to $x_1 + ix_2$. Then the hyperbolic metric on B^3 restricts to give the plane hyperbolic metric on the $\mathbb{D}$. We will see shortly that much more is true: the restriction of the hyperbolic metric to the intersection of B^3 with a sphere orthogonal to the unit sphere gives a hyperbolic plane metric.

14.2 Inversion

We are used to thinking of extending the complex plane $\mathbb{C}$ by adding a point ∞ and identifying the resulting space with the Riemann sphere $\mathbb{P}$. We can also do this in higher dimensions. The N-dimensional Euclidean space $\mathbb{R}^N$ is extended by adjoining a point ∞ to obtain $\mathbb{R}^N_\infty$. This is homeomorphic to the unit sphere S^N in $\mathbb{R}^{N+1}$ and we can use stereographic projection to identify $\mathbb{R}^N_\infty$ with S^N. We will only consider the 3-dimensional case although the results apply in higher dimensions.

As in the case of the Riemann sphere we will define circles to include straight lines, and spheres in $\mathbb{R}^3_\infty$ to include planes. So a *sphere* in $\mathbb{R}^3_\infty$ is either of the form

$$S(\boldsymbol{c}, r) = \{\boldsymbol{x} \in \mathbb{R}^3 : ||\boldsymbol{x} - \boldsymbol{c}|| = r\}$$

for $\boldsymbol{c} \in \mathbb{R}^3$ and $r > 0$, or else

$$\Pi(\boldsymbol{u}, t) = \{\boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{x} \cdot \boldsymbol{u} = t\} \cup \{\infty\}$$

for a unit vector $\boldsymbol{u}$ and $t \in \mathbb{R}$. We call $\boldsymbol{c}$ and r the *Euclidean centre* and *Euclidean radius* of the sphere $S(\boldsymbol{c}, r)$, while $\boldsymbol{u}$ is a unit normal to the plane $\Pi(\boldsymbol{u}, t)$.

We define inversion is a sphere Σ in $\mathbb{R}^3_\infty$ exactly as in §9.2 . Two distinct points $\boldsymbol{x}, \boldsymbol{x}'$ are *inverse points for* Σ if every circle orthogonal to Σ through $\boldsymbol{x}$ also passes through $\boldsymbol{x}'$. Also, when $\boldsymbol{x} \in \Sigma$, we say that $\boldsymbol{x}$ and $\boldsymbol{x}$ itself are inverse points.

Proposition 14.2 Inversion in spheres
For each sphere Σ in $\mathbb{R}^3_\infty$ and each point $\boldsymbol{x} \in \mathbb{R}^3_\infty$ there is a unique point $J(\boldsymbol{x})$ with $\boldsymbol{x}$ and $J(\boldsymbol{x})$ inverse points for Σ.

The map J is called *inversion in* Σ. It is an involution that reverses orientation.

Proof:

We already know from Proposition 9.4 that inversion in a circle maps the plane to itself. We will use this to extend the result to higher dimensions.

Suppose first that $\Sigma = S(\boldsymbol{c}, r)$. Any plane π through $\boldsymbol{x}$ and $\boldsymbol{c}$ cuts Σ in a circle σ and we know from Proposition 9.4 that there is a point $\boldsymbol{x}'$ in π with $\boldsymbol{x}$ and $\boldsymbol{x}'$ inverse points for σ. Indeed we know that the inverse point is given by

$$J(\boldsymbol{x}) = \boldsymbol{c} + \left(\frac{r^2}{||\boldsymbol{x} - \boldsymbol{c}||^2} \right) (\boldsymbol{x} - \boldsymbol{c}) \ .$$

This expression makes sense for any point $\boldsymbol{x} \in \mathbb{R}^3_\infty$, so we only need to show that it does have the properties we require. (Note that we should interpret this formula as saying that $J(\boldsymbol{c}) = \infty$ and $J(\infty) = \boldsymbol{c}$.)

If γ is any circle through $\boldsymbol{x}$ that crosses Σ orthogonally, then there is a plane π through γ. In this plane, we know that $\boldsymbol{x}$ and $J(\boldsymbol{x})$ are inverse points for the circle $\sigma = \Sigma \cap \pi$ in the plane π. Consequently, γ must pass through $J(\boldsymbol{x})$.

An entirely similar argument applies when Σ is a plane $\Pi(\boldsymbol{u}, t)$. Then we have

$$J(\boldsymbol{x}) = \boldsymbol{x} + 2(t - \boldsymbol{x} \cdot \boldsymbol{u})\boldsymbol{u} \ .$$

□

We already know a lot about inversion in 2-dimensions. Since inversion in higher dimensions is defined in terms of inversions in 2-dimensions, we readily obtain a variety of results about inversion.

Proposition 14.3 Inversion preserves spheres.
Let J be inversion in a sphere Σ. Then J maps any sphere U in $\mathbb{R}^N_\infty$ onto another sphere.

Proof:

Let Σ be the sphere $S(\boldsymbol{c}, r)$ and U the sphere $S(\boldsymbol{d}, s)$. Let ℓ be the straight line through $\boldsymbol{c}$ and $\boldsymbol{d}$. Then any plane π through ℓ cuts Σ is a circle σ and U in a circle u. We know that inversion in the circle σ sends u to another circle, so we see that $J(U)$ cuts π in a circle. This is true for every plane π through ℓ, so $J(U)$ must be a sphere.

A similar but simpler argument applies when Σ or U are planes. □

Corollary 14.4 Inversion preserves circles.
Let J be inversion in a sphere Σ. Then J maps any circle γ in $\mathbb{R}^N_\infty$ onto another circle.

Proof:

The circle γ is the intersection of two spheres. Inversion maps each of these to another sphere and the intersection of these two spheres is again a circle. □

Proposition 14.5 Inversion preserves angles.
Let J be inversion in a sphere Σ. If two curves α and β in $\mathbb{R}^3_\infty$ cross at an angle θ, then $J(\alpha)$ and $J(\beta)$ also cross at an angle θ.

Proof:

Let J be inversion in the sphere $S(\boldsymbol{c}, r)$. We wish to calculate the derivative $J'(\boldsymbol{x})$. To simplify the algebra, translate and enlarge the sphere so that it becomes the unit sphere $S(\boldsymbol{0}, 1)$. Note that this does not alter any angles.

Now $J : \boldsymbol{x} \mapsto \left(\frac{1}{||\boldsymbol{x}||^2}\right) \boldsymbol{x}$ satisfies

$$\begin{aligned} J(\boldsymbol{x}+\boldsymbol{h}) &= \left(\frac{1}{||\boldsymbol{x}||^2 + 2\boldsymbol{x}\cdot\boldsymbol{h} + ||\boldsymbol{h}||^2}\right)(\boldsymbol{x}+\boldsymbol{h}) \\ &= J(\boldsymbol{x}) + \left(\frac{1}{||\boldsymbol{x}||^2}\right)\left(\boldsymbol{h} - \left(\frac{2\boldsymbol{x}\cdot\boldsymbol{h}}{||\boldsymbol{x}||^2}\right)\boldsymbol{x}\right) + o(\boldsymbol{h}) \ . \end{aligned}$$

So we see that J is differentiable at $\boldsymbol{x}$ with

$$J'(\boldsymbol{x}) : \boldsymbol{h} \mapsto \left(\frac{1}{||\boldsymbol{x}||^2}\right)\left(\boldsymbol{h} - \left(\frac{2\boldsymbol{x}\cdot\boldsymbol{h}}{||\boldsymbol{x}||^2}\right)\boldsymbol{x}\right) \ .$$

This is a scalar multiple of reflection in the plane orthogonal to $\boldsymbol{x}$ and so preserves angles (and reverses orientation). □

Exercise:

19. Show from the formula

$$J(\boldsymbol{x}) = \boldsymbol{c} + \left(\frac{r^2}{||\boldsymbol{x}-\boldsymbol{c}||^2}\right)(\boldsymbol{x}-\boldsymbol{c}) \ .$$

for inversion in the sphere $S(\boldsymbol{c}, r)$ that inversion maps a sphere to another sphere.

---

15 EXTENDING MÖBIUS TRANSFORMATIONS TO HYPERBOLIC SPACE

15.1 Inversions and the hyperbolic metric

Let Σ be a sphere that is orthogonal to the unit sphere S^2 in $\mathbb{R}^3$. Then Σ meets S^2 in a circle σ. Proposition 14.3 shows that inversion J in Σ fixes σ and maps the unit sphere to another sphere. Proposition 14.5 shows that this image sphere is also orthogonal to Σ. Hence J maps the unit sphere onto itself. Therefore, J maps the ball B^3 onto itself.

Conversely, any circle σ on the unit sphere S^2 is the intersection of S^2 with a sphere Σ orthogonal to S^2.

Throughout this section we will only be concerned with spheres Σ orthogonal to the unit sphere, and the circle $\sigma = \Sigma \cap S^2$ where it meets the unit sphere.

Proposition 15.1 Inversions are hyperbolic isometries.
Let J be inversion in a sphere Σ orthogonal to the unit sphere S^2. Then J is an isometry for the hyperbolic metric but reverses orientation.

Proof:

Let $\boldsymbol{x}$ be a point of B^3 and choose any plane π through the origin and $\boldsymbol{x}$. Then π intersects Σ in a circle γ and the map J acts on π as inversion in γ. Proposition 11.1 shows that inversion in γ is an isometry for the hyperbolic metric on the disc $\pi \cap B^3$. Hence J must be an isometry at $\boldsymbol{x}$.

We have already seen that the derivative of J is orientation reversing. □

Exercise:

20. Let J be inversion in a sphere Σ and Q inversion in the unit sphere S^2. Show that Σ is orthogonal to S^2 if and only if $J \circ Q = Q \circ J$.

Lemma 15.2
For each point $\boldsymbol{a} \in B^3$ there is an inversion J in a sphere orthogonal to S^2 that interchanges the origin and $\boldsymbol{a}$.

Proof:

If $\boldsymbol{a} = \mathbf{0}$, then inversion in any plane through the origin (that is reflection in such a plane) will do.

Otherwise, let $Q(\boldsymbol{a}) = \dfrac{\boldsymbol{a}}{||\boldsymbol{a}||^2}$ and set

$$\Sigma = \left\{ \boldsymbol{x} : ||\boldsymbol{x} - Q(\boldsymbol{a})|| = \frac{\sqrt{1 - ||\boldsymbol{a}||^2}}{||\boldsymbol{a}||} \right\} .$$

Then it is simple to check that inversion in Σ maps $\mathbf{0}$ to $\boldsymbol{a}$. □

Proposition 15.3 Hyperbolic geodesics
Between any two distinct points of B^3 there is an unique path with shortest hyperbolic length. This is a section of a circle orthogonal to the unit sphere S^2.

Proof:

Suppose that $\boldsymbol{a}$ and $\boldsymbol{b}$ are two distinct points in the ball B^3. We already know that the radial path is the shortest path from $\boldsymbol{a}$ to $\boldsymbol{b}$ when $\boldsymbol{a} = \mathbf{0}$ because of Proposition 14.1 .

For any other value of $\boldsymbol{a}$, the lemma shows that there is an inversion J in a sphere orthogonal to S^2 with $J(\mathbf{a}) = \mathbf{0}$. Since this inversion is a hyperbolic isometry, we see that the shortest path from $\boldsymbol{a}$ to $\boldsymbol{b}$ must be the image of a radial path under J. Now Corollary 14.4 completes the proof. □

Suppose now that T is a Möbius transformation. This maps the unit sphere S^2 to itself. We know from §9 that T can be written as the composition of inversions in an even number of circles, say $\sigma_1, \sigma_2, \ldots, \sigma_N$. For each of these circles, there is a sphere Σ_n orthogonal to S^2 which intersects S^2 in the circle σ_n. Let J_n be inversion in the sphere Σ_n. Then the composition $J_N \circ J_{N-1} \circ \ldots \circ J_3 \circ J_2 \circ J_1$ acts on the sphere S^2 as the Möbius transformation T. However, this composition gives a map $\widetilde{T}$ from all of $\mathbb{R}^3_\infty$ to itself. This map agrees with T on the sphere S^2 and maps the ball B^3 onto itself. So we have a way to extend the Möbius transformation T to the ball. This extension was introduced by Poincaré and is often called the *Poincaré extension* of the Möbius transformation.

This extension is unique. For suppose that $\boldsymbol{x}$ is any point of B^3. Choose a circle γ through $\boldsymbol{x}$ that crosses S^2 orthogonally at two points $\boldsymbol{a}$ and $\boldsymbol{b}$. Then each inversion J in a sphere orthogonal to S^2 maps γ to another circle orthogonal to S^2. Consequently, $\widetilde{T}(\boldsymbol{x})$ lies on the circle orthogonal to S^2 that joins the two points $T(\boldsymbol{a})$ and $T(\boldsymbol{b})$ in S^2. This is true for every choice of γ so we see that the point $\widetilde{T}(\boldsymbol{x})$ is completely determined by the map T acting on the sphere. In particular, if we write T as a composition of inversions in two different ways we must obtain the same extension $\widetilde{T}$.

Proposition 15.4 Extensions of Möbius transformations.
For every Möbius transformation $T : S^2 \to S^2$ the extension $\widetilde{T} : \mathbb{R}^3_\infty \to \mathbb{R}^3_\infty$ maps the unit ball B^3 onto itself and is an orientation preserving isometry for the hyperbolic metric.

Proof:

Proposition 15.1 shows that each inversion J_n is an isometry for the hyperbolic metric, hence the composition $\widetilde{T}$ is also. Since T is the composition of an even number of inversions, we see that $\widetilde{T}$ is the composition of an even number of the orientation reversing isometries J_n. Hence $\widetilde{T}$ is orientation preserving. □

Now we will prove that every orientation preserving isometry of B^3 for the hyperbolic metric is an extension of a Möbius transformation.

Theorem 15.5 Möbius transformations as isometries of hyperbolic 3-space.
Every orientation preserving isometry of hyperbolic 3-space B^3 is $\widetilde{T}$ for some Möbius transformation $T : S^2 \to S^2$.

Proof:

Suppose that $A : B^3 \to B^3$ is an orientation preserving isometry for the hyperbolic metric on B^3. Then $A(\mathbf{0}) \in B^3$ so Lemma 15.2 gives an inversion J in a sphere orthogonal to S^2 with $J(A(\mathbf{0})) = \mathbf{0}$. Hence, $A' = J \circ A$ is an isometry of B^3 that fixes $\mathbf{0}$.

For each unit vector $\mathbf{u} \in S^2$ we know that the path

$$\gamma_{\boldsymbol{u}} : [0, 1) \mapsto t\boldsymbol{u}$$

is a hyperbolic geodesic with $\rho(\mathbf{0}, \gamma_{\boldsymbol{u}}(t)) = \log(1 + t)/(1 - t)$. Hence, $A' \circ \gamma_{\boldsymbol{u}}$ must also be a hyperbolic geodesic. Since it starts at the origin, we must have

$$A'(t\boldsymbol{u}) = t\boldsymbol{v}$$

for some unit vector $\boldsymbol{v}$. Write $\boldsymbol{v} = \alpha(\boldsymbol{u})$.

Now observe that

$$\lim_{t\to 0+} \frac{\rho(t\boldsymbol{u}_1, t\boldsymbol{u}_2)}{t} = 2||\boldsymbol{u}_1 - \boldsymbol{u}_2||$$

for any unit vectors $\boldsymbol{u}_k$. Since A' is an isometry, this shows that

$$||\alpha(\boldsymbol{u}_1) - \alpha(\boldsymbol{u}_2)|| = ||\boldsymbol{u}_1 - \boldsymbol{u}_2|| \ .$$

It follows that α is an orthogonal linear map in O(3). This shows that $A' : \mathbb{R}^3_\infty \to \mathbb{R}^3_\infty$ is orthogonal.

Finally we know that every orthogonal linear map is the composition of reflections in plane through the origin. These are inversions, so we see that A', and hence A, is the composition of inversions in spheres orthogonal to S^2. □

We have now proved that the groups Möb and $\mathrm{Isom}^+(B^3)$ are isomorphic with a Möbius transformation T corresponding to the extension $\widetilde{T}$. This means that the Möbius group Möb acts on hyperbolic 3-space B^3. Usually we will not distinguish between T and its extension $\widetilde{T}$.

15.2 The upper half-space.

We can redo the arguments above for any ball in $\mathbb{R}^3_\infty$ and obtain a hyperbolic metric on the ball for which the orientation preserving isometries are the Möbius transformations. The most important example is when the ball is the upper half-space:

$$\mathbb{R}^3_+ = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_3 > 0\} \ .$$

The boundary of this is the extended complex plane $\mathbb{C}_\infty = \mathbb{R}^2_\infty$. We can show that any Möbius transformation acting on this boundary extends to an orientation preserving isometry of the upper half-space for the hyperbolic metric with density:

$$\lambda(\boldsymbol{x}) = \frac{1}{x_3} \ .$$

We can also deduce the results for the upper half-space directly from those for the ball B^3 for inversion in the sphere

$$\Sigma = \left\{\boldsymbol{x} : ||\boldsymbol{x} + \boldsymbol{e}_3|| = \sqrt{2}\right\} \qquad \text{where } \boldsymbol{e}_3 = (0, 0, 1)$$

maps the upper half-space onto the ball and *vice versa*.

16 ISOMETRIES OF $\mathbb{H}^3$

We have seen how to put a hyperbolic metric on the unit ball B^3 in $\mathbb{R}^3$ or the upper half-space $\mathbb{R}^3_+$. We will denote both of these by $\mathbb{H}^3$ and call them *hyperbolic* 3*-space*. The orientation preserving isometries for hyperbolic 3-space have been identified with the group of Möbius transformations acting on the boundary $\partial\mathbb{H}^3$. In this section we wish to study these isometries in more detail.

16.1 Examples in Hyperbolic Geometry

The hyperbolic metric is given by:

$$ds = \frac{2}{1-||\boldsymbol{x}||^2}||d\boldsymbol{x}|| \quad \text{on } B^3 \qquad \text{and} \qquad ds = \frac{1}{x_3}||d\boldsymbol{x}|| \quad \text{on } \mathbb{R}^3_+ \ .$$

The hyperbolic geodesics are the arcs of circles orthogonal to the boundary $\partial\mathbb{H}^3$.

The plane $\{\boldsymbol{x} : x_3 = 0\}$ meets the ball B^3 in the unit disc and the hyperbolic metric on B^3 restricts to the hyperbolic plane metric on this disc. A similar result holds for the intersection of any other sphere orthogonal to $\partial\mathbb{H}^3$ with $\mathbb{H}^3$. For suppose that Σ is the intersection with B^3 of a sphere orthogonal to ∂B^3. Then the boundary of Σ is a circle $\sigma \subset \partial B^3$. We know that there is a Möbius transformation T that maps this circle σ to the unit circle $\mathbb{T}$. Since T acts isometrically on B^3, it must map Σ to the unit disc. Hence Σ with the hyperbolic metric is isometric to the hyperbolic plane. We call such an intersection of a sphere orthogonal to $\partial\mathbb{H}^3$ with $\mathbb{H}^3$ a *hyperbolic plane in* $\mathbb{H}^3$.

Two hyperbolic planes that meet in $\mathbb{H}^3$ intersect in a hyperbolic geodesic. We think of geodesics as the straight lines for hyperbolic geometry.

In order to develop our sense of what hyperbolic 3-space is like, we will prove a series of simple results. In all of these we choose whichever model (B^3 or $\mathbb{R}^3_+$) is easiest and apply isometries to make the calculations simple.

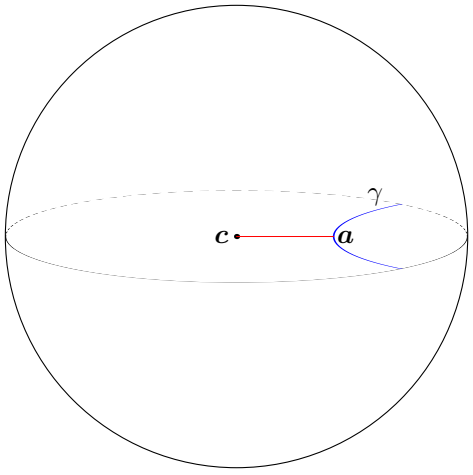
A ball $\{\boldsymbol{x} \in \mathbb{H}^3 : \rho(\boldsymbol{x}, \boldsymbol{c}) < \rho_o\}$ is a Euclidean ball which lies entirely within $\mathbb{H}^3$. Note that the Euclidean centre will not normally be $\boldsymbol{c}$ nor the Euclidean radius ρ_o.

Take $\mathbb{H}^3$ as B^3 and consider first the case where $\boldsymbol{c} = \boldsymbol{0}$. Then Proposition 14.1 shows that

$$\rho(\boldsymbol{x}, \boldsymbol{0}) = \log\left(\frac{1+||\boldsymbol{x}||}{1-||\boldsymbol{x}||}\right) \ .$$

Hence the ball is the set $\{\boldsymbol{x} : ||\boldsymbol{x}|| < \tanh\frac{1}{2}\rho_o\}$. Each Möbius transformation is a composition of inversions and these map balls to balls, so the result continues to hold for any $\boldsymbol{c} \in B^3$.

There is a unique point of a geodesic γ closest to a point $\boldsymbol{c}$ in $\mathbb{H}^3$.



Take $\mathbb{H}^3$ to be B^3 and $\boldsymbol{c} = \mathbf{0}$. The geodesic γ is then the arc of a circle orthogonal to the unit sphere ∂B^3. This circle has a unique point $\boldsymbol{a}$ with smallest norm and, by the previous remark, this is closest hyperbolically to $\mathbf{0}$. Note that the shortest hyperbolic path from $\mathbf{0}$ to $\boldsymbol{a}$ is a radial line. So we see that the hyperbolic geodesic from a point $\boldsymbol{c}$ to the closest point of a hyperbolic geodesic γ is orthogonal to γ.

For two geodesics α, β in $\mathbb{H}^3$ which do not have a common endpoint on $\partial\mathbb{H}^3$ there are unique points $\boldsymbol{a} \in \alpha$ and $\boldsymbol{b} \in \beta$ with $\rho(\boldsymbol{a}, \boldsymbol{b})$ minimal.

By applying an isometry we may assume that $\mathbb{H}^3 = \mathbb{R}^3_+$ and α is the geodesic $\{(0, 0, x_3) : x_3 > 0\}$. Consider the point $\boldsymbol{x} = (\sin\theta, 0, \cos\theta) \in \mathbb{R}^3_+$. The shortest path from $\boldsymbol{x}$ to α is the arc

$$\{(\sin\phi, 0\cos\phi) : 0 \leqslant \phi \leqslant \theta\} \ .$$

This has hyperbolic length

$$\rho_o = \int_0^\theta \frac{1}{\cos\phi}\, d\phi \ .$$

Now make the change of variables $t = \tan\frac{1}{2}\phi$, so that $\cos\phi = \frac{1-t^2}{1+t^2}$ and $\frac{dt}{d\phi} = \frac{1}{2}(1+t^2)$. This gives:

$$\rho_o = \int_0^{\tan\frac{1}{2}\theta} \frac{1+t^2}{1-t^2}\,\frac{2}{1+t^2}\, dt = \int_0^{\tan\frac{1}{2}\theta} \frac{2}{1-t^2}\, dt = \tanh^{-1}(\tan\tfrac{1}{2}\theta) \ .$$

Now the Möbius transformation $T : z \mapsto kz$ $(k > 0)$ acts on $\mathbb{R}^3_+$ as

$$T : (x_1, x_2, x_3) \mapsto (kx_1, kx_2, kx_3)$$

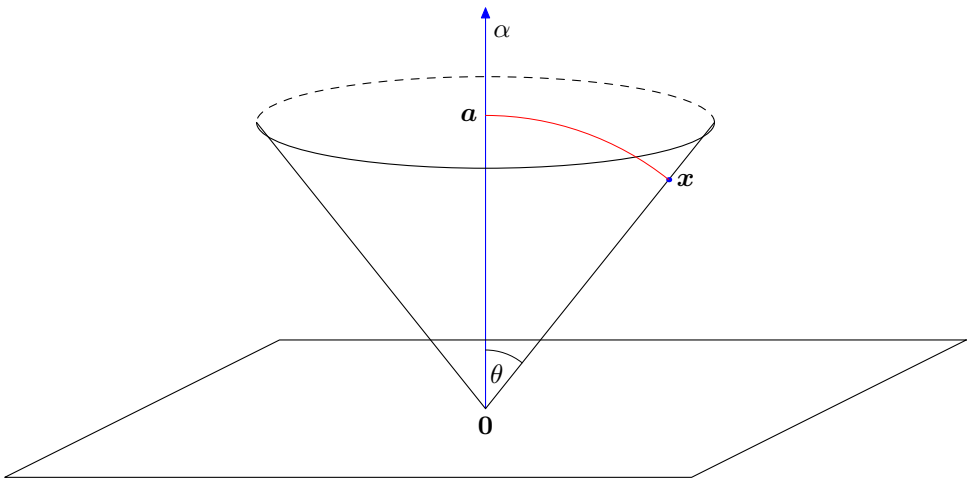
and this must be an isometry. So all of the points $(k\sin\theta, 0, k\cos\theta)$ are at a hyperbolic distance ρ_o from α. Similarly, $T : z \mapsto e^{i\psi}z$ acts on $\mathbb{R}^3_+$ as

$$T : (x_1 + ix_2, x_3) \mapsto (e^{i\psi}(x_1 + ix_2), x_3)$$

and this must be an isometry. Hence all of the points

$$\{\boldsymbol{x} \in \mathbb{R}^3_+ : x_3 = ||\boldsymbol{x}||\cos\theta\}$$

are at a hyperbolic distance $\rho_o = \tanh^{-1}\left(\tan\frac{1}{2}\theta\right)$ from α. This is a cone about the axis α.



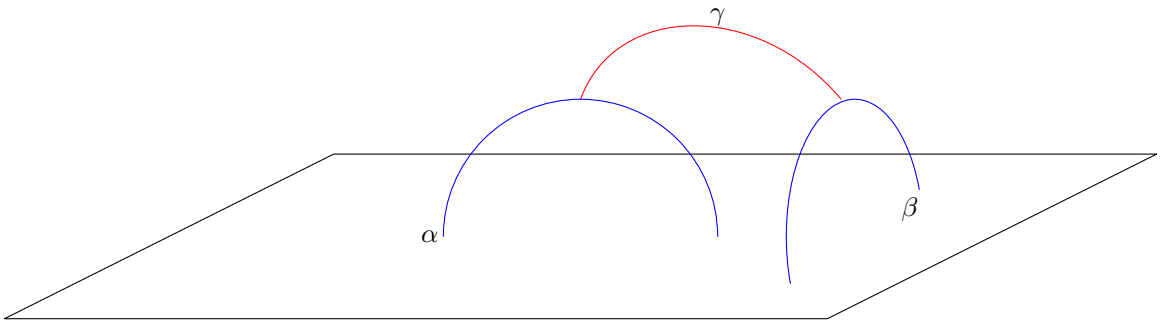
Exercise:

21. Draw the set of points that lie within a fixed hyperbolic distance ρ_o of a geodesic α in the unit disc $\mathbb{D}$ and in the unit ball B^3.

Now suppose that β is a geodesic in $\mathbb{R}^3_+$ joining points of $\partial\mathbb{R}^3_+$ neither of which is 0 or ∞. Then there will be a unique point $\boldsymbol{b} \in \beta$ with the angle

$$\sin^{-1}\left(\frac{x_3}{||\boldsymbol{x}||}\right)$$

minimal. This is the point closest to α. It is clear that the shortest path from α to β is a geodesic that crosses both α and β orthogonally.



If two geodesics have a common endpoint, then they contain points that are arbitrarily close.

Let α and β be geodesics in $\mathbb{R}^3_+$ with ∞ as their common endpoint. Then

$$\alpha = \{(a_1, a_2, x_3) : x_3 > 0\} \quad ; \qquad \beta = \{(b_1, b_2, x_3) : x_3 > 0\} \ .$$

Now it is clear that

$$\rho((a_1, a_2, x_3), (b_1, b_2, x_3)) \leqslant \frac{||(a_1, a_2) - (b_1, b_2)||}{x_3}$$

so there are points of the two geodesics that are arbitrarily close together.

Two geodesics α and β in $\mathbb{H}^3$ are either:

(a) *Parallel:* when they have a common endpoint.

(b) *Cross:* when they intersect at a point of $\mathbb{H}^3$.

(c) *Skew:* when they have no common endpoint and do not intersect.

We have shown that two skew lines have a common normal joining the closest points of the two lines. In case (b) there is also a common normal through the point of intersection. However, there is no such normal for parallel lines unless the lines are identical.

16.2 Axes of Isometries

Let T be a Möbius transformation. When we think of T acting on the Riemann sphere, it has two fixed points, say a and b. When we think of T acting on the hyperbolic 3-space $\mathbb{H}^3$, then there is a geodesic α joining a and b. T maps this geodesic to itself since it fixes the endpoints. We call α the *axis of the Möbius transformation.*

The Möbius transformation T acts isometrically on $\mathbb{H}^3$, so T must move the points of the axis a fixed hyperbolic distance. This is called the *translation length of* T. It is 0 for elliptic transformations but non-zero for hyperbolic and loxodromic transformations.

Exercise:

22. Show that the translation length of the transformation $M_k : z \mapsto kz$ is $\log |k|$. Hence show how to find the translation length of the Möbius transformation

$$z \mapsto \frac{2z+1}{5z+3} \ .$$

A parabolic transformation does not have an axis. If there were any geodesic that were mapped onto itself, preserving direction, then both endpoints would be fixed. This can not occur if there is just one fixed point on $\partial\mathbb{H}^3$.

17 INVOLUTIONS

A Möbius transformation R which has finite order must be elliptic. In particular, an *involution* which has $R^2 = I$ must be elliptic. We can conjugate R so that its fixed points are at $\mathbf{0}, \infty \in \mathbb{R}^3_+$. The axis α of R is then the positive x_3-axisjoining $\mathbf{0}$ to ∞. Then $R : z \mapsto -z$ on $\mathbb{P}$ and

$$R : (x_1, x_2, x_3) \mapsto (-x_1, -x_2, x_3) \qquad \text{for } \boldsymbol{x} \in \mathbb{R}^3_+ \ .$$

This fixes every point of the axis.

Note that, if π is a hyperbolic plane that contains the axis α of the involution R, then R maps this plane to itself but interchanges the half-spaces on either side of it. On the plane π, the involution acts as inversion in the geodesic α. In §11 we saw that every Möbius transformation of the disc could be written as the composite of two inversions. Here we will prove that each Möbius transformation on $\mathbb{P}$ can be written as the composite of two involutions.

Proposition 17.1 Isometries of $\mathbb{H}^3$ are compositions of two involutions.
Every Möbius transformation can be expressed as $R_2 \circ R_1$ for two elliptic involutions R_1, R_2.

Proof:

The identity is R^2 for every involution R.

Suppose that P is parabolic. Then we may assume that it is $P : z \mapsto z + 1$. This acts as

$$P : (x_1, x_2, x_3) \mapsto (x_1 + 1, x_2, x_3)$$

on $\mathbb{R}^3_+$. For this, take

$$R_1 : z \mapsto -z \qquad \text{and} \qquad R_2 : z \mapsto 1 - z \ .$$

Then R_1, R_2 are involutions with $P = R_2 \circ R_1$. Note that the axes of R_1, R_2 are

$$\{(0, 0, x_3) : x_3 > 0\} \qquad \text{and} \qquad \{(\tfrac{1}{2}, 0, x_3) : x_3 > 0\}$$

which have a common endpoint at the fixed point of P.

Suppose that T is a Möbius transformation with 2 fixed points. We may assume that these are 0 and ∞. So $T : z \mapsto \lambda^2 z$ for some $\lambda \neq 0$. Then T acts on $\mathbb{R}^3_+$ as

$$T : (x_1 + ix_2, x_3) \mapsto (\lambda^2(x_1 + ix_2), |\lambda^2| x_3) \ .$$

For this, take

$$R_1 : z \mapsto \frac{1}{z} \qquad \text{and} \qquad R_2 : z \mapsto \frac{\lambda^2}{z} \ .$$

These are involutions with $T = R_2 \circ R_1$. Note that the axes of R_1, R_2 are

$$\{(\cos\theta, 0, \sin\theta) : 0 < \theta < \pi\} \qquad \text{and} \qquad \{(\lambda\cos\theta, 0, |\lambda|\sin\theta) : 0 < \theta < \pi\} \ .$$

These are identical when $\lambda = \pm 1$ and $T = I$. They cross at $(0, 0, 1)$ when $|\lambda| = 1$ and T is elliptic. Otherwise, they are skew and T is loxodromic or hyperbolic with its axis normal to both. □

Proposition 17.2
Let R_1, R_2 be involutions with axes α_1, α_2 in $\mathbb{H}^3$. Then

(a) If $\alpha_1 = \alpha_2$ then $R_2 \circ R_1 = I$.

(b) If α_1 and α_2 are parallel, then $R_2 \circ R_1$ is parabolic with the common endpoint of α_1 and α_2 as its fixed point on $\partial\mathbb{H}^3$.

(c) If α_1 α_2 cross at a point P, then $R_2 \circ R_1$ is elliptic with axis through P perpendicular to α_1 and α_2.

(d) If α_1, α_2 are skew, then $R_2 \circ R_1$ is loxodromic or hyperbolic with axis normal to both α_1 and α_2.

Proof:

(a) is obvious. For (b) conjugate so that in $\mathbb{R}^3_+$ we have:

$$\alpha_1 = \{(a_1, x_3) : x_3 > 0\} \qquad \text{and} \qquad \alpha_1 = \{(a_2, x_3) : x_3 > 0\}$$

for some $a_1, a_2 \in \mathbb{C}$. Then

$$R_1 : z \mapsto 2a_1 - z \qquad \text{and} \qquad R_2 : z \mapsto 2a_2 - z$$

and $R_2 \circ R_1 : z \mapsto 2(a_2 - a_1) + z$ is parabolic with its single fixed point at ∞.

For (c) or (d), we know that there is a unique geodesic γ normal to both α_1 and α_2. Conjugate so this is the x_3-axis in $\mathbb{R}^3_+$. The geodesic α_j is then a half-circle that crosses the x_3-axis orthogonally, so it must join points $\pm w_j$. Consequently,

$$R_j : z \mapsto \frac{w_j^2}{z}$$

and $R_2 \circ R_1 : z \mapsto (w_2^2/w_1^2)z$ is either elliptic, hyperbolic or loxodromic. □

Exercise:

23. Let R_1, R_2 be involutions with axes α_1, α_2 in $\mathbb{H}^3$. Show that $R_2 \circ R_1$ is hyperbolic when both α_1 and α_2 lie in a hyperbolic plane.

Finally, let us look at the group $G = \langle T \rangle$ generated by a single orientation preserving isometry T of $\mathbb{H}^3$. The last proposition shows that we can write T as the composite $R_2 \circ R_1$ of two involutions with axes α_1, and α_2. We will concentrate on the case where T is loxodromic or hyperbolic, since the other cases are simpler. There is then a common normal γ to α_1 and α_2 and this is the axis for T. If we parametrise γ by hyperbolic length, then we can assume that $\gamma(0)$ is the point where α_1 meets γ and $\gamma(\tau)$ is the point where α_2 meets γ. This means that

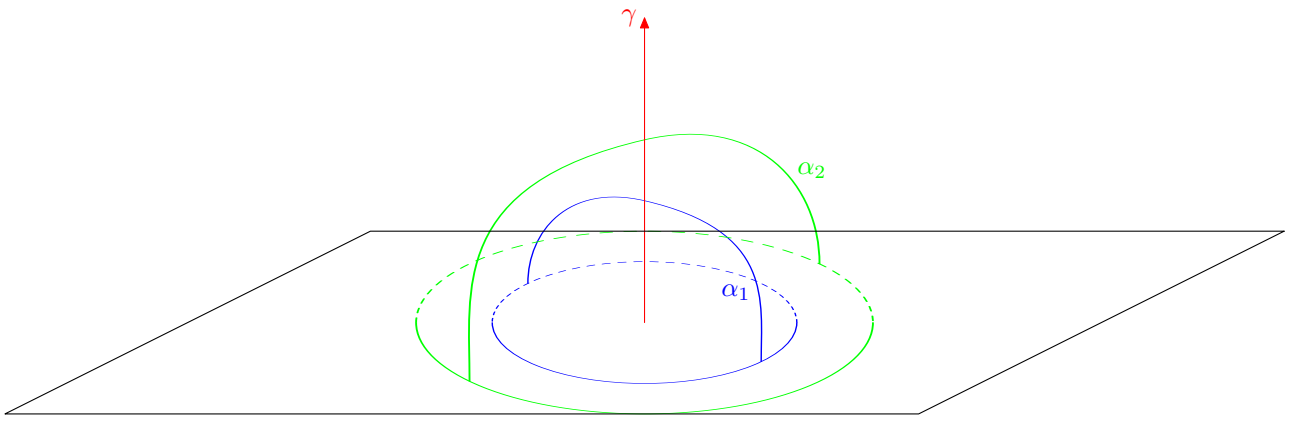
$$R_1(\gamma(t)) = \gamma(-t) \; ; \quad R_2(\gamma(t)) = \gamma(2\tau - t)) \; ; \quad T(\gamma(t)) = R_2(R_1(\gamma(t))) = \gamma(2\tau + t) \; .$$

So the translation length of T is 2τ: twice the distance between the axes α_1 and α_2.

Each point $\boldsymbol{x} \in \mathbb{H}^3$ is closest to some point of γ and those points which are closest to $\gamma(t)$ fill a hyperbolic plane $\pi(t)$. Since T is an isometry, it maps $\pi(t)$ onto $\pi(2\tau + t)$. Hence the set

$$F = \bigcup \{\pi(t) : 0 \leqslant t < 2\tau\}$$

is a fundamental set for the group $G = \langle T \rangle$. T maps the plane $\pi(0)$ that bounds one side onto the plane $\pi(2\tau)$ that bounds the other.



Note also that half of F:

$$\bigcup \{\pi(t) : 0 \leqslant t \leqslant \tau\}$$

is a fundamental set of the larger group $\langle R_1, R_2 \rangle$.

When T is parabolic, there is a fundamental set bounded by two hyperbolic planes that touch at the fixed point of T on $\partial\mathbb{H}^3$. When T is elliptic of finite order, there is a fundamental set bounded by two hyperbolic half-planes that meet on the axis of T.

Exercise:

24. Suppose that T is a Möbius transformation that maps the unit disc $\mathbb{D}$ onto itself. Then T also acts as an isometry of the hyperbolic 3-space B^3. How are fundamental sets for $G = \langle T \rangle$ acting on $\mathbb{D}$ related to fundamental sets for G acting on B^3?

---

18 KLEINIAN GROUPS

Möbius transformations are represented by 2×2 complex matrices so a group of Möbius transformations is discrete if it is a discrete subset of the set of all 2×2 matrices. A *Kleinian group* is a discrete subgroup of Möb. We will think of these groups acting as isometries of the hyperbolic 3-space $\mathbb{H}^3$.

Every Fuchsian group is certainly discrete when we think of it as a subgroup of Möb rather than a subgroup of Möb($\mathbb{D}$). We will see many more examples later.

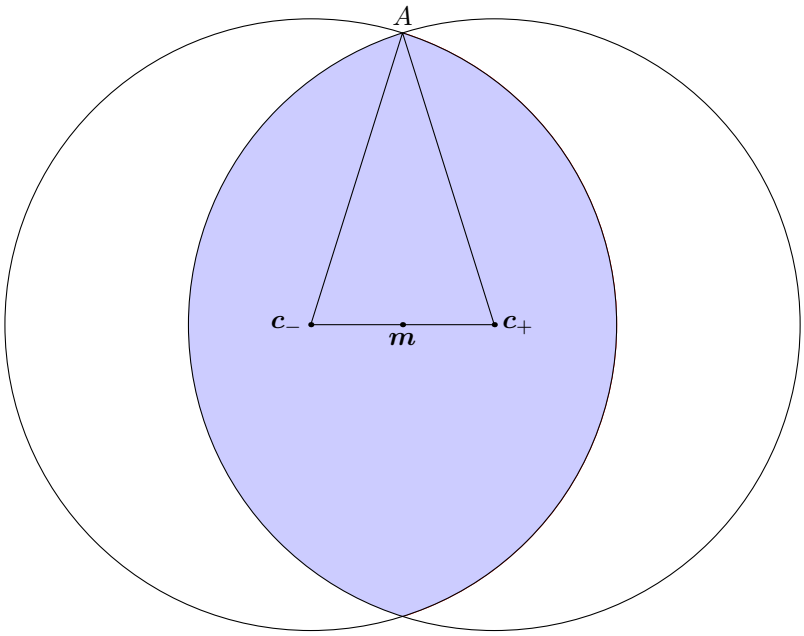
18.1 Finite Kleinian Groups

Any finite subgroup of Möb is certainly a Kleinian group. However, we will show that these finite groups are all conjugate to finite subgroups of SO(3). So we already know which groups can arise: cyclic and dihedral groups together with the tetrahedral, octahedral and icosahedral groups.

Lemma 18.1
Let S be a non-empty finite subset of $\mathbb{H}^3$. Then there is a unique closed hyperbolic ball $\overline{B}(\boldsymbol{c}, \rho)$ of smallest hyperbolic radius that contains S.

Proof:
Set ρ_o be the infimum of the radii ρ for which there is a centre $\boldsymbol{c} \in \mathbb{H}^3$ with $S \subset \overline{B}(\boldsymbol{c}, \rho)$. Then there is a sequence ρ_n that decreases to ρ_0 and a sequence of centres $\boldsymbol{c}_n$ with $S \subset \overline{B}(\boldsymbol{c}_n, \rho_n)$. The centres $\boldsymbol{c}_n$ all lie within ρ_1 of any chosen point of S, so we can find a convergent subsequence using the Bolzano – Weierstrass Theorem. We will assume that $\boldsymbol{c}_n \to \boldsymbol{c}$ as $n \to \infty$. Then $S \subset \overline{B}(\boldsymbol{c}, \rho_o)$. This shows that there is at least one closed ball of smallest radius containing S.

Now suppose that there are two different closed balls with smallest radius ρ_0 which contain S, say $\overline{B}(\boldsymbol{c}_+, \rho_o)$ and $\overline{B}(\boldsymbol{c}_-, \rho_o)$. Let $\boldsymbol{m}$ be the midpoint of the hyperbolic geodesic from $\boldsymbol{c}_+$ to $\boldsymbol{c}_-$. We can apply an isometry to $\mathbb{H}^3$ to move $\boldsymbol{m}$ to the origin. Then $\boldsymbol{c}_+ = -\boldsymbol{c}_-$. In the picture below, S must lie in the intersection of the two balls $\overline{B}(\boldsymbol{c}_+, \rho_o)$ and $\overline{B}(\boldsymbol{c}_-, \rho_o)$, so it must lie in the shaded region. This is contained in the closed ball centred on $\boldsymbol{m}$ with radius $\rho(\boldsymbol{m}, A)$. This radius is less than ρ_o, which is a contradiction. □



Proposition 18.2 Finite Kleinian groups are conjugate to subgroups of SO(3).
Every finite subgroup of Möb is conjugate in the Möbius group to a subgroup of SO(3).

Proof:

We will consider the finite group G acting on the unit ball B^3. It has a finite orbit $\Omega = G(\boldsymbol{x})$ for any point $\boldsymbol{x} \in B^3$. The lemma shows that this is contained within a unique smallest closed hyperbolic ball, say $\overline{B}(\boldsymbol{c}, \rho_o)$.

Each $T \in G$ acts isometrically on B^3 and permutes the elements of the orbit Ω. So

$$\Omega = T(\Omega) \subset T(\overline{B}(\boldsymbol{c}, \rho_o)) = \overline{B}(T(\boldsymbol{c}), \rho_o) \ .$$

This implies that $T(\boldsymbol{c}) = \boldsymbol{c}$, so every element of G fixes $\boldsymbol{c}$.

Now conjugate by a Möbius transformation that maps $\boldsymbol{c}$ to the origin. Then G becomes a group of hyperbolic isometries that fix the origin. These must be elements of SO(3). □

This proposition shows that we need to consider infinite Kleinian groups in order to obtain new and interesting examples of such groups. To do this we need to think more carefully about the action of a Kleinian group on hyperbolic 3-space.

18.2 Discontinuous Action

Let G be a subgroup of Möb $= \mathrm{Isom}^+(\mathbb{H}^3)$. The group G *acts discontinuously at* $\boldsymbol{x}_o \in \mathbb{H}^3$ if there is some $\delta > 0$ for which $\{T \in G : \rho(\boldsymbol{x}_o, T(\boldsymbol{x}_o)) < \delta\}$ is finite. The group G *acts discontinuously on* $\mathbb{H}^3$ if it acts discontinuously at each point of $\mathbb{H}^3$.

If G acts discontinuously at $\boldsymbol{x}_o$, then the stabilizer: $\mathrm{Stab}(\boldsymbol{x}_o) = \{T \in G : T(\boldsymbol{x}_o) = \boldsymbol{x}_0\}$ is a finite group and so conjugate to a finite subgroup of SO(3).

Lemma 18.3
Let G act discontinuously at a point $\boldsymbol{x}_0 \in \mathbb{H}^3$. Then, for any compact set $K \subset \mathbb{H}^3$, the set $\{T \in G : T(K) \cap K \neq \emptyset\}$ is finite.

Proof:

Since K is compact, it is certainly bounded, so there is a ρ_o with $K \subset B(\boldsymbol{x}_o, \rho_o)$. If $K \cap T(K) \neq \emptyset$, then we can find $\boldsymbol{a} \in K$ with $T(\boldsymbol{a}) \in K$. So

$$\rho(T(\boldsymbol{x}_o), \boldsymbol{x}_o) \leqslant \rho(T(\boldsymbol{x}_o), T(\boldsymbol{a})) + \rho(T(\boldsymbol{a}), \boldsymbol{x}_o) = \rho(\boldsymbol{x}_o, \boldsymbol{a}) + \rho(T(\boldsymbol{a}), \boldsymbol{x}_o) < 2\rho_o \ .$$

Suppose that there were infinitely many such elements T of G. Then the Bolzano – Weierstrass theorem shows that we can find a sequence of distinct elements (T_n) with $T_n(\boldsymbol{x}_o)$ converging to some point $\boldsymbol{y} \in \mathbb{H}^3$ as $n \to \infty$.

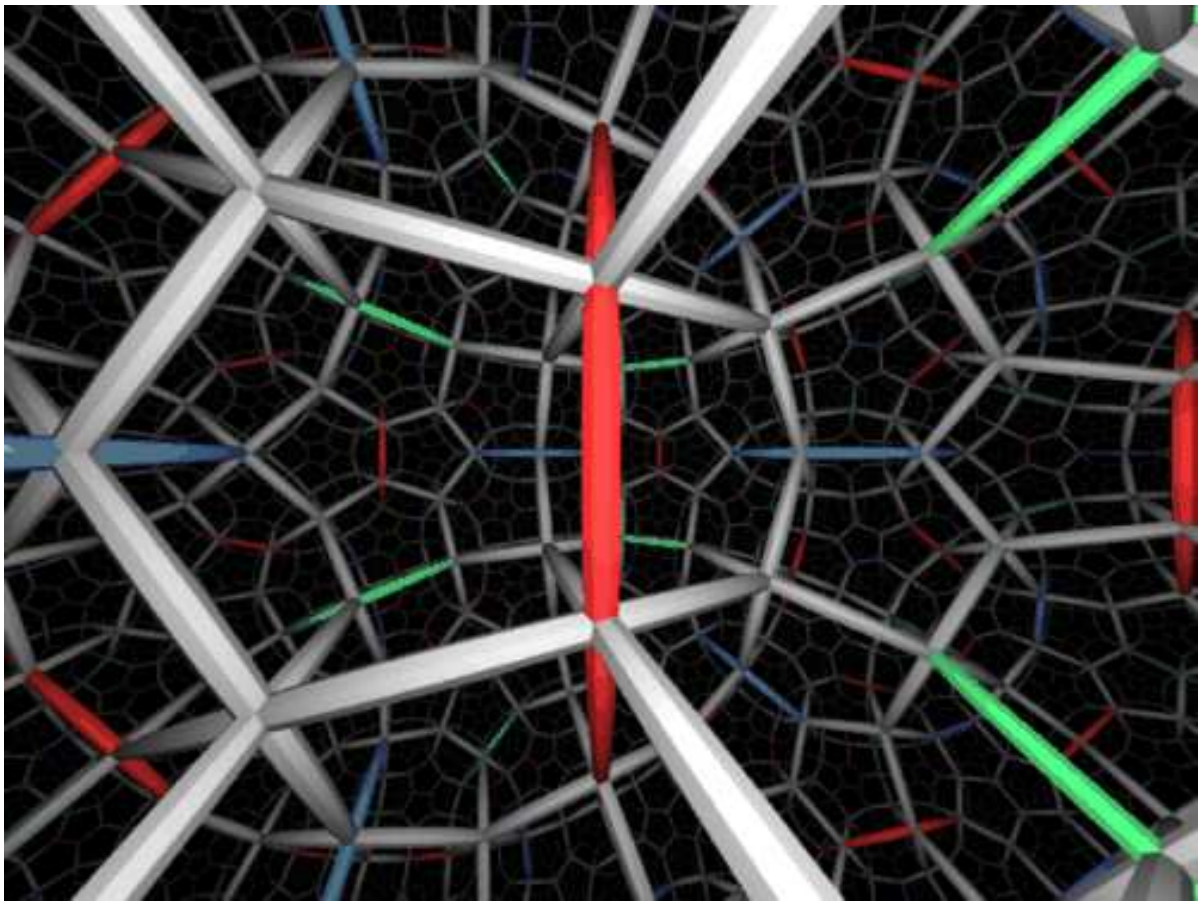
In particular, there is a natural number N with $\rho(T_n(\boldsymbol{x}_o), \boldsymbol{y}) < \frac{1}{2}\delta$ for $n \geqslant N$. Consequently the elements $S_n = T_N^{-1} \circ T_n$ of G satisfy

$$\rho(S_n(\boldsymbol{x}_o), \boldsymbol{x}_o) = \rho(T_n(\boldsymbol{x}_o), T_N(\boldsymbol{x}_o)) < \tfrac{1}{2}\delta + \tfrac{1}{2}\delta = \delta$$

for $n \geqslant N$. This means that there are infinitely many $S_n \in G$ with $\rho(S_n(\boldsymbol{x}_o), \boldsymbol{x}_o) < \delta$ which contradicts G acting discontinuously at $\boldsymbol{x}_o$. □

Suppose that F is a fundamental set for G acting on $\mathbb{H}^3$. Then the copies $T(F)$ for $T \in G$ tessellate all of $\mathbb{H}^3$. If a neighbourhood of $\boldsymbol{x}_o$ meets only finitely many copies of F, then G acts discontinuously at $\boldsymbol{x}_o$. This means it is often fairly straightforward to show that a group acts discontinuously by exhibiting a suitable fundamental set.

For example, we can construct a regular dodecahedron in $\mathbb{H}^3$ with each pentagonal face having each angle a right-angle and with the angles between faces being right-angles. The copies of this dodecahedron obtained by reflecting (inverting) in the faces tessellates all of $\mathbb{H}^3$. The group of Möbius transformations that are symmetries of this tessellation clearly acts discontinuously on $\mathbb{H}^3$.



Dodecahedral tessellation of $\mathbb{H}^3$. (See *http://www.geom.uiuc.edu/graphics/pix/Special_Topics/* .)

We will now show that acting discontinuously on $\mathbb{H}^3$ is equivalent to being a discrete group of Möbius transformations.

Theorem 18.4 Discrete if and only if acts discontinuously.
A group of Möbius transformations is discrete if and only if it acts discontinuously on $\mathbb{H}^3$.

Proof:

Suppose first that G is not discrete. Then there is a sequence (T_n) of non-identity transformations in G with $T_n \to I$ as $n \to \infty$. This implies that $T_n(\boldsymbol{x}_o) \to \boldsymbol{x}_o$ for any point $\boldsymbol{x}_o \in \mathbb{H}^3$. So G does not act discontinuously at $\boldsymbol{x}_o$.

Suppose now that G does not act discontinuously at some point $\boldsymbol{x}_o \in \mathbb{H}^3$. Then there is a $\delta > 0$ with infinitely many $T \in G$ satisfying $\rho(\boldsymbol{x}_o, T(\boldsymbol{x}_o)) < \delta$. It is more convenient to work on the upper half-space, so conjugate so that $\boldsymbol{x}_o$ is mapped to $\boldsymbol{k} = (0, 0, 1) \in \mathbb{R}^3_+$.

For each $\boldsymbol{a} \in \mathbb{R}^3_+$ define $S_{\boldsymbol{a}} : \boldsymbol{x} \mapsto a_3\boldsymbol{x} + (a_1, a_2, 0)$. Then $S_{\boldsymbol{a}}(\boldsymbol{k}) = \boldsymbol{a}$. (This corresponds to the Möbius transformation on $\mathbb{P}$ given by $z \mapsto a_3 z + (a_1 + ia_2)$.) Use the Bolzano – Weierstrass theorem to find a sequence of distinct transformations $T_n \in G$ with $T_n(\boldsymbol{k}) \to \boldsymbol{y}$ as $n \to \infty$. Set

$$S_n = S_{T_n(\boldsymbol{k})} \quad \text{and} \quad R_n = S_n^{-1} \circ T_n \ .$$

(Note that S_n and R_n need not be in G.) Then we have $R_n(\boldsymbol{k}) = \boldsymbol{k}$ so each R_n is in the group $\mathrm{Stab}(\boldsymbol{k})$. This stabilizer is conjugate to the group SO(3), which is compact. Hence we can find a subsequence $R_{n'}$ which converges to a transformation R_∞ as $n' \to \infty$.

Now we have $T_n(\boldsymbol{k}) \to \boldsymbol{y}$, so $S_n \to S_{\boldsymbol{y}}$ as $n \to \infty$. Hence, on the subsequence, we have

$$T_{n'} = S_{n'} \circ R_{n'} \to S_{\boldsymbol{y}} \circ R_\infty \qquad \text{as } n \to \infty \ .$$

This shows that G is not discrete. □

We can also consider groups acting discontinuously on other spaces and prove similar results.

Look at: *http://www.plunk.org/∼hatch/HyperbolicApplet/* .

19 LIMITS OF ORBITS

Let G be a Kleinian group acting on the ball B^3 and choose a point $\boldsymbol{x}_o \in B^3$. Then we have shown that G acts discontinuously on $\mathbb{H}^3 = B^3$. So only finitely many points of the orbit $G(\boldsymbol{x}_o)$ lie within any hyperbolic ball. This implies that $||T(\boldsymbol{x}_o)||$ tends to 1 as T runs through G. We can think of the points $G(\boldsymbol{x}_o)$ as lying inside the space $\mathbb{R}^3$ with the Euclidean metric (or, better still, inside $\mathbb{R}^3_\infty$ with the chordal metric) and consider the limit points of the orbit.

Recall that a *limit point* of a set $\Omega \subset \mathbb{R}^3$ is a point $\boldsymbol{u}$ for which there is a sequence $(\boldsymbol{w}_n)$ of distinct points $\boldsymbol{w}_n \in \Omega$ which converge to $\boldsymbol{u}$, so $||\boldsymbol{w}_n - \boldsymbol{u}|| \to 0$ as $n \to \infty$. The set of limit points of the orbit $G(\boldsymbol{x}_o)$ is called the *limit set* of G and will be denoted by $\Lambda(\boldsymbol{x}_o)$. We will see shortly that this limit set is independent of the point $\boldsymbol{x}_o$ we choose.

Note that when we talk about limit points of the orbit we are using the Euclidean (or, chordal) metric. The hyperbolic metric is not defined on the boundary so it does not make sense to ask for limit points in the hyperbolic metric.

Proposition 19.1 The limit set is closed and G-invariant.
For any Kleinian group acting on B^3 the limit set $\Lambda(\boldsymbol{x}_o)$ is a closed, G-invariant subset of $\partial\mathbb{H}^3 = S^2$.

Proof:
$\Lambda(\boldsymbol{x}_o)$ is obviously closed since it is $\overline{G(\boldsymbol{x}_o)} \setminus B^3$. Since the orbit $\Omega = G(\boldsymbol{x}_o)$ satisfies $T(\Omega) = \Omega$ for each $T \in G$, and T is continuous on all of $\mathbb{R}^3_\infty$, we see that $T(\Lambda(\boldsymbol{x}_o)) = \Lambda(\boldsymbol{x}_o)$. So the limit set is G-invariant. □

Suppose that T is a loxodromic or hyperbolic transformation in the group G. Then $T^n(\boldsymbol{x}_o)$ tends to one of the fixed points of T as $n \to \infty$ and $T^{-n}(\boldsymbol{x}_o)$ tends to the other. Hence, both the fixed points of T are in the limit set for G. Similarly the fixed point of a parabolic transformation is in the limit set. However the fixed points of an elliptic transformation need not lie in the limit set.

Exercise:

25. Give an example of an elliptic element of a Kleinian group with fixed points that do not lie in the limit set. Give an example of a Kleinian group for which the limit set is empty.
26. Let G be a Kleinian group with an invariant disc $\Delta \subset \mathbb{P}$. Show that the limit set of G is a subset of $\partial\Delta$.

We will now show that $\Lambda(\boldsymbol{x}_o)$ is independent of the point $\boldsymbol{x}_o$. For suppose that $\boldsymbol{x}_1$ is another point of B^3. Then $\rho(\boldsymbol{x}_o, \boldsymbol{x}_1)$ is finite. Each $T \in G$ is an isometry for the hyperbolic metric, so $\rho(T(\boldsymbol{x}_o), T(\boldsymbol{x}_1)) = \rho(\boldsymbol{x}_o, \boldsymbol{x}_1)$. Although these points $T_n(\boldsymbol{x}_o)$, and $T_n(\boldsymbol{x}_1)$ stay the same hyperbolic distance apart they get closer together for the Euclidean metric as the points get closer to the boundary. Hence $T_n(\boldsymbol{x}_o)$ and $T_n(\boldsymbol{x}_1)$ will converge to the same point of ∂B^3.

Lemma 19.2
Let $(\boldsymbol{x}_n)$ and $(\boldsymbol{y}_n)$ be two sequences of points in B^3 with $\rho(\boldsymbol{x}_n, \boldsymbol{y}_n) \leqslant K$ for each $n \in \mathbb{N}$. If the sequence of points $(\boldsymbol{x}_n)$ in B^3 converges in the Euclidean metric to a limit point $\boldsymbol{u} \in \partial B^3$, then the sequence $(\boldsymbol{y}_n)$ will also converge to $\boldsymbol{u}$ for the Euclidean metric.

Proof:
The hyperbolic density at a point $\boldsymbol{x}$ is

$$\lambda(\boldsymbol{x}) = \frac{2}{1 - ||\boldsymbol{x}||^2} = 2\cosh^2 \tfrac{1}{2}\rho(\boldsymbol{0}, \boldsymbol{x}) \ .$$

So, for any given $\varepsilon > 0$, we can find ρ_o with

$$\lambda(\boldsymbol{x}) > \frac{1}{\varepsilon} \qquad \text{for } \rho(\boldsymbol{0}, \boldsymbol{x}) > \rho_o - K \ .$$

Suppose that γ is a hyperbolic geodesic joining the points $\boldsymbol{a}, \boldsymbol{b}$. Then

$$\rho(\boldsymbol{a}, \boldsymbol{b}) = L(\gamma) = \int_\gamma \lambda(\boldsymbol{x}) \, ||d\boldsymbol{x}|| \geqslant \inf \{\lambda(\boldsymbol{x}) : \boldsymbol{x} \in \gamma\} \, ||\boldsymbol{b} - \boldsymbol{a}|| \ .$$

Hence, if $\rho(\boldsymbol{a}, \boldsymbol{b}) \leqslant K$ and $\rho(\boldsymbol{0}, \boldsymbol{a}) > \rho_o$, then

$$K \geqslant \rho(\boldsymbol{a}, \boldsymbol{b}) \geqslant \frac{1}{\varepsilon} ||\boldsymbol{b} - \boldsymbol{a}|| \ .$$

Applying this to the pairs of points $\boldsymbol{x}_n, \boldsymbol{y}_n$ gives the result. □

Proposition 19.3 Limit set is independent of the base point.
The limit sets $\Lambda(\boldsymbol{x}_o)$ and $\Lambda(\boldsymbol{x}_1)$ are equal for any Kleinian group G and any points $\boldsymbol{x}_o, \boldsymbol{x}_1 \in B^3$.

Proof:

A point $\boldsymbol{u}$ is in $\Lambda(\boldsymbol{x}_o)$ if there is a sequence (T_n) in G with $||T_n(\boldsymbol{x}_o) - \boldsymbol{u}|| \to 0$ as $N \to \infty$. Now each T_n acts isometrically on B^3, so

$$\rho(T_n(\boldsymbol{x}_o), T_n(\boldsymbol{x}_1)) = \rho(\boldsymbol{x}_o, \boldsymbol{x}_1) \ .$$

The lemma shows that $||T_n(\boldsymbol{x}_1) - \boldsymbol{u}|| \to 0$ as $N \to \infty$. So $\boldsymbol{u} \in \Lambda(\boldsymbol{x}_1)$. □

We already know that every fixed point of a hyperbolic or loxodromic transformation does lie in the limit set. For almost all Kleinian groups the fixed points of loxodromic and hyperbolic transformations are dense in the limit set. The exceptional groups are ones with very simple structure. They are called *elementary groups* and we will not be concerned with them.

Exercise:

27. Let G be the group generated by the single parabolic transformation $P : z \mapsto z + 1$. Show that the limit set is $\{\infty\}$ but that there are no hyperbolic or loxodromic transformations in G.

Lemma 19.4
Let α be a hyperbolic geodesic that passes through a point $\boldsymbol{c} \in B^3$. At least one of the endpoints $\boldsymbol{u}$ of α satisfies

$$||\boldsymbol{u} - \boldsymbol{c}|| \leqslant \frac{1 - ||\boldsymbol{c}||^2}{\sqrt{2}||\boldsymbol{c}||} = \frac{\sqrt{2}}{\sinh \rho(\boldsymbol{0}, \boldsymbol{c})} \ .$$

Proof:

Draw the Euclidean sphere π that passes through $\boldsymbol{c}$, is orthogonal to ∂B^3 and crosses the radius from $\boldsymbol{0}$ to $\boldsymbol{c}$ normally. This Euclidean sphere has radius r and centre $\boldsymbol{k}$. It meets B^3 in a hyperbolic plane.

Since π cuts the unit sphere orthogonally, we must have

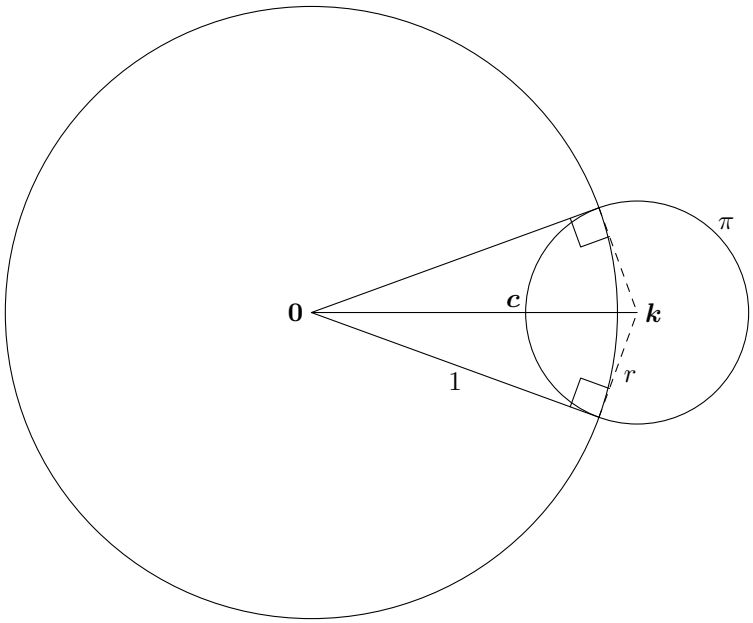
$$||\boldsymbol{k}||^2 = (||\boldsymbol{c}|| + r)^2 = 1^2 + r^2 \ .$$

So

$$r = \frac{1 - ||\boldsymbol{c}||^2}{2||\boldsymbol{c}||} \ .$$

Any hyperbolic geodesic through $\boldsymbol{c}$ goes inside π in one direction, so one endpoint will lie within the sphere π. This endpoint $\boldsymbol{u}$ satisfies

$$||\boldsymbol{u} - \boldsymbol{c}|| \leqslant \sqrt{2} r \ .$$



□

Proposition 19.5
Let G be a Kleinian group that contains a hyperbolic or loxodromic transformation. Then the fixed points of all the hyperbolic and loxodromic transformations in G are dense in the limit set for G.

Proof:

Let A be the hyperbolic or loxodromic transformation in G. Let α be its axis joining the two fixed points $\boldsymbol{a}_+, \boldsymbol{a}_-$. We know that both these fixed points lie in the limit set. Choose a point $\boldsymbol{x}_o \in B^3$ that lies on the axis α. Then Proposition 19.3 shows that it suffices to prove that the fixed points are dense in $\Lambda(\boldsymbol{x}_o)$.

Let $\boldsymbol{u} \in \Lambda(\boldsymbol{x}_o)$. Then there is a sequence (T_n) in G with $||T_n(\boldsymbol{x}_o) - \boldsymbol{u}|| \to 0$ as $n \to \infty$. Now the conjugate $T_n \circ A \circ T_n^{-1}$ is hyperbolic or loxodromic and has axis $T_n(\alpha)$ which passes through $T_n(\boldsymbol{x}_o)$. The lemma shows that one of the endpoints $T_n(\boldsymbol{a}_+)$ or $T_n(\boldsymbol{a}_-)$ satisfies

$$||T_n(\boldsymbol{a}_\pm) - \boldsymbol{u}|| \leqslant \frac{\sqrt{2}}{\sinh \rho(\boldsymbol{0}, T_n(\boldsymbol{x}_o))} \ .$$

As $n \to \infty$, so $\rho(\boldsymbol{0}, T_n(\boldsymbol{x}_o)) \to \infty$. Hence, some sequence of endpoints $T_n(\boldsymbol{a}_\pm)$ converges to $\boldsymbol{u}$. □

Recall that a subset Q of a metric space is *perfect* if no point of Q is isolated.

Corollary 19.6 Limit sets are perfect.
The limit set of a Kleinian group G that contains a hyperbolic or loxodromic transformation is either finite or perfect.

Proof:

Suppose that the limit set is not finite. Let A be a hyperbolic or loxodromic transformation in G with axis α. For each point $\boldsymbol{v} \in \partial B^3 = S^2$, except the endpoints of α, we know that $A^n(\boldsymbol{v})$ tends to one endpoint $\boldsymbol{u}_+$ of the axis α as $n \to +\infty$ and to the other $\boldsymbol{u}_-$ as $n \to -\infty$.

Since the limit set is infinite, Lemma 19.4 shows that there is another fixed point $\boldsymbol{v}$ of a hyperbolic or loxodromic transformation that is not fixed by A. Then $A^n(\boldsymbol{v}) \to \boldsymbol{u}_\pm$ as $n \to \pm\infty$. Hence both the fixed points $\boldsymbol{u}_\pm$ of A are not isolated in the limit set.

Every point of the limit set is a limit of fixed points of hyperbolic or loxodromic transformations in G, so no point in the limit set is isolated. □

Corollary 19.7 Limit sets are finite or uncountable.
The limit set of a Kleinian group G that contains a hyperbolic or loxodromic transformation is either finite or uncountable.

Proof:
For the limit set Λ is a closed subset of the sphere S^2, so it is compact. The previous proposition shows that it is perfect. Now the result follows from Cantor's theorem that a perfect, compact, metric space is uncountable.

Suppose that Λ were countable and enumerate its points as $(x_n)_{n\in\mathbb{N}}$. We will construct a decreasing sequence of non-empty, perfect, closed subsets K_n of Λ with $x_n \notin K_n$. Then the intersection $\bigcap K_n$ can not be empty, since Λ is compact. However, it does not contain any of the points x_n so it must be empty.

Set $K_0 = \Lambda$. Suppose that K_n has been defined and is a non-empty, perfect, closed subset of Λ. If $x_{n+1} \notin K_n$ then take $K_{n+1} = K_n$. Otherwise, $x_{n+1} \in K_n$. Since K_n is perfect, there must be another point, say y, in K_n. Set K_{n+1} to be the closure of

$$K_n \cap B(y, \tfrac{1}{2}\rho(x_{n+1}, y)) \ .$$

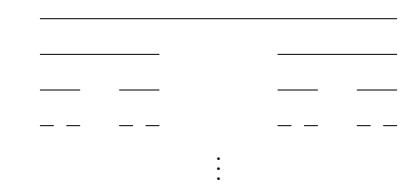
No point of this closure is isolated, so K_{n+1} is a non-empty, perfect, closed set with $x_{n+1} \notin K_{n+1} \subset K_n$. This completes the inductive construction. □

20 HAUSDORFF DIMENSION

20.1 Cantor Sets

Define a sequence of sets $C_n \subset [0,1]$ as follows:

$$C_0 = [0,1]$$
$$C_1 = [0,\tfrac{1}{3}] \cup [\tfrac{2}{3},1]$$
$$C_2 = [0,\tfrac{1}{9}] \cup [\tfrac{2}{9},\tfrac{1}{3}] \cup [\tfrac{2}{3},\tfrac{7}{9}] \cup [\tfrac{8}{9},1]$$
$$C_3 = [0,\tfrac{1}{27}] \cup [\tfrac{2}{27},\tfrac{1}{9}] \cup [\tfrac{2}{9},\tfrac{7}{27}] \cup [\tfrac{8}{27},\tfrac{1}{3}] \cup [\tfrac{2}{3},\tfrac{19}{27}] \cup [\tfrac{20}{27},\tfrac{7}{9}] \cup [\tfrac{8}{9},\tfrac{25}{27}] \cup [\tfrac{26}{27},1]$$
$$\vdots$$



At each stage, C_n consists of 2^n disjoint intervals each of length 3^{-n}. C_{n+1} is obtained from C_n by removing the open middle third from each of these intervals. The *Cantor set* C is the intersection $C = \bigcap_{n\in\mathbb{N}} C_n$.

The Cantor set is clearly closed and bounded in $\mathbb{R}$, so it is compact. It is non-empty, since the endpoints of each component interval of C_n lie in C. More carefully, we can write any $x \in [0,1]$ in base 3 as

$$x = (0.x_1x_2x_3\ldots)_3 \quad = \quad \frac{x_1}{3} + \frac{x_2}{3^2} + \frac{x_3}{3^3} + \ldots$$

where the digits $x_1, x_2, x_3, \ldots$ are each 0 or 1 or 2. Then $x \in C_n$ if and only if it can be written as

$$x = \frac{x_1}{3} + \frac{x_2}{3^2} + \frac{x_3}{3^3} + \ldots \qquad \text{with } x_1, x_2, \ldots, x_n \in \{0,2\}\ .$$

Note that this is true even for the endpoints of C_n . For example,

$$\tfrac{1}{3} = (0.022222\ldots)_3 \qquad \text{and} \qquad \tfrac{2}{3} = (0.200000\ldots)_3\ .$$

Consequently, the Cantor set consists of all $x \in [0,1]$ that can be written as

$$x = \sum_{k=1}^{\infty} \frac{x_k}{3^k} \qquad \text{with each } \ x_k \in \{0,2\}\ .$$

This shows that the Cantor set has the same cardinality as the unit interval, for the map

$$C \to [0,1]\ ; \ \ \sum \frac{x_k}{3^k} \mapsto \sum \frac{x_k/2}{2^k}$$

is a bijection. Hence the Cantor set is uncountable. It also shows that the map

$$\phi : \mathbb{Z}_2^{\mathbb{N}} \to C\ ; \ \ (y_k) \mapsto \sum \frac{2y_k}{3^k}$$

is a bijection, so we can identify C with the infinite group $\mathbb{Z}_2^{\mathbb{N}}$.

Exercise:

28. Let $\boldsymbol{y} = (y_k)$ and $\boldsymbol{z} = (z_k)$ be sequences in $\mathbb{Z}_2^{\mathbb{N}}$. Show that

$$d(\boldsymbol{y},\boldsymbol{z}) = \begin{cases} 0 & \text{when } \boldsymbol{y} = \boldsymbol{z}; \\ 2^{-n} & \text{when } n = \min\{k : y_k \neq z_k\} \end{cases}$$

is a metric on $\mathbb{Z}_2^{\mathbb{N}}$. Show that the map $\phi : \mathbb{Z}_2^{\mathbb{N}} \to C$ defined above is a homeomorphism from $\mathbb{Z}_2^{\mathbb{N}}$ with this metric.

The Cantor set is easily seen to be perfect. This gives another proof that the set C is uncountable.

The sets C_n for $N \geqslant 1$ are disconnected, so the Cantor set is disconnected. Indeed, the Cantor set is *totally disconnected*: the only connected subsets of C are the singletons. For suppose that x, y were two different points of C. Then $|x - y| > 3^{-n}$ for some $n \in \mathbb{N}$, so x and y are in different components of C_n. Any non-empty, compact, metric space which is both perfect and totally disconnected is called *a Cantor set*. All such sets are, in fact, homeomorphic to the Cantor set (but we will not prove this).

The most important property of the Cantor set is its self-similarity. The maps

$$s_0 : C \to C \; ; x \mapsto \tfrac{1}{3}x \quad \text{and} \quad s_1 : C \to C \; ; x \mapsto \tfrac{1}{3}x + \tfrac{2}{3}$$

send C homeomorphically onto the subsets $C \cap [0, \frac{1}{3}]$ and $C \cap [\frac{2}{3}, 1]$. Each of them is an contraction with scale factor $\frac{1}{3}$.

Exercise:

29. How do these self-similarities act on $\mathbb{Z}_2^{\mathbb{N}}$?

We can use these self-similarities to find the "dimension" of the Cantor set. For now we will do this informally. Later we will define the Hausdorff dimension of a set and prove the results properly. Consider a subset X of $\mathbb{R}^N$ and let $\mathbb{V}_d(X)$ denote the d-dimensional volume of X. So $\mathbb{V}_1(X)$ is the length of X; $\mathbb{V}_2(X)$ the area of X; $\mathbb{V}_3(X)$ the volume of X. If X is a set of dimension $k = 1, 2, 3, \ldots$, then we expect that

$$\mathbb{V}_d(X) = \begin{cases} 0 & \text{for } k < d; \\ \infty & \text{for } k > d. \end{cases}$$

If s is a contraction with scale factor λ then

$$\mathbb{V}_d(s(X)) = \lambda^d \mathbb{V}_d(X) \; .$$

For the Cantor set C we know that C is the disjoint union of $s_0(C)$ and $s_1(C)$. Hence, we would expect

$$\begin{aligned} \mathbb{V}_d(C) &= \mathbb{V}_d(s_0(C)) + \mathbb{V}_d(s_1(C)) \\ &= \left(\tfrac{1}{3}\right)^d \mathbb{V}_d(C) + \left(\tfrac{1}{3}\right)^d \mathbb{V}_d(C) = 2\left(\tfrac{1}{3}\right)^d \mathbb{V}_d(C) \; . \end{aligned}$$

So the d-dimensional measure can only be finite and non-zero when

$$1 = 2\left(\tfrac{1}{3}\right)^d \; , \text{ that is } \; d = \frac{\log 2}{\log 3} = 0.63093\ldots .$$

Hence the Cantor set has fractional dimension $0.63093\ldots$.

20.2 Hausdorff Dimension

Let M be a metric space with metric d. We will be particularly interested in subsets of $\mathbb{R}^N$ with the Euclidean metric or $\mathbb{R}^N_\infty$ with the chordal metric, for example the Cantor set or limit sets of Kleinian groups. The *diameter of* M is

$$\operatorname{diam}(M) = \sup\{d(x, y) : x, y \in M\} \; .$$

A collection $\{U_1, U_2, U_3, \ldots\}$ of subsets of M is a *δ-cover for M* when $M = \bigcup_{n\in\mathbb{N}} U_n$ and each set U_n has diameter at most δ. We set

$$\mathcal{H}^d_\delta(M) = \inf\left\{ \sum_{n\in\mathbb{N}} \operatorname{diam}(U_n)^d : (U_n)_{n\in\mathbb{N}} \text{ is a } \delta\text{-cover for } M \right\} \; .$$

As we decrease δ, so the class of allowed δ-covers is reduced. Hence, $\mathcal{H}^d_s(M)$ increases as $\delta \searrow 0$. Now we define the *d-dimensional Hausdorff measure of M* as:

$$\mathcal{H}^d(M) = \lim_{\delta \to 0} \mathcal{H}^d_\delta(M) = \sup\left\{\mathcal{H}^d_\delta(M) : \delta > 0\right\} \ .$$

This measures the d-dimensional size of the set M. It is usually either 0 or ∞. Indeed, we will show that, for each set M, there is at most one dimension d for which $\mathcal{H}^d(M)$ is finite and non-zero.

Suppose that M is a metric space and $0 \leqslant s < d$. For each $0 < \delta < 1$ and any δ-cover (U_n) of M we have

$$\sum \operatorname{diam}(U_n)^d \leqslant \delta^{d-s} \sum \operatorname{diam}(U_n)^s \ .$$

So $\mathcal{H}^d_\delta(M) \leqslant \delta^{d-s}\mathcal{H}^s_\delta(M)$. This shows that

$$\begin{array}{lll} \text{if } \mathcal{H}^s(M) < \infty & \text{then} & \mathcal{H}^d(M) = 0 \quad \text{for } s < d \ ; \\ \text{if } \mathcal{H}^d(M) > 0 & \text{then} & \mathcal{H}^s(M) = \infty \quad \text{for } s < d \ . \end{array}$$

So there is a critical value at which the Hausdorff dimension $\mathcal{H}^s(M)$ jumps from ∞ to 0. This value is called the *Hausdorff dimension* $\dim_{\mathcal{H}} M$ *of* M. Note that

$$\mathcal{H}^s(M) = \begin{cases} \infty & \text{for } s < \dim_{\mathcal{H}} M; \\ 0 & \text{for } s > \dim_{\mathcal{H}} M. \end{cases}$$

but the Hausdorff measure at the critical value $\dim_{\mathcal{H}} M$ may be any number between 0 and ∞ including both endpoints. It is usually very hard to calculate the Hausdorff measure at this critical value and only a little easier to find the Hausdorff dimension.

A map $f : M \to N$ between two metric spaces is *K-Lipschitz* if

$$d(f(x), f(y)) \leqslant K d(x, y) \qquad \text{for all } x, y \in M \ .$$

Such a map is certainly uniformly continuous. A map $f : M \to N$ is *Lipschitz* if it is K-Lipschitz for some finite K. A map $f : M \to N$ is *bi-Lipschitz* if it is Lipschitz and it has an inverse which is also Lipschitz. Any map $f : \mathbb{R}^N \to \mathbb{R}^N$ that is differentiable with bounded derivative is certainly Lipschitz by the mean value theorem. So, for example, every Möbius transformation is bi-Lipschitz.

Proposition 20.1
If $f : M \to N$ is a K-Lipschitz map, then

$$\mathcal{H}^d(f(M)) \leqslant K^d \mathcal{H}^d(M) \ .$$

Proof:

Suppose that (U_n) is a δ-cover for M. Then each $f(U_n)$ has diameter at most $K\operatorname{diam}(U_n)$, so $(f(U_n))$ is a $K\delta$-cover for $f(M)$. Furthermore,

$$\sum \operatorname{diam}(f(U_n))^d \leqslant K^d \sum \operatorname{diam}(U_n)^d \ .$$

So, $\mathcal{H}^d_{K\delta}(f(M)) \leqslant K^d \mathcal{H}^d_\delta(M)$. Taking the limits as $\delta \searrow 0$ gives the result. □

Corollary 20.2 Lipschitz maps preserve Hausdorff dimension
If $f : M \to N$ is a Lipschitz map, then

$$\dim_{\mathcal{H}} f(M) \leqslant \dim_{\mathcal{H}} M \ .$$

Proof:

Suppose that f is K-Lipschitz. The Proposition shows that $\mathcal{H}^d(f(M)) \leqslant K^d \mathcal{H}^d(M)$ for every value of the dimension d. Hence $\dim_{\mathcal{H}} f(M) \leqslant \dim_{\mathcal{H}} M$. □

This Corollary shows that the Hausdorff dimension of a set is unchanged by bi-Lipschitz maps. In particular, the Hausdorff dimension of the limit set $\Lambda(G)$ of a Kleinian group G is unchanged by a Möbius transformation. So $\Lambda(TGT^{-1})$ and $\Lambda(G)$ have the same Hausdorff dimension.

Exercise:

30. A map $f : M \to N$ is α-Hölder continuous if there is a constant $C < \infty$ with

$$d(f(x), f(y)) \leqslant C d(x, y)^\alpha \qquad \text{for all } \ x, y \in M \ .$$

Show that, for such a map,

$$\dim_{\mathcal{H}} f(M) \leqslant \frac{1}{\alpha} \dim_{\mathcal{H}} M \ .$$

The Hausdorff measure $\mathcal{H}^d$ is an (outer) measure. We will not use measure theory but we will need the following very simple consequences:

If $M \subset N$ then $\mathcal{H}^d(M) \leqslant \mathcal{H}^d(N)$.

If M is the disjoint union of two subsets M_1, M_2, then $\mathcal{H}^d(M) = \mathcal{H}^d(M_1) + \mathcal{H}^d(M_2)$.

We are now in a position to prove properly that the Cantor set C has Hausdorff dimension $\log 2 / \log 3$. Let $C_0 = C \cap [0, \frac{1}{3}]$ and $C_1 = C \cap [\frac{2}{3}, 1]$. The contraction maps

$$s_0 : C \to C_0 \ ; x \mapsto \tfrac{1}{3}x \ \text{ and } \ s_1 : C \to C_1 \ ; x \mapsto \tfrac{1}{3}x + \tfrac{2}{3}$$

are each Lipschitz maps with Lipschitz constant $\frac{1}{3}$. Their inverses are also Lipschitz with constants 3. So

$$\mathcal{H}^d(C_0) = \mathcal{H}^d(C_1) = \left(\tfrac{1}{3}\right)^d \mathcal{H}^d(C) \ .$$

This shows that

$$\mathcal{H}^d(C) = \mathcal{H}^d(C_0) + \mathcal{H}^d(C_1) = 2\left(\tfrac{1}{3}\right)^d \mathcal{H}^d(C)$$

for every dimension d. If, for some value d, we have $0 < \mathcal{H}^d(C) < \infty$ then we see that $2\left(\frac{1}{3}\right)^d = 1$ and so the Hausdorff dimension d must satisfy

$$2\left(\tfrac{1}{3}\right)^d = 1 \qquad \text{that is} \qquad d = \frac{\log 2}{\log 3} \ .$$

However, we need to show that for this dimension we do indeed have $0 < \mathcal{H}^d(C) < \infty$.

First observe that the set C_k consists of 2^k intervals each of length 3^{-k}. These intervals form a δ-cover of C for $\delta = 3^{-k}$. So

$$\mathcal{H}^d_\delta(C) \leqslant 2^k (3^{-k})^d = (2 \times 3^{-d})^k = 1$$

because $d = \log 3 / \log 2$. Letting $k \nearrow \infty$ this gives $\mathcal{H}^d(C) \leqslant 1$.

For the converse, suppose that (U_n) is a δ-cover of C. We will show that

$$\sum \operatorname{diam}(U_n)^d \geqslant \tfrac{1}{2} \ .$$

This will prove that $\mathcal{H}^d(C) \geqslant \frac{1}{2}$ as required.

First observe that, by expanding the sets U_n by a small amount, we can ensure that they are open. Since C is compact, there is then a finite subcover. So we can assume that (U_n) is a finite collection of open sets that cover C. Choose an integer K so that each U_n has diameter greater than 3^{-K}. Suppose that U_n is one of these sets, with

$$3^{-k-1} \leqslant \operatorname{diam}(U_n) < 3^{-k}$$

for some natural number $k < K$. Now any two components of C_k are distance at least 3^{-k} apart, so U_n can not meet more than one of them. Let V_n be this component. It has length 3^{-k}, so $\operatorname{diam}(U_n) \geqslant \frac{1}{3}\operatorname{diam}(V_n)$. Then (V_n) is a cover for C and

$$\sum \operatorname{diam}(U_n)^d \geqslant \sum \left(\tfrac{1}{3}\operatorname{diam}(V_n)\right)^d = \frac{1}{3^d}\sum \operatorname{diam}(V_n)^d = \frac{1}{2}\sum \operatorname{diam}(V_n)^d \ .$$

Hence it will suffice to prove that

$$\sum \operatorname{diam}(V_n)^d \geqslant 1$$

for a finite cover of C by the intervals V_n.

Now consider one of the intervals V_n. The intersection $C_K \cap V_n$ consists of 2^{K-k} intervals each of length 3^{-K}, which we will denote by $(W_j)_{j=1}^{2^{K-k}}$. Then

$$\operatorname{diam}(V_n)^d = (3^{-k})^d = 2^{-k} \qquad \text{and} \qquad \sum_{j=1}^{2^{K-k}} \operatorname{diam}(W_j)^d = 2^{K-k}(3^{-K})^d = 2^{-k}(2\times 3^{-d})^K = 2^{-k}$$

are equal. So it suffices to prove that

$$\sum \operatorname{diam}(W_j)^d \geqslant 1$$

when (W_j) is a cover of C by the component intervals of C_K. Since these are a cover, all 2^K intervals must appear in the sum and the result is clear.

21 CALCULATING THE HAUSDORFF DIMENSION

Proposition 21.1
If a metric space M has $\dim_{\mathcal{H}} M < 1$, then M is totally disconnected.

Proof:
Consider first a subset X of $\mathbb{R}$ with $\dim_{\mathcal{H}} X < 1$. Any non-empty open interval U has $\mathcal{H}^1(U) = 1$, so X can not contain any such interval. This means that between any two points $x, y \in X$ there is a point $c \notin X$.

Now suppose that x, y are two distinct points of the metric space M. Then

$$f : M \to \mathbb{R} \; ; \quad z \mapsto d(x, z)$$

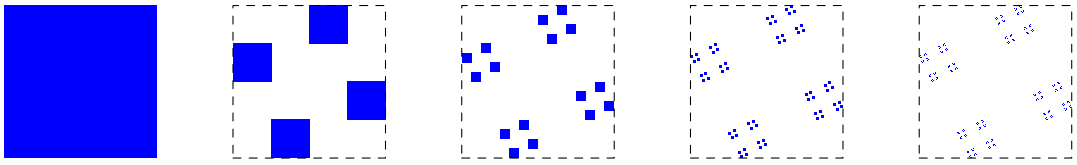
is 1-Lipschitz, so Corollary 20.2 shows that $\dim_{\mathcal{H}} f(M) < 1$. Hence there is a $c \notin f(M)$ with $0 = f(x) < c < f(y)$. The sets $f^{-1}(-\infty, c)$ and $f^{-1}(c, \infty)$ are then disjoint open sets containing x and y respectively with union M. □

The converse fails as the example of *Cantor dust* shows. We construct Cantor dust as follows:

$D_0 = [0, 1] \times [0, 1] \subset \mathbb{R}^2$;

If D_n has been defined and consists of a finite number of disjoint closed squares, then D_{n+1} is obtained by dividing each square into 16 smaller squares and keeping only the 4 shown in the diagram below.

Cantor dust is the intersection $\bigcap D_n$.



D_0 D_1 D_2 D_3 D_4

Each D_n consists of 4^n squares with side length 4^{-k} and hence diameter $4^{-k}\sqrt{2}$. Covering D by these squares shows that

$$\mathcal{H}^1(D) \leqslant \sqrt{2} \; .$$

The projection π_1 onto the first co-ordinate is a 1-Lipschitz map, with $\pi_1(D) = [0, 1]$, so

$$\mathcal{H}^1(D) \geqslant \mathcal{H}^1([0, 1]) = 1 \; .$$

Therefore we see that D has Hausdorff dimension 1. However, it is clear that D is totally disconnected.

21.1 Invariant Sets

Let M be a metric space and $\mathcal{K}(M)$ the collection of non-empty, compact subsets of M. If K is one of these compact sets then

$$K^\delta = \bigcup \{B(x, \delta) : x \in K\}$$

is an open set containing K for each $\delta > 0$. It is called the *δ-neighbourhood of K*. The *Hausdorff distance* D on $\mathcal{K}(M)$ is

$$D(K, L) = \inf\{\delta > 0 : K \subset L^\delta \text{ and } L \subset K^\delta\} \; .$$

Lemma 21.2 Hausdorff distance
The Hausdorff distance D is a metric on $\mathcal{K}(M)$.

Proof:

Let $K, L \in \mathcal{K}(M)$. Choose a point $x \in L$. The compact set K is bounded, so there is some $\delta > 0$ with $K \subset B(x, \delta) \subset L^\delta$. A similar argument with K and L interchanged shows that the Hausdorff distance $D(K, L)$ is finite.

If $D(K, L) = 0$, then $K \subset \bigcap\{L^\delta : \delta > 0\} = L$ and $L \subset K$, so $K = L$.

It is clear that $D(K, L) = D(L, K)$.

For a third set $M \in \mathcal{K}(M)$ suppose that $M \subset L^\varepsilon$ and $L \subset K^\delta$. Then

$$M \subset L^\varepsilon \subset (K^\delta)^\varepsilon \subset K^{\delta+\varepsilon} \ .$$

So we see that $M \subset K^{D(M,L)+D(L,K)}$. Hence we obtain the triangle inequality:

$$D(M, K) \leqslant D(M, L) + D(L, K) \ .$$

□

A *contraction* on the metric space M is a map $C : M \to M$ for which there is a constant c with $0 \leqslant c < 1$ and

$$d(C(x), C(y)) \leqslant c\, d(x, y) \qquad \text{for all } x, y \in M \ .$$

The contraction mapping theorem shows that such a map has a unique fixed point in M, provided that M is complete.

Let $C_1, C_2, \ldots, C_N$ be a finite collection of contractions with constants $c_1, c_2, \ldots, c_N$ respectively. A subset F of M is an *invariant set for* $C_1, C_2, \ldots, C_N$ when

$$F = \bigcup_{n=1}^{N} C_n(F) \ .$$

Proposition 21.3 Invariant sets
A finite set $C_1, C_2, \ldots, C_N$ of contractions on the Euclidean space $\mathbb{E}^M$ have a non-empty, compact, invariant set. There is only one such non-empty, compact, invariant set.

Proof:

For each compact subset $K \in \mathcal{K}(M)$ define $\mathcal{C}(K) = \bigcup_{n=1}^{N} C_n(K)$. Each map C_n is continuous, so $C_n(K)$ is compact and hence $\mathcal{C}(K) \in \mathcal{K}(M)$. We are looking for a set F with $\mathcal{C}(F) = F$.

For two sets $K, L \in \mathcal{K}(M)$ with $K \subset L^\delta$ we have $C_n(K) \subset C_n(L^\delta) \subset C_n(L)^{c_n \delta}$. Hence,

$$\mathcal{C}(K) \subset \mathcal{C}(L)^{c\delta} \qquad \text{where } c = \max\{c_1, c_2, \ldots, c_N\} \ .$$

This implies that

$$D(\mathcal{C}(K), \mathcal{C}(L)) \leqslant c\, D(K, L) \ .$$

Hence $\mathcal{C}$ is a contraction.

If K and L were both non-empty, compact, invariant sets, then $D(\mathcal{C}(K), \mathcal{C}(L)) \leqslant cD(K, L)$, so $K = L$. It remains to show that there is at least one invariant set.

Consider the closed ball $Q = \overline{B}(\mathbf{0}, R)$. Its image under $\mathcal{C}$ lies within a distance $\max\{d(\mathbf{0}, C_n(\mathbf{0}))\} + cR$ of the origin. Hence, if we choose R large enough we will have $\mathcal{C}(Q) \subset Q$. This implies that

$$Q \supset \mathcal{C}(Q) \supset \mathcal{C}^2(Q) \supset \mathcal{C}^3(Q) \supset \mathcal{C}^4(Q) \supset \ldots \ .$$

The intersection $F = \bigcap \mathcal{C}^n(Q)$ is therefore a non-empty compact set. It is clear that $\mathcal{C}(F) = F$, so F is invariant. □

For a contraction mapping $C : M \to M$ we know that the sequence $x, C(x), C^2(x), C^3(x), \ldots$ converges to the fixed point. Hence the proof above show that the sets

$$K, \mathcal{C}(K), \mathcal{C}^2(K), \mathcal{C}^3(K), \ldots$$

converge to the invariant set in the Hausdorff metric for any starting set $K \in \mathcal{K}(M)$. If we choose K as a compact set with $\mathcal{C}(K) \subset K$, then

$$K \supset \mathcal{C}(K) \supset \mathcal{C}^2(K) \supset \mathcal{C}^3(K) \supset \ldots$$

and the intersection $\bigcap \mathcal{C}^n(K)$ is the invariant set.

Exercise:

31. What is the unique non-empty, compact, invariant set for a single contraction $C : M \to M$?

    Let C_0, C_1 be the contractions on $\mathbb{R}$ given by

    $$C_0 : x \mapsto \tfrac{1}{3}x \qquad \text{and} \qquad C_1 : x \mapsto \tfrac{1}{3}x + \tfrac{2}{3} \ .$$

    Find the unique non-empty, compact, invariant set for C_0, C_1. Show that there are other non-empty invariant sets.

We can also use the argument in the last proposition to give an upper bound on the Hausdorff dimension of the invariant set.

Proposition 21.4
Let $C_1, C_2, \ldots, C_N$ be contractions for the Euclidean metric on $\mathbb{E}^M$ with F as the non-empty, compact, invariant set. Let $c_n < 1$ be constants with

$$d(C_n(x), C_n(y)) \leqslant c_n d(x, y) \qquad \text{for all } x, y \in \mathbb{E}^M \ .$$

Then $\dim_{\mathcal{H}} F \leqslant d$ where d is the unique solution to the equation

$$\sum_{n=1}^{N} c_n^d = 1 \ .$$

Proof:

First note that the function $f : s \to \sum c_n^s$ is strictly decreasing, so there is a unique positive number d with $\sum_{n=1}^{N} c_n^d = 1$.

The set F is bounded and so of finite diameter. Fix a natural number K and consider sequences $\boldsymbol{n} = (n(1), n(2), \ldots, n(K))$ of K integers with $1 \leqslant n(k) \leqslant N$. For any such sequence $\boldsymbol{n}$ define a set $\mathcal{C}_{\boldsymbol{n}}(F)$ as

$$\mathcal{C}_{\boldsymbol{n}}(F) := C_{n(1)} \circ C_{n(2)} \circ \ldots \circ C_{n(K)}(F) \ .$$

This has diameter at most $c_{n(1)} c_{n(2)} \ldots c_{n(K)} \mathrm{diam}(F)$. If $c = \max\{c_n : n = 1, 2, \ldots, N\} < 1$ then this diameter is at most $c^K \mathrm{diam}(F)$.

Since F is invariant, we have

$$F = \bigcup_{\boldsymbol{n}} \mathcal{C}_{\boldsymbol{n}}(F)$$

where the union is over all sequences of K integers. Hence the sets $\mathcal{C}_{\boldsymbol{n}}(F)$ form a δ-cover for F provided that $c^K \mathrm{diam}(F) \leqslant \delta$. For this cover we have

$$\sum_{\boldsymbol{n}} \mathrm{diam}(\mathcal{C}_{\boldsymbol{n}}(F))^d \leqslant \sum_{\boldsymbol{n}} (c_{n(1)} c_{n(2)} \dots c_{n(K)} \mathrm{diam}(F))^d = \mathrm{diam}(F)^d \ .$$

The last equality follows from the choice of d to satisfy $\sum c_n^d = 1$. So $\mathcal{H}^d_\delta(F) \leqslant \mathrm{diam}(F)^d$ for every $\delta > 0$. Consequently, $\mathcal{H}^d(F) \leqslant \mathrm{diam}(F)^d$ and the Hausdorff dimension of F is at most d. □

The proof above should be compared with the proof that the Hausdorff d-measure of the Cantor set is finite when $d = \log 2 / \log 3$. The other part of the proof, showing that the Hausdorff d-measure is larger than 0, can also be generalised, although we need stronger restrictions on the contractions and the metric space M. The Theorem below gives a useful result in this direction. We will not prove it since it would involve a little probability.

A map $C : M \to M$ is a *similarity* with scale factor $c \in [0, 1)$ if

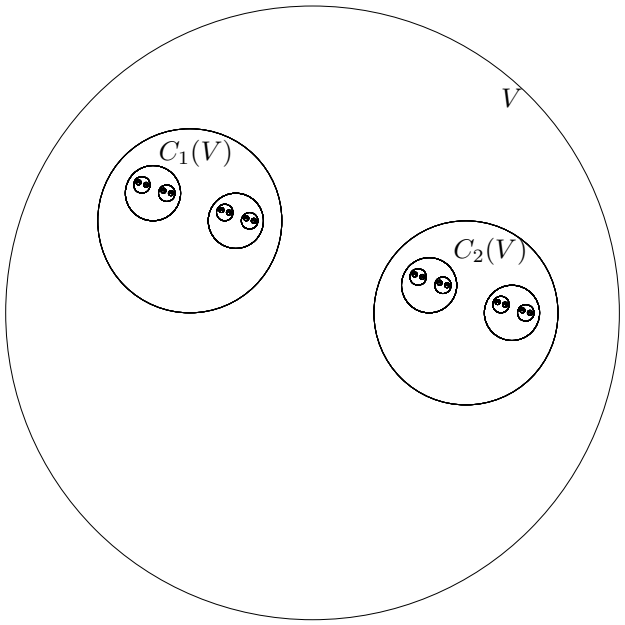
$$d(C(x), C(y)) = c\, d(x, y) \qquad \text{for all } x, y \in M \ .$$

Such a similarity is certainly a contraction map.

Theorem 21.5
Let $C_1, C_2, \dots, C_N$ be similarities on the Euclidean space $\mathbb{E}^M$ with scale factors $c_1, c_2, \dots, c_N$. These similarities have a unique non-empty, compact invariant set F. Let d be the unique number with $\sum c_n^d = 1$. Suppose that there is a bounded open set V with $C_1(V), C_2(V), \dots, C_N(V)$ disjoint and $\bigcup C_n(V) \subset V$. Then the Hausdorff d-measure $\mathcal{H}^d(F)$ is greater than 0 and so the Hausdorff dimension of F is at least d.

Combining this with the previous Proposition shows that the Hausdorff dimension is exactly d and that $0 < \mathcal{H}^d(F) < \infty$.

For a proof of this result, which is not examinable, see "Fractal Geometry", by K. Falconer, pp, 119-120. □



22 EXAMPLES OF HAUSDORFF DIMENSION

We can now determine the Hausdorff dimension of many self-similar sets.

22.1 The Cantor Set

The two similarities

$$C_0 : \mathbb{R} \to \mathbb{R} \; ; \quad x \mapsto \tfrac{1}{3}x \quad \text{and} \quad C_1 : \mathbb{R} \to \mathbb{R} \; ; \quad x \mapsto \tfrac{1}{3}x + \tfrac{2}{3}$$

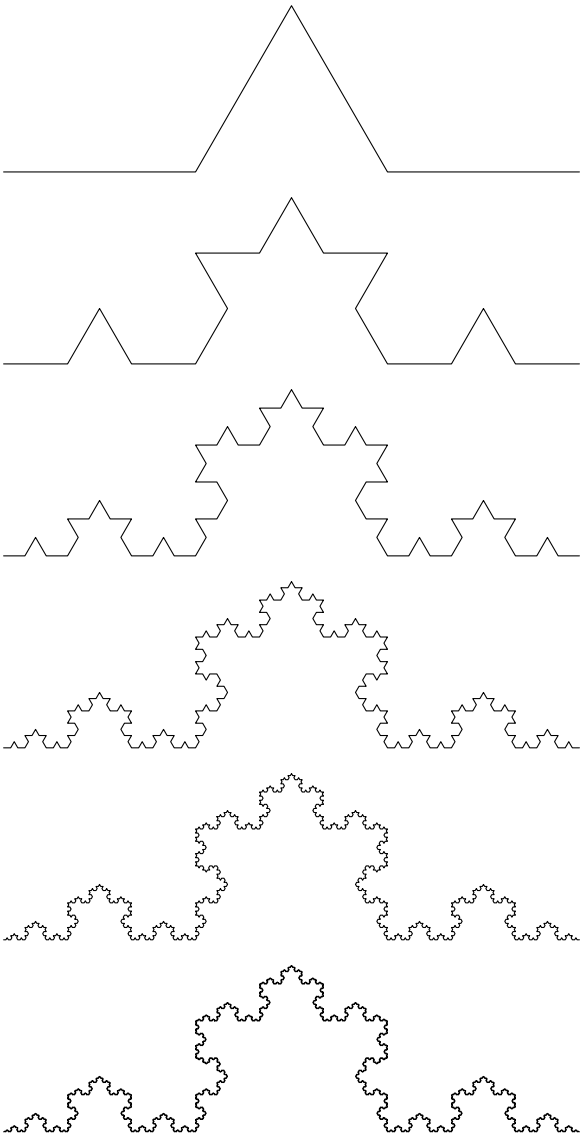
have the Cantor set as an invariant set. For we see that

$$\begin{aligned} \mathcal{C}([0,1]) &= C_1 = [0, \tfrac{1}{3}] \cup [\tfrac{2}{3}, 1] \\ \mathcal{C}^2([0,1]) &= C_2 \\ \mathcal{C}^3([0,1]) &= C_3 \qquad \textit{etc.} \end{aligned}$$

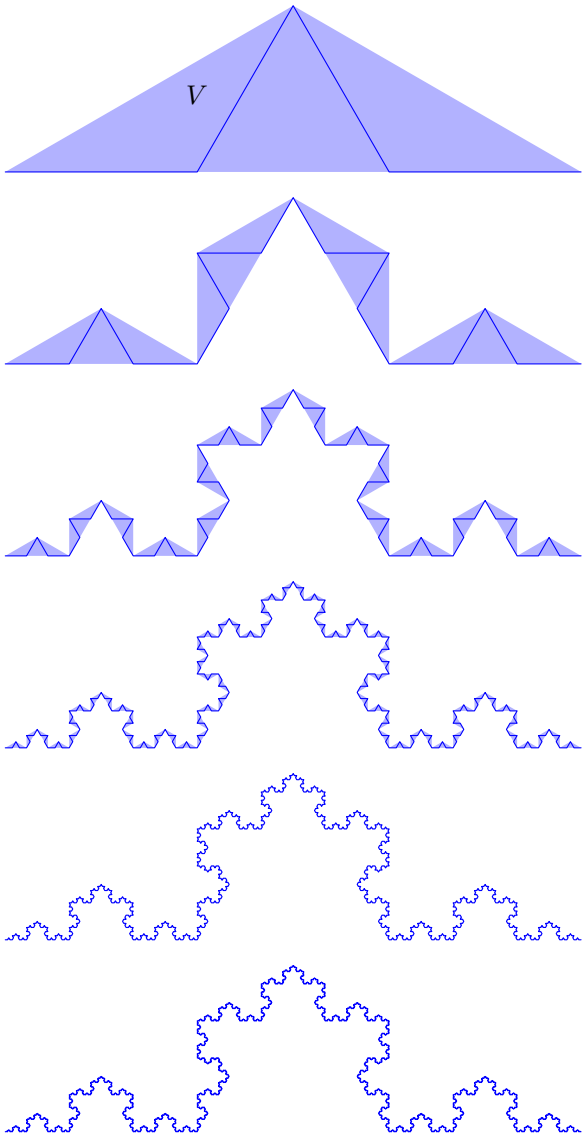
and the invariant set is the limit of this sequence. We can take the open interval $(0,1)$ as the set V in Theorem 21.5 . Hence the dimension d of the Cantor set is the unique solution of $\sum c_n^d = 1$. This is $2 \times (\frac{1}{3})^d = 1$, so $d = \log 2 / \log 3$.

22.2 The von Koch snowflake

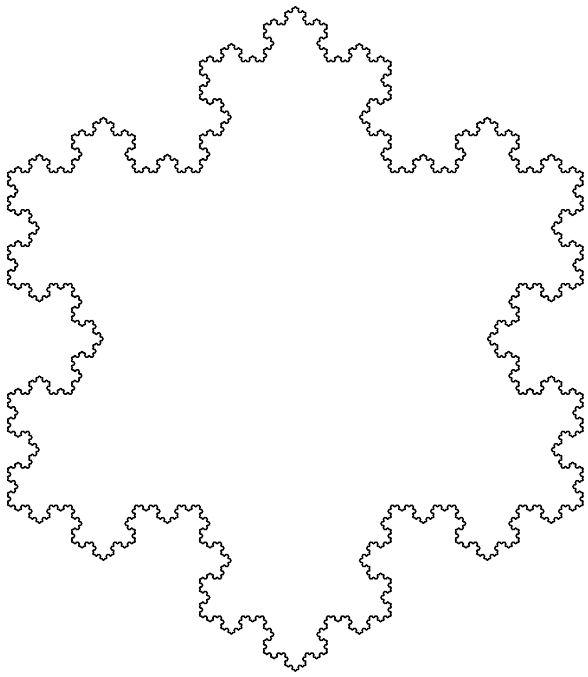
This is the curve constructed by an iterative process starting from the unit interval $[0,1] \subset \mathbb{R}^2$. At each stage, each straight line segment is replaced by 4 line segments each $\frac{1}{3}$ as long, as in the diagram. For this we have 4 similarities each with scale factor $\frac{1}{3}$. We can take the open isosceles triangle with base $(0,1)$ and height $\frac{1}{2}\sqrt{3}$ as the set V. Then $4(\frac{1}{3})^d = 1$, so the Hausdorff dimension of the von Koch snowflake is $\log 4/\log 3$.



The von Koch snowflake

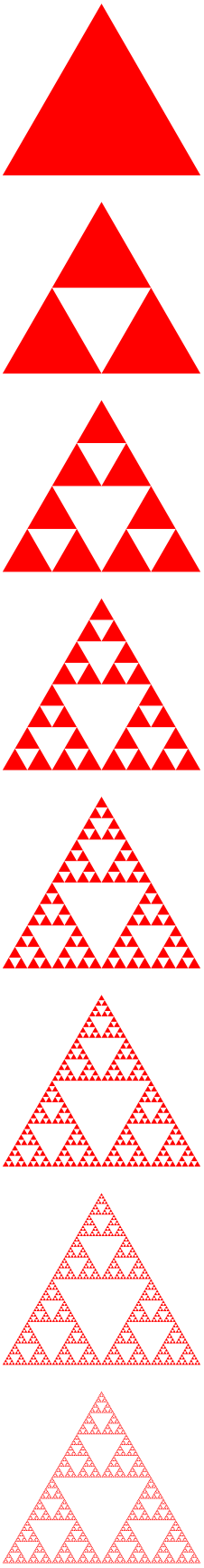


Images of V.



Three copies of the von Koch curve making a snowflake.

22.3 The Sierpiński Gasket



This is the set constructed as follows. We begin with a filled in equilateral triangle. At each stage we replace each equilateral triangle by 3 triangles each $\frac{1}{3}$ the size. So we have 3 similarities each with scale factor $\frac{1}{2}$. We can take the interior of the initial triangle as the set V. Then $3(\frac{1}{2})^d = 1$, so the Hausdorff dimension of Sierpiński's gasket is $\log 3/\log 2$.

23 SCHOTTKY GROUPS

23.1 Fuchsian Groups

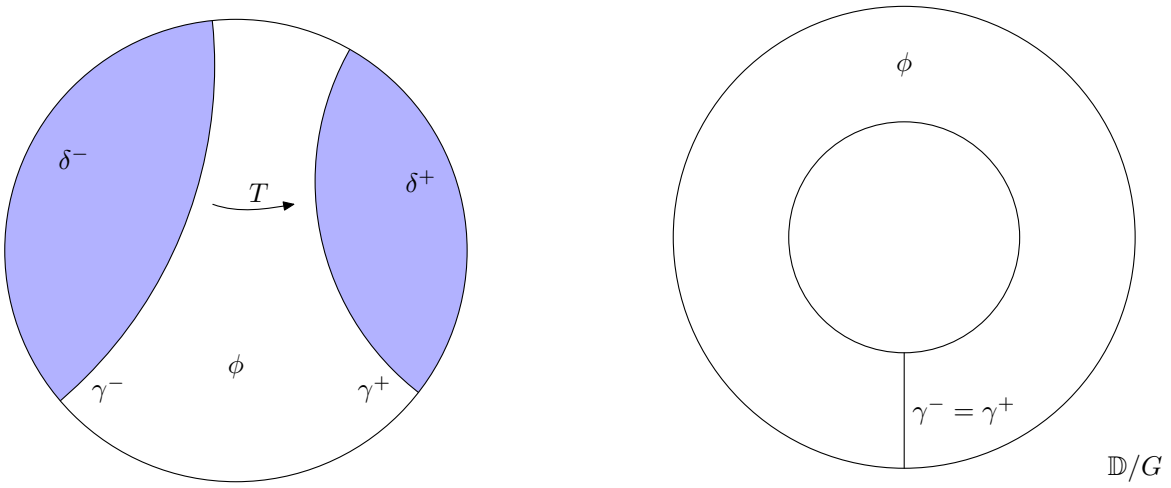
We begin with the following simple exercise.

Proposition 23.1
Let γ^-, γ^+ be two hyperbolic geodesics in $\mathbb{D}$ that do not meet either in $\mathbb{D}$ or on its boundary. These bound two disjoint half-planes δ^-, δ^+. There is a Möbius transformation T that maps γ^- onto γ^+ and δ^- onto $\mathbb{D} \setminus \overline{\delta^+}$.

Proof:
There is a hyperbolic geodesic ν normal to both γ^- and γ^+. This has endpoints w^-, w^+ where we may assume that w^- is in the boundary of δ^- and w^+ in the boundary of δ^+. Conjugate by a Möbius transformation A chosen to map $\mathbb{D}$ onto $\mathbb{R}^2_+$, w^- to 0 and w^+ to ∞. Then ν is mapped to the imaginary axis. So the geodesics $A(\gamma^\pm)$ must be half-circles perpendicular to this: $A(\gamma^\pm) = \{z \in \mathbb{R}^2_\infty : |z| = R^\pm\}$. Now the map

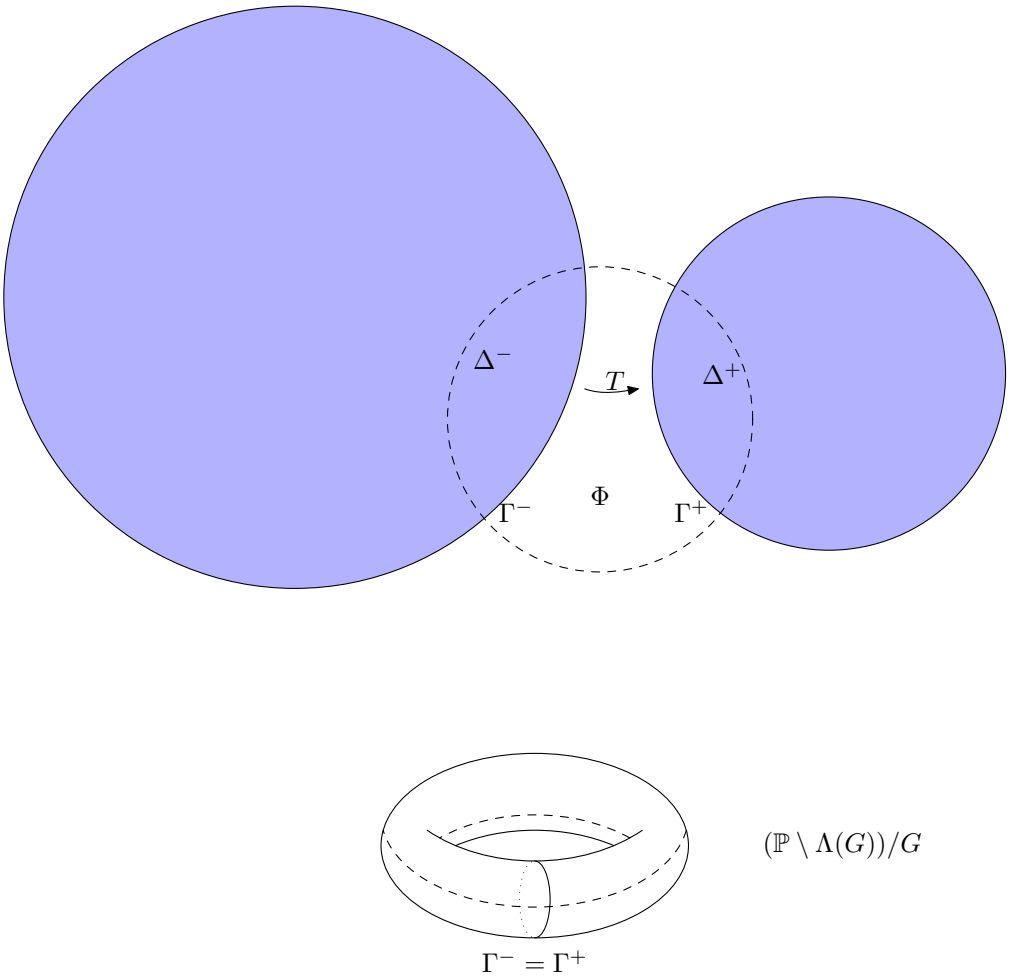
$$U : z \mapsto \left(\frac{R^+}{R^-}\right) z$$

is a hyperbolic Möbius transformation that maps $A(\gamma^-)$ onto $A(\gamma^+)$. So we may take $T = A^{-1} \circ U \circ A$. □



The region $f = \mathbb{D} \setminus (\delta^- \cup \overline{\delta^+})$ between γ^- and γ^+ is a fundamental set for the group G generated by the single Möbius transformation T. The quotient $\mathbb{D}/G$ is obtained by identifying the two sides γ^- and γ^+ of f to get an annulus (ring) as shown above.

We can also think of $\mathbb{D}$ as a subset of the Riemann sphere $\mathbb{P}$. The geodesics $\gamma^\pm$ are parts of circles $\Gamma^\pm$ orthogonal to the unit circle $\partial\mathbb{D}$. These circles enclose two disjoint discs $\Delta^\pm$. The Möbius transformation T acts on all of the Riemann sphere and maps Δ^- onto $\mathbb{P} \setminus \overline{\Delta^+}$. Let G be the group generated by T. Then the limit set $\Lambda(G)$ consists of just the two fixed points of T. The group G acts discontinuously on all of the remainder of $\mathbb{P}$ and the set $\Phi = \mathbb{P} \setminus (\Delta^- \cup \overline{\Delta^+})$ is a fundamental set. The quotient $(\mathbb{P} \setminus \Lambda(G))/G$ is obtained from Φ by identifying the two circles Γ^- and Γ^+ to obtain a torus. We will sometimes abuse the notation by writing $(\mathbb{P} \setminus \Lambda(G))/G$ as $\mathbb{P}/G$.



Note that inversion J in the unit circle $\partial\mathbb{D}$ maps $\mathbb{D}$ to the complementary disc $J(\mathbb{D})$ and sends Φ to itself with the part ϕ inside $\mathbb{D}$ going to the part $J(\phi)$ outside $\mathbb{D}$. The torus $(\mathbb{P} \setminus \Lambda(G))/G$ is obtained by taking the two annuli $\mathbb{D}/G$ and $J(\mathbb{D})/G$ and joining them along their boundaries.

We can also think of the Möbius transformation acting on hyperbolic 3-space $\mathbb{H}^3$. The circles $\Gamma^\pm$ are the boundaries of hyperbolic planes $\mathcal{G}^\pm$ in $B^3 = \mathbb{H}^3$. These bound disjoint half-spaces $\mathcal{H}^\pm$. The Möbius transformation T maps $\mathcal{D}^-$ onto $\mathbb{H}^3 \setminus \overline{\mathcal{D}^+}$. The set $\mathcal{F} = \mathbb{H}^3 \setminus (\mathcal{D}^- \cup \overline{\mathcal{D}^+})$ is a fundamental set for G. The quotient $\mathbb{H}^3/G$ is obtained from $\mathcal{F}$ by identifying the two planes $\mathcal{G}^-$ and $\mathcal{G}^+$ to get a solid torus. Then $(\mathbb{P} \setminus \Lambda(G))/G$ is the boundary of $\mathbb{H}^3/G$.

We can do the same for Fuchsian groups G generated by more than one element. The group G of Möbius transformations acts on $\mathbb{D}$ and we obtain a surface $\mathbb{D}/G$ for the quotient. This surface is orientable since each Möbius transformation is orientation preserving. Similarly, G acts on $J(\mathbb{D}) = \mathbb{P} \setminus \overline{\mathbb{D}}$. The quotient $J(\mathbb{D})/G$ is another surface, called the *dual* of $\mathbb{D}/G$. The inversion J in $\partial\mathbb{D}$ induces an orientation reversing bijection from $\mathbb{D}/G$ to $J(\mathbb{D})/G$.

When we think of G acting on all of the Riemann sphere we get a quotient $(\mathbb{P} \setminus \Lambda(G))/G$ that consists of $\mathbb{D}/G$ and $J(\mathbb{D})/G$ joined together along their boundaries. Similarly, when we think of G acting on $\mathbb{H}^3$ we get a quotient $\mathbb{H}^3/G$ which is a 3-dimensional solid that has $(\mathbb{P} \setminus \Lambda(G))/G$ as its boundary. (Recall that we have seen examples of Fuchsian groups which have all of the unit circle in the limit set $\Lambda(G)$. In this case, the two parts $\mathbb{D}/G$ and $J(\mathbb{D})/G$ are separated by the quotient of the limit set.)

23.2 Schottky Groups

In the previous section all of our circles were orthogonal to the unit circle. In this section we do not insist on this and produce Kleinian groups rather than Fuchsian groups. These are the Schottky groups. As in the last section, we will begin by considering groups generated by a single Möbius transformation.

Proposition 23.2
Let Γ^-, Γ^+ be two disjoint circles bounding two disjoint discs Δ^-, Δ^+ in $\mathbb{P}$. Then there is a Möbius transformation T that maps Γ^- onto Γ^+ and Δ^- onto $\mathbb{P} \setminus \overline{\Delta^+}$.

Proof:
We could prove this by adapting the proof of Proposition 23.1 for hyperbolic 3-space. However, for variety, we will give a different argument that works entirely on the Riemann sphere.

Let $J^\pm$ be inversion in the circle $\Gamma^\pm$. Then $S = J^+ \circ J^-$ is a non-identity Möbius transformation. Let z_o be one fixed point. Then $J^+(J^-(z_o)) = S(z_o) = z_o$ so $J^-(z_o) = J^+(z_o)$. Also,

$$S(J^-(z_o)) = J^+ \circ J^- \circ J^-(z_o) = J^+(z_o) = J^-(z_o)$$

so $J^-(z_o)$ is also fixed by S. Since Γ^- and Γ^+ are disjoint, the points z_o and $J^-(z_o)$ must be distinct. So S has 2 fixed points. Note that J^- and J^+ both interchange these two fixed points.

Conjugate by a Möbius transformation A that sends z_o to 0 and $J^-(z_o)$ to ∞. Then $A \circ J^\pm \circ A^{-1}$ is inversion in the circle $A(\Gamma^\pm)$ and must interchange 0 and ∞. So $A(\Gamma^\pm)$ must be a circle $\{z \in \mathbb{P} : |z| = r^\pm\}$ for some $0 < r^\pm < \infty$. Consequently the map

$$U : z \mapsto e^{i\theta} \left(\frac{r^+}{r^-}\right) z$$

maps $A(\Gamma^-)$ onto $A(\Gamma^+)$. (It is hyperbolic if $e^{i\theta} = 1$ and loxodromic otherwise.) The map $T = A \circ U \circ A^{-1}$ now has the required properties. □

Let $\Gamma^-, \Gamma^+, \Delta^-, \Delta^+$ and T be as in the proposition. The group G generated by T has a limit set $\Lambda(G)$ consisting of the two fixed points of T. The set $\Phi = \mathbb{P} \setminus (\Delta^- \cup \overline{\Delta^+})$ is a fundamental set for the group G acting on $\mathbb{P} \setminus \Lambda(G)$. The quotient $\mathbb{P}/G$ (or, more accurately, $(\mathbb{P} \setminus \Lambda(G))/G$) is obtained from this fundamental set by identifying the two circles Γ^- and Γ^+. So the quotient is a torus (the surface of a ring doughnut).

We can also think of the Möbius transformation acting on hyperbolic 3-space $\mathbb{H}^3$. The circles $\Gamma^\pm$ are the boundaries of hyperbolic planes $\mathcal{G}^\pm$ in $B^3 = \mathbb{H}^3$. These bound disjoint half-spaces $\mathcal{D}^\pm$. The Möbius transformation T maps $\mathcal{D}^-$ onto $\mathbb{H}^3 \setminus \overline{\mathcal{D}^+}$. The set $\mathcal{F} = \mathbb{H}^3 \setminus (\mathcal{D}^- \cup \overline{\mathcal{D}^+})$ is a fundamental set for G. The quotient $\mathbb{H}^3/G$ is obtained from $\mathcal{F}$ by identifying the two planes $\mathcal{G}^-$ and $\mathcal{G}^+$ to get a solid torus (the body of a ring doughnut). Then $\mathbb{P}/G$ is the boundary of $\mathbb{H}^3/G$.

The special case when the circles Γ_n are orthogonal to the unit circle was the Fuchsian case dealt with in the previous section.

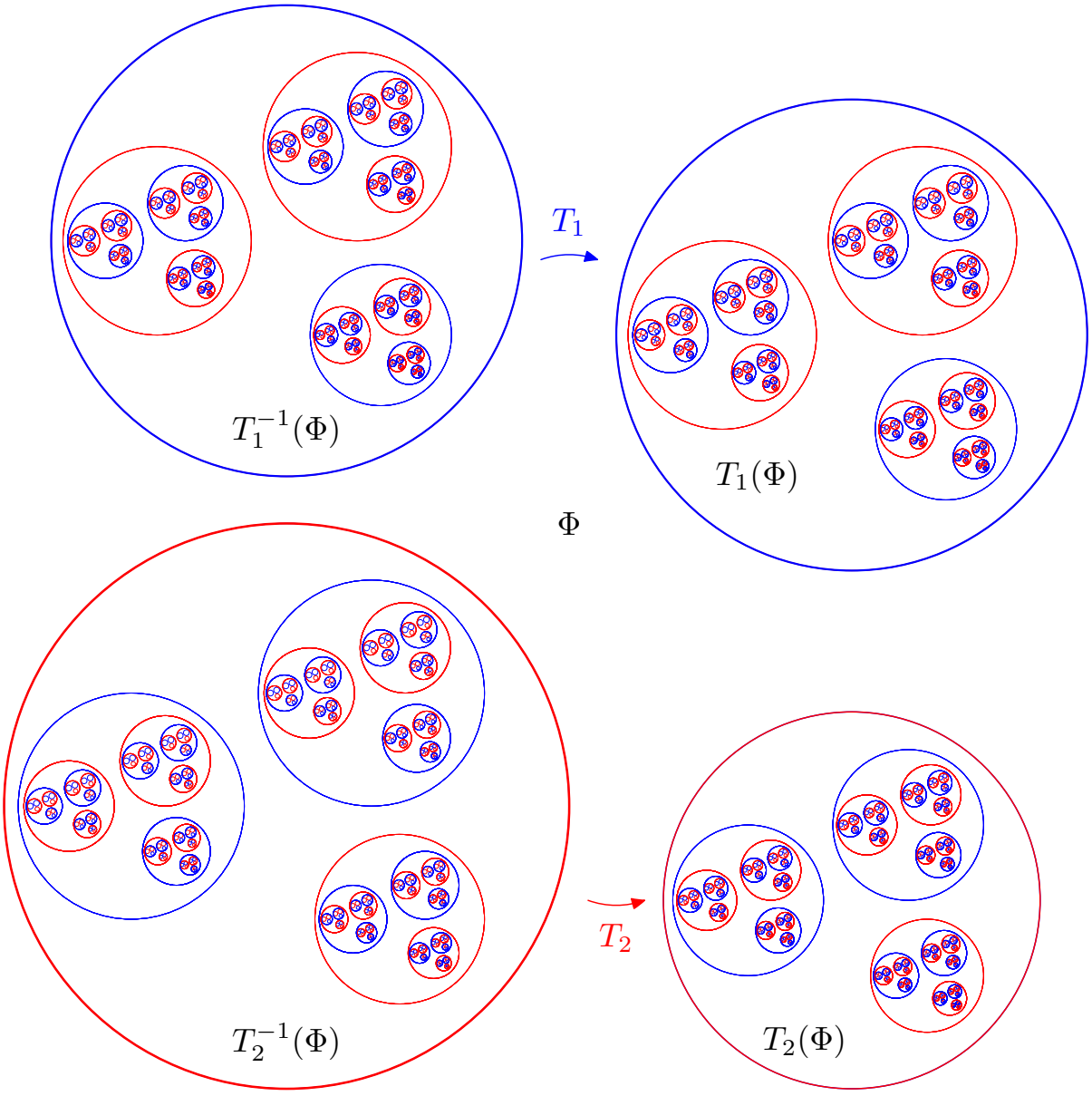
We can do the same when we have more than one pair of circles. In this case we get a Schottky group generated by several Möbius transformations. For $k = 1, 2, \ldots, K$, let Γ_k^- and Γ_k^+ be disjoint circles bounding $2K$ disjoint discs Δ_k^- and Δ_k^+. The proposition shows that there are Möbius transformations T_k that map Γ_k^- onto Γ_k^+ and Δ_k^- onto $\mathbb{P} \setminus \overline{\Delta_k^+}$. The group G generated by $T_1, T_2, \ldots, T_K$ is called the *(classical) Schottky group* for the discs. We have already considered the Schottky groups for $K = 1$ but Schottky groups with more than one generator are more interesting.

Let Φ be the set

$$\Phi = \mathbb{P} \setminus \left(\bigcup_{k=1}^{K} \Delta_k^- \cup \overline{\Delta_k^+} \right) .$$

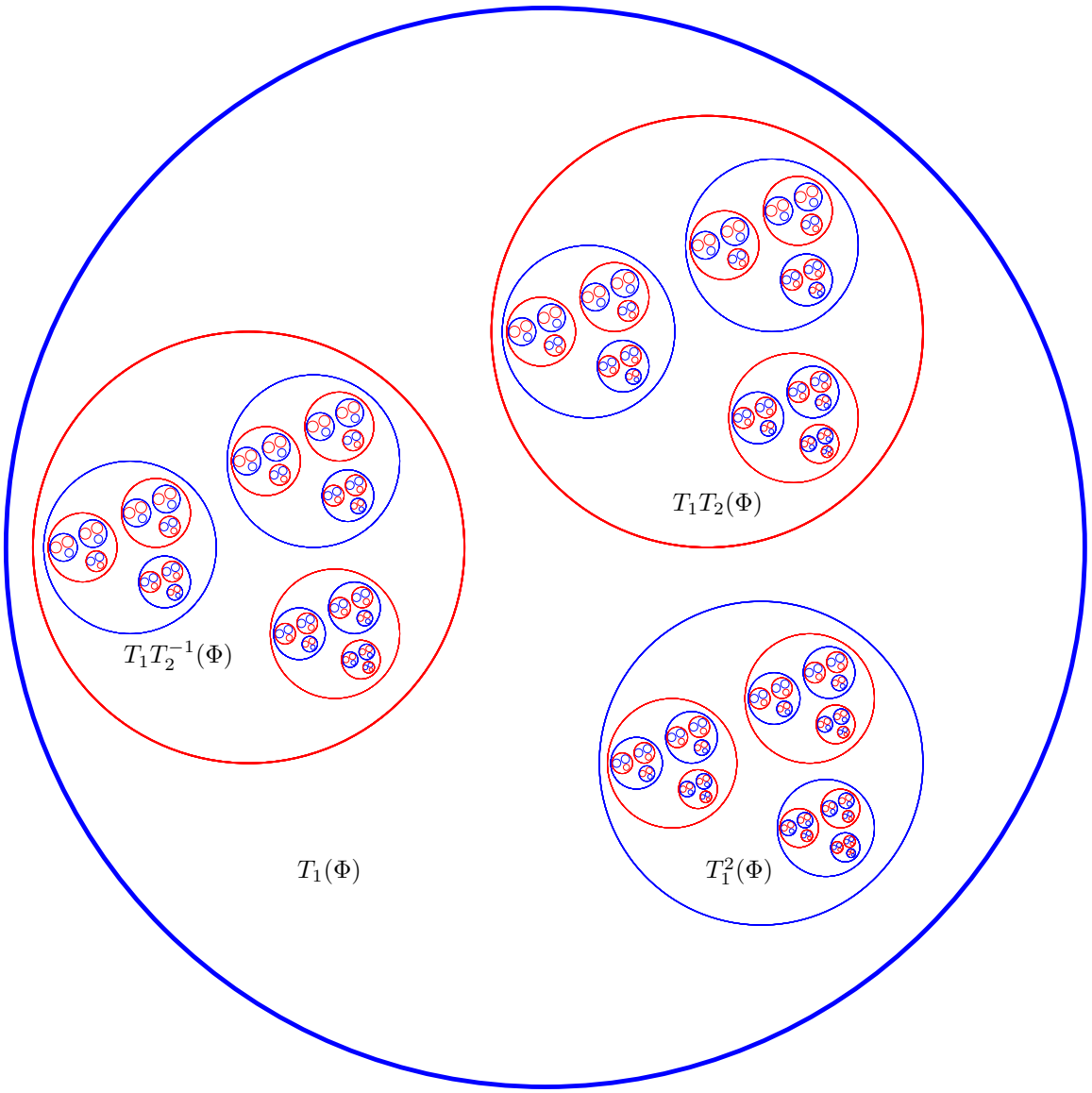
This is the region outside all of the discs $\Delta_k^{\pm}$ except that we have included half of the bounding circles. We will show that Φ is a fundamental set for the Schottky group G acting on $\mathbb{P} \setminus \Lambda(G)$.

The image $T_k(\Phi)$ lies inside Δ_k^+ and is bounded by the $2K$ circles $T_k(\Gamma_j^{\pm})$. One of these is $T_k(\Gamma_k^-) = \Gamma_k^+$ but all of the others are strictly inside Δ_k^+. Now apply another one of the generators, say T_j. The image $T_j(T_k(\Phi))$ is a subset of $T_j(\Delta_k^+)$, which is itself a disc inside Δ_j^+. Hence we get a pattern of nested discs as shown below.



The Tessellation for the Schottky Group.

We will see shortly that the intersection of any chain of nested discs is a point and the closure of these points is the limit set $\Lambda(G)$. The remainder of the Riemann sphere: $\mathbb{P} \setminus \Lambda(G)$, is tessellated by the images $g(\Phi)$ for $g \in G$. These are clearly locally finite so we see that G acts discontinuously on $\mathbb{P} \setminus \Lambda(G)$.

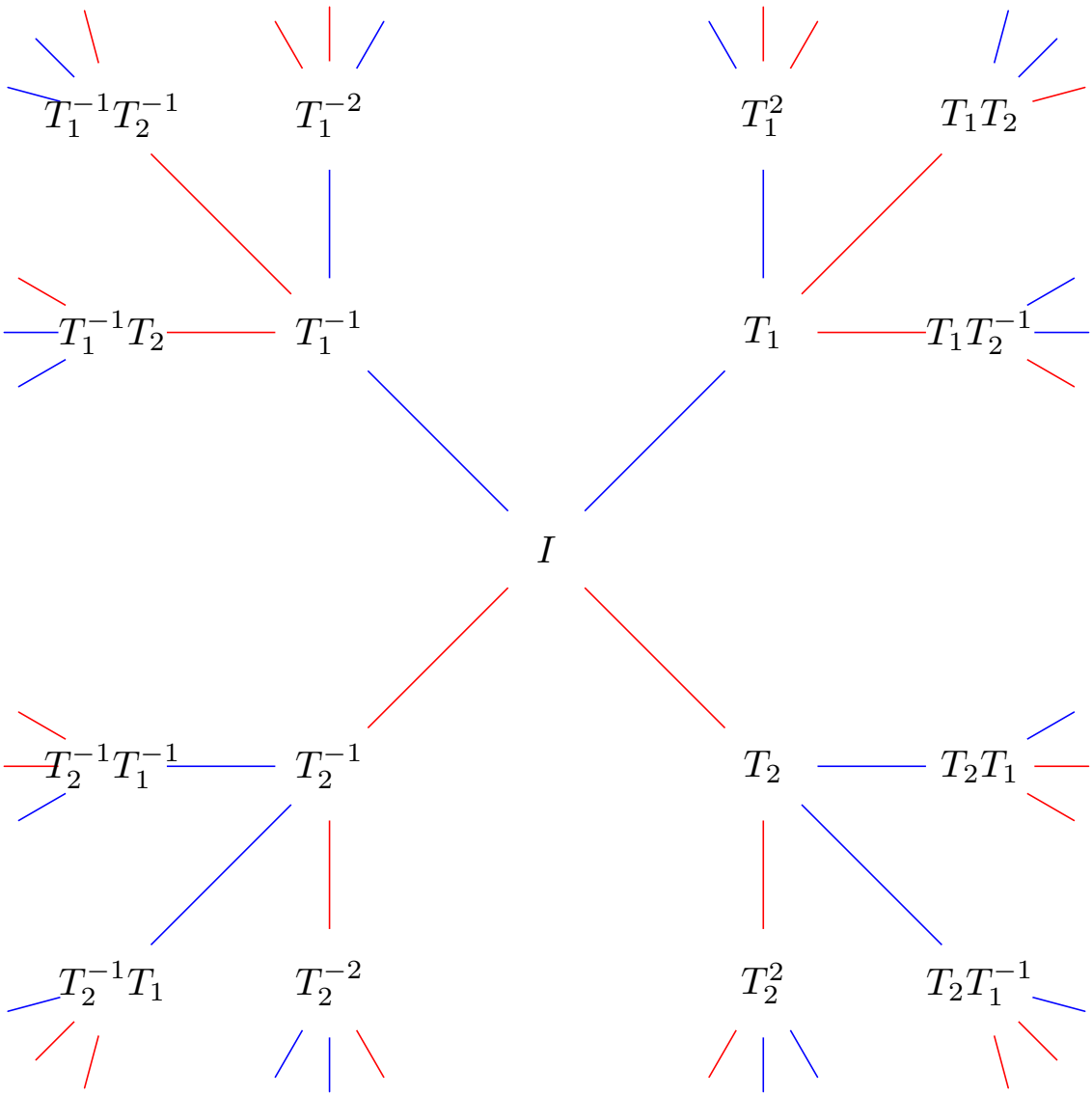


Enlargement of the region Δ_1^+.

Theorem 23.3 Schottky groups are free groups.
The Schottky group generated by the Möbius transformations pairing discs Δ_k^- and Δ_k^+ for $k = 1, 2, 3 \ldots, K$ with the closure of all these $2K$ discs disjoint, is a free group on the generators $(T_k)_{k=1}^K$.

Proof:

To see this we need to look more closely at the tessellation shown above. It is simplest to see the structure of this tessellation if we simplify the diagram. So draw a graph — the *Cayley graph for G* — by putting a vertex for each image $g(\Phi)$ and labelling it with the element g of G. Join the vertices labelled g and $g \circ T_k$ for $k = 1, 2, \ldots, K$. We have done this below for the case of two pairs of discs and coloured the edges from g to $g \circ T_1$ in blue and the edges from g to $g \circ T_2$ in red. Two vertices g and h in this graph are adjacent when $g = hA$ for A one of the generators T_k or their inverses T_k^{-1}.



The Cayley graph for a Schottky Group.

Consider a product

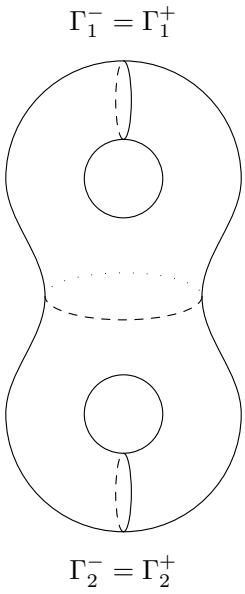
$$A_1 \circ A_2 \circ A_3 \circ \ldots \circ A_N$$

where $N \geqslant 0$ and each A_n is one of T_k or T_k^{-1}. If two successive terms A_n, A_{n+1} have $A_n \circ A_{n+1} = I$, then we can cancel them and reduce the length N by 2. Repeat this until there are no such pairs. We need to show that no such product is the identity I (except for the trivial product of no elements). In the graph, the path

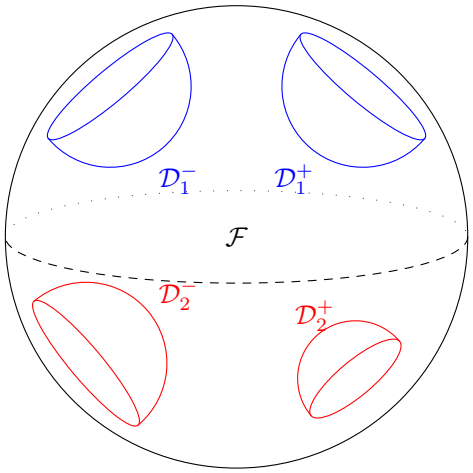
$$I,\ A_1,\ A_1 \circ A_2,\ A_1 \circ A_2 \circ A_3,\ \ldots,\ A_1 \circ A_2 \circ A_3 \circ \ldots \circ A_N = g$$

follows edges from I to g and never turns back on itself. Since there are no loops in the graph, this implies that the path can not return to its starting point. So $g \neq I$ unless the product is trivial. Hence G is a free group. □

The quotient $(\mathbb{P}\setminus\Lambda(G))/G$ is obtained from the fundamental set Φ by identifying the pairs of circles Δ_k^- and Δ_k^+ for $k = 1, 2, \ldots, K$. This gives a sphere with K handles.



We can also think of the Schottky group acting on the hyperbolic 3-space $\mathbb{H}^3$. The circles $\Gamma_k^\pm$ are the boundaries of hyperbolic planes $\mathcal{G}_k^\pm$ in $B^3 = \mathbb{H}^3$. These bound disjoint half-spaces $\mathcal{D}_k^\pm$. The Möbius transformations T_k maps $\mathcal{D}_k^-$ onto $\mathbb{H}^3 \setminus \overline{\mathcal{D}_k^+}$. The set $\mathcal{F} = \mathbb{H}^3 \setminus (\mathcal{D}^- \cup \overline{\mathcal{D}^+})$ is a fundamental set for G. The quotient $\mathbb{H}^3/G$ is obtained from $\mathcal{F}$ by identifying the two planes $\mathcal{G}_k^-$ and $\mathcal{G}_k^+$ for $k = 1, 2, \ldots, K$. This gives a solid ball with K handles. Then $(\mathbb{P} \setminus \Lambda(G))/G$ is the boundary of $\mathbb{H}^3/G$.



23.2 The Limit Set for a Schottky Group

The pictures above suggest that nested sequences of the discs obtained as images of the basic discs $\Delta_k^\pm$ under the group G are single points and that these points are dense in the limit set. We will prove this and also obtain a bound on the Hausdorff dimension of the limit set.

It is convenient to work in $\mathbb{H}^3$. The hyperbolic planes $\mathcal{D}_k^-$ and $\mathcal{D}_k^+$ do not meet either in $\mathbb{H}^3$ nor on its boundary, so they are all at least a non-zero hyperbolic distance t apart. This means that any curve in the fundamental set $\mathcal{F}$ that joins one of these planes to another must have length at least t.

We can write any $g \in G$ as a product

$$g = A_1 \circ A_2 \circ \ldots \circ A_N$$

where $N \geqslant 0$, each A_n is one of T_k or T_k^{-1} and $A_n \circ A_{n+1} \neq I$ for $n = 1, 2, \ldots, N-1$. Since G is a free group, the length N of such a product is determined by g. We will call it the *length* of g. There is 1 element of length 0, $2K$ of length 1, and $2K(2K-1)^{N-1}$ of length N.

Choose a base point in $\mathcal{F}$, say the origin $\mathbf{0}$. Let γ be the shortest hyperbolic path from this base point to the set $g(\mathcal{F})$. This path begins in $\mathcal{F}$, then crosses into $A_1(\mathcal{F})$, then $A_1 \circ A_2(\mathcal{F})$ and so on until it crosses into $A_1 \circ A_2 \circ \ldots \circ A_{N-1}(\mathcal{F})$ and finally meets $g(\mathcal{F})$. This means that it must cross each of the regions $A_1(\mathcal{F})$, $A_1 \circ A_2(\mathcal{F})$, $\ldots$, $A_1 \circ A_2 \circ \ldots \circ A_N(\mathcal{F})$ from one of the bounding planes to another. Hence the hyperbolic length of γ must be at least Nt.

Lemma 23.4
Let $\mathcal{D}$ be a hyperbolic plane at a hyperbolic distance ρ from the origin in $B^3 = \mathbb{H}^3$. Then the Euclidean diameter of $\mathcal{D}$ is at most $2/\sinh\rho$.

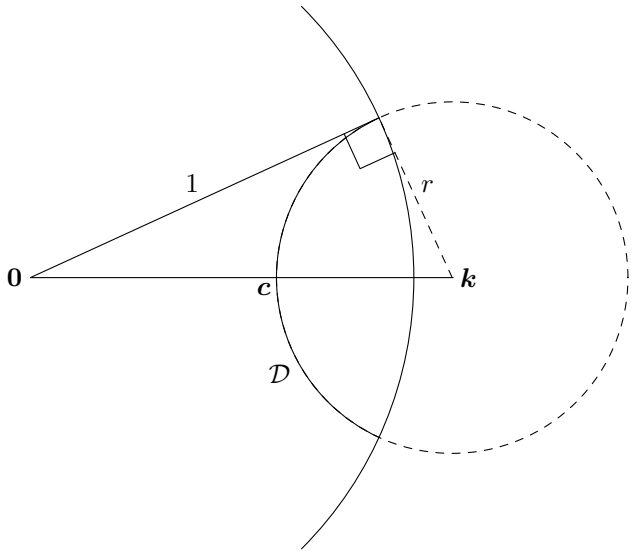
This is essentially the same result as Lemma 19.4 . The inequality is only useful when ρ is large. For small ρ the observation that $\operatorname{diam}(\mathcal{D}) \leqslant 2$ is better.

Proof:
The hyperbolic plane $\mathcal{D}$ is part of a Euclidean sphere with centre $\boldsymbol{k}$ and Euclidean radius r. The shortest path from the origin to $\mathcal{D}$ is a radial line from $\mathbf{0}$ to $\boldsymbol{k}$ that crosses the plane $\mathcal{D}$ at a point $\boldsymbol{c}$ with $||\boldsymbol{c}|| = \tanh\frac{1}{2}\rho$. Since $\mathcal{D}$ is orthogonal to the unit sphere, Pythagoras' Theorem shows that

$$1 + r^2 = ||\boldsymbol{k}||^2 = (||\boldsymbol{c}|| + r)^2 \ .$$

So

$$r = \frac{1 - ||\boldsymbol{c}||^2}{2||\boldsymbol{c}||} = \frac{1 - \tanh^2\frac{1}{2}\rho}{2\tanh\frac{1}{2}\rho} = \frac{1}{\sinh\rho} \ .$$



The Euclidean diameter of $\mathcal{D}$ is at most $2r$. □

It follows from this lemma that an element g of length N in G has $g(\mathcal{F})$ at a hyperbolic distance $\rho \geqslant Nt$ from $\mathbf{0}$ and hence the Euclidean diameter of $g(\mathcal{F})$ is at most $2/\sinh Nt$. When we look at this on the Riemann sphere it shows that $\operatorname{diam}(g(\Phi)) \leqslant 2/\sinh Nt$.

Consider the tessellation given by the images $g(\Phi)$ for g in the Schottky group G. We wish to study a chain of these images $g_n(\Phi)$ where every successive pair $g_n(\Phi)$ and $g_{n+1}(\Phi)$ meet along a circle. Now the fundamental domain Φ only meets the images $A(\Phi)$ where A is one of T_k or T_k^{-1}. Hence $g_n(\Phi)$ and $g_{n+1}(\Phi)$ only meet when $g_{n+1} = g_n \circ A_n$ where $A_n \in \{T_k, T_k^{-1} : k = 1, 2, \ldots, K\}$. Thus we have

$$g_n = A_1 \circ A_2 \circ \ldots \circ A_n \ .$$

Note that we can not have $A_n \circ A_{n+1} = I$ or else $g_{n-1} = g_{n+1}$ and the two images $g_{n-1}(\Phi)$ and $g_{n+1}(\Phi)$ are the same. The images $g_n(\Phi)$ touch each other along a sequence of circles that are nested inside one another. (Look at the diagram of the Tessellation for the Schottky Group.) We can use the previous lemma to estimate the size of these circles as

$$\operatorname{diam}(g_n(\Phi)) \leqslant \frac{2}{\sinh nt} \ .$$

This certainly shows that their Euclidean diameter tends to 0 as $n \to \infty$.

We can also use this idea to find an upper estimate for the Hausdorff dimension of the limit set $\Lambda(G)$. Suppose that C_n is the circle separating $g_n(\Phi)$ from $g_{n+1}(\Phi)$. Then C_n surrounds a disc containing all of the images $g_m(\Phi)$ for $m > n$. Hence this disc must contain a limit point of G. So every sequence (g_n) as above gives us a point of the limit set. However, no point of any of the images $g(\Phi)$ can be in the limit set since there are only finitely many copies $h(\Phi)$ for $h \in G$ that border it. Therefore the limit set is the complement of all these images $g(\Phi)$.

For any natural number N, the complement of the union:

$$\bigcup \{g(\Phi) : g \in G \ \text{ has length at most } N\}$$

certainly contains the limit set. Our argument shows that the complement consists of $2K(2K-1)^{N-1}$ discs each with Euclidean diameter at most $\frac{2}{\sinh Nt}$. So the Hausdorff d-dimensional measure of $\Lambda(G)$ is at most

$$\mathcal{H}^d_\delta(\Lambda(G)) \leqslant 2K(2K-1)^{N-1} \left(\frac{2}{\sinh Nt}\right)^d$$

for $\delta > \frac{2}{\sinh Nt}$. When N is large this is approximately

$$2K(2K-1)^{N-1} \left(4e^{-Nt}\right)^d = \frac{2K \times 4^d}{2K-1} \left((2K-1)e^{-td}\right)^N \ .$$

So we see that the d-dimensional Hausdorff measure is 0 for

$$d > \frac{\log(2K-1)}{t} \ .$$

This proves that the Hausdorff dimension of the limit set is at most

$$\frac{\log(2K-1)}{t} \ .$$

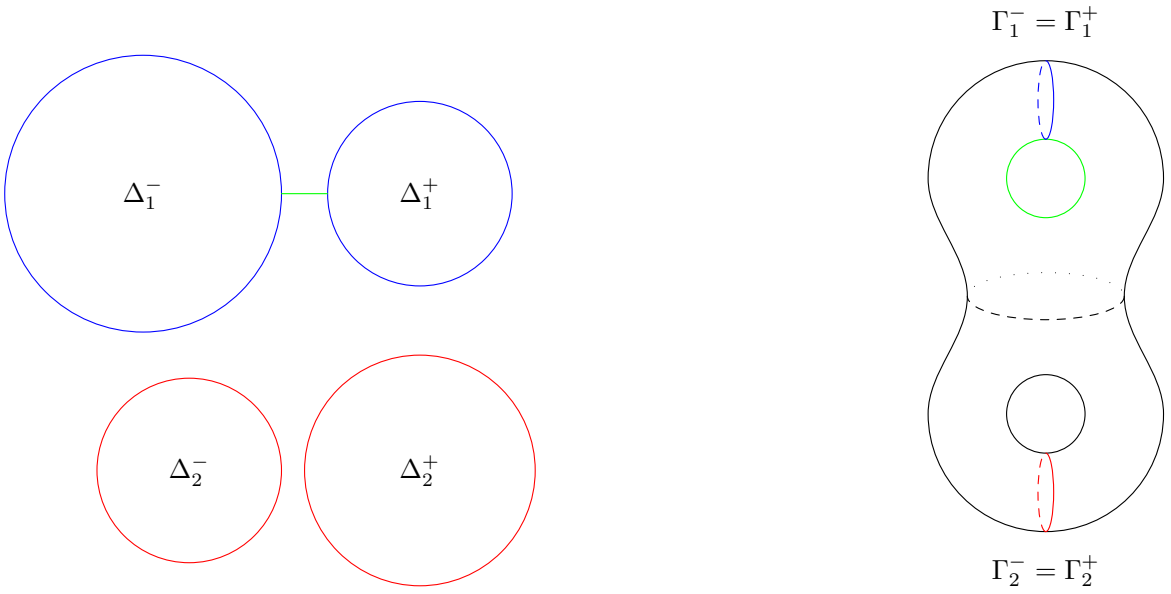
(Much more delicate arguments show that the Hausdorff dimension of the limit set is actually equal to the infimum of those d for which the series $\sum_{g \in G} \exp(-d\rho(\mathbf{0}, g(\mathbf{0})))$ converges.)

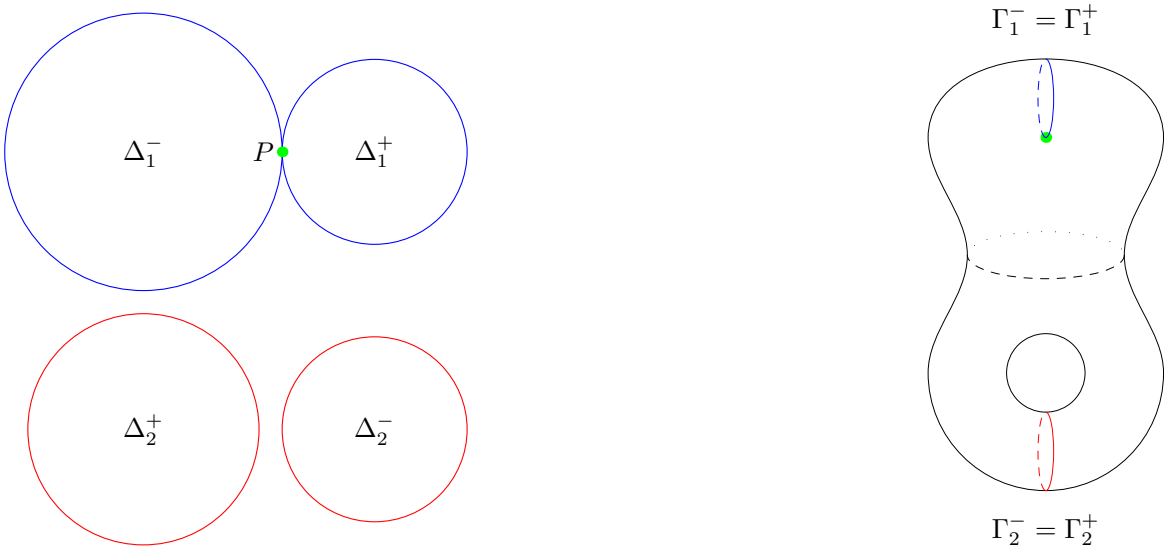
24 DEGENERATE SCHOTTKY GROUPS

24.1 How Schottky Groups Degenerate

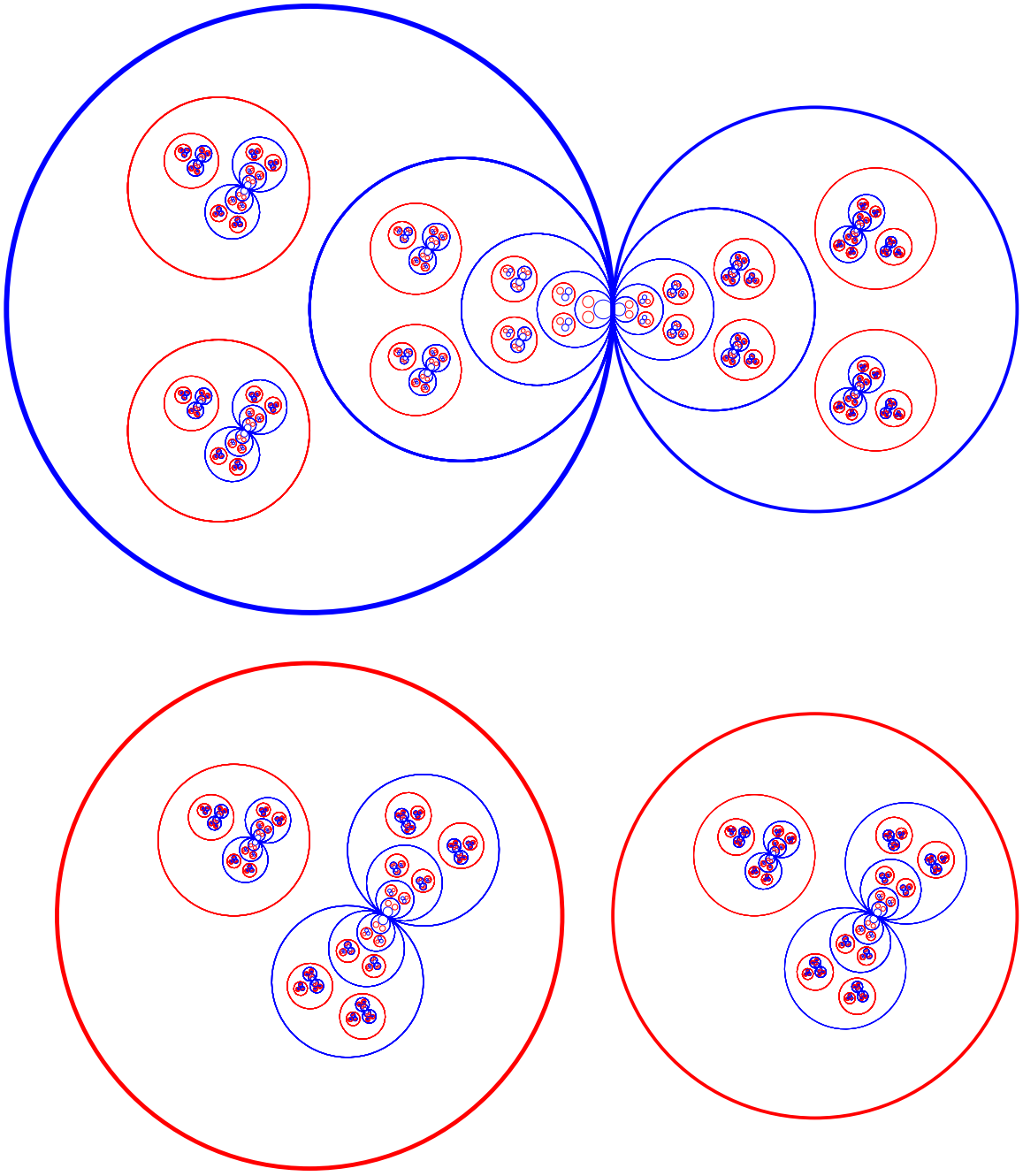
So far we have insisted that the discs $\Delta_k^{\pm}$ for our Schottky groups are disjoint. It is interesting to ask what happens as they move closer to one another and, eventually touch or cross. There are still Möbius transformations that pair the circles $\Gamma_k^{\pm}$. However, the group generated by them need not be discrete. For example, if the circles crossed and the map T_k were elliptic, then G could only be discrete if T_k were of finite order.

Consider the Schottky group for the disjoint discs $(\Delta_k^{\pm})_{k=1}^K$. Take one pair, say Δ_1^- and Δ_1^+, and allow them to move closer. This means that the generator T_1 also varies. In the picture below the shortest (chordal) path between the two discs is marked in green. As the discs move closer together, so this green curve becomes shorter. Ultimately, the green curve reduce to a point and the two discs Δ_1^- and Δ_1^+ touch at a point P. In the quotient, the point P corresponds to a singular point of the surface.





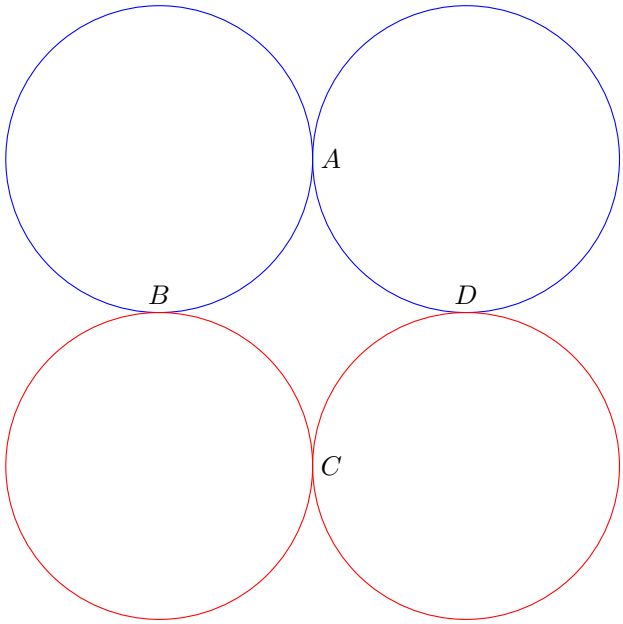
The transformation T_1 also changes as we vary the discs. The most interesting case is when, in the limit, T_1 becomes a parabolic transformation that fixes P. Then we get a tessellation as shown below.

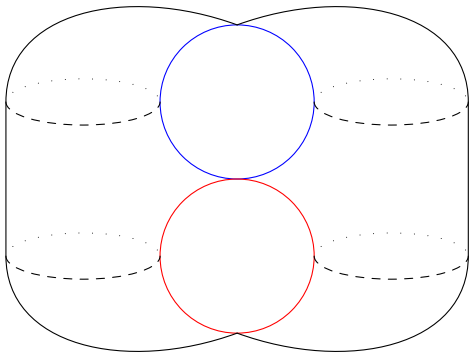


We can also let other pairs of circles touch. Let us consider the case where we have 2 pairs of circles $\Gamma_1^{\pm}$ and $\Gamma_2^{\pm}$ as shown below. We will also assume that the transformations T_1, T_2 are parabolic with

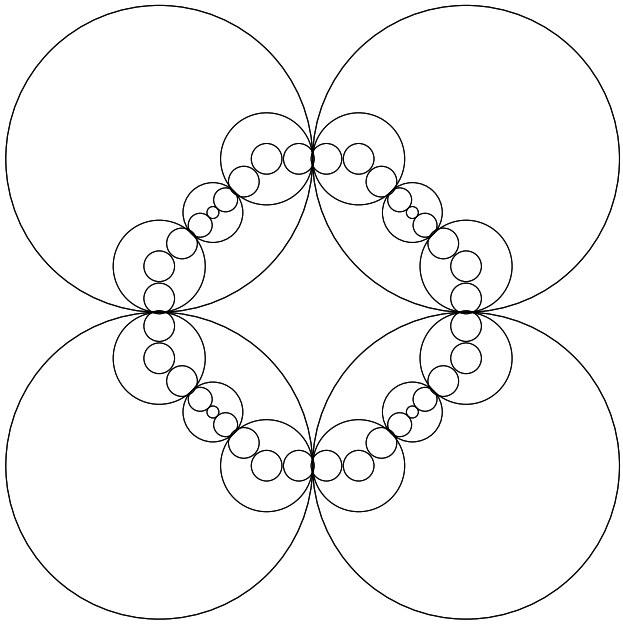
$$T_1 : A \mapsto A \quad ; \quad T_2 : C \mapsto C \; ;$$
$$T_1 : B \mapsto D \quad ; \quad T_2 : B \mapsto D \; .$$

The fundamental set then breaks into two parts, both quadrilaterals bounded by arcs of each of the four circles. When we identify the edges using T_1 and T_2, each of these quadrilaterals becomes a sphere with 3 punctures.





The limit set in this case is a closed Jordan curve contained in the chains of circles shown below.



Exercise:

32. Show that, in the special case where the 4 points A, B, C, D lie on a circle that crosses the 4 circles $\Gamma_k^{\pm}$ orthogonally, the group is a Fuchsian group.

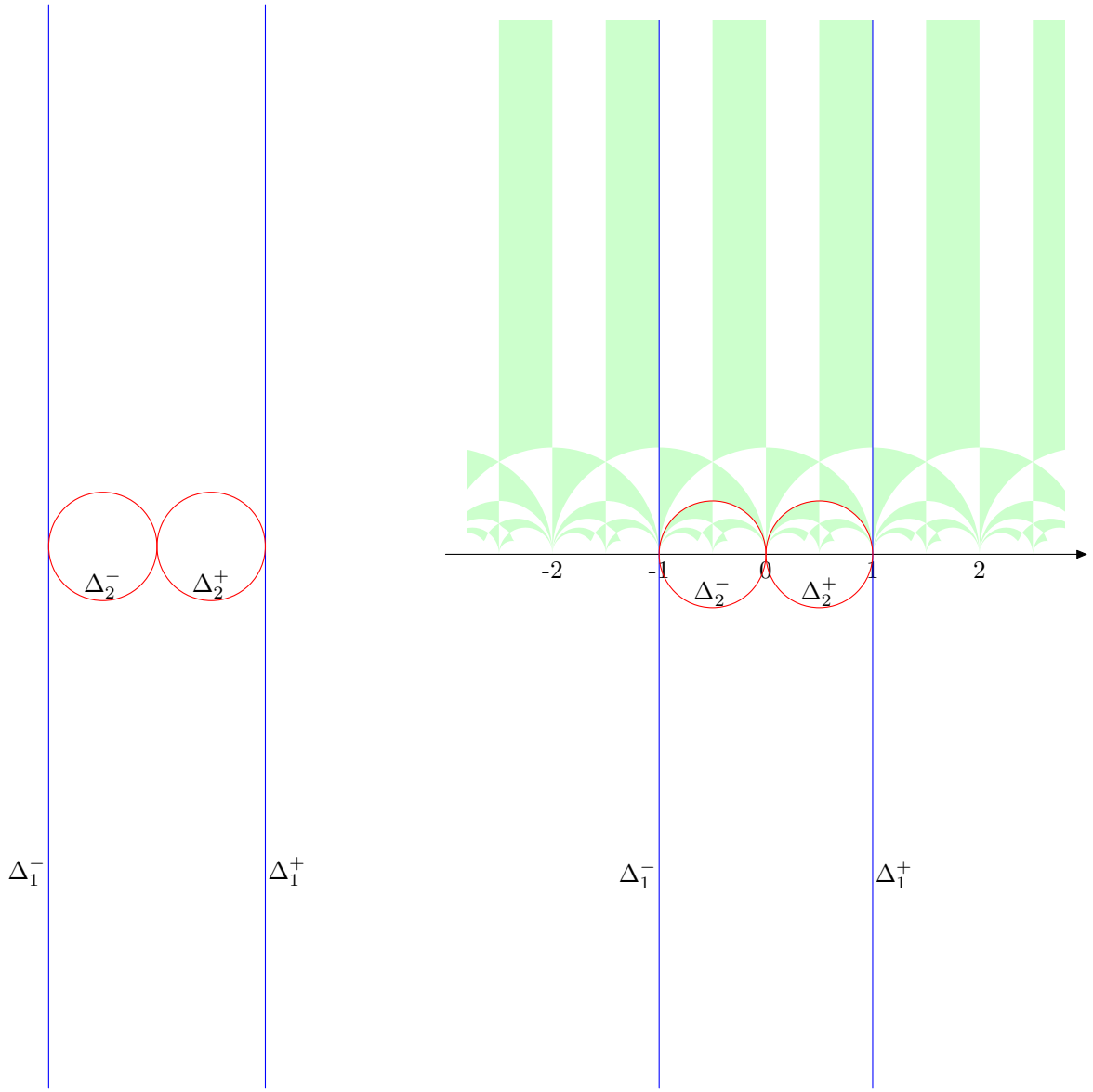
A special case of this is when our discs are:

$$\Delta_1^- = \{z \in \mathbb{C}_\infty : \mathrm{Re}(z) < -1\} \qquad \Delta_1^+ = \{z \in \mathbb{C}_\infty : \mathrm{Re}(z) > 1\}$$
$$\Delta_2^- = \{z \in \mathbb{C}_\infty : |z + \tfrac{1}{2}| < \tfrac{1}{2}\} \qquad \Delta_2^+ = \{z \in \mathbb{C}_\infty : |z - \tfrac{1}{2}| < \tfrac{1}{2}\}$$

The transformations are

$$T_1 : z \mapsto z + 2 \qquad \text{and} \qquad T_2 : z \mapsto -\frac{1}{z}\ .$$

This is then a Fuchsian group preserving the upper and lower half-planes. It is a subgroup of the modular group we looked at earlier. Its limit set is the circle $\mathbb{R} \cup \{\infty\}$.

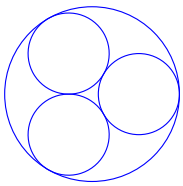


An even more degenerate case is when each disc touches the 3 others. This is illustrated below on the left. The group is generated by two parabolic transformations fixing points where two circles touch. The limit set is the *Apollonian gasket*. This is the pattern obtained as follows:

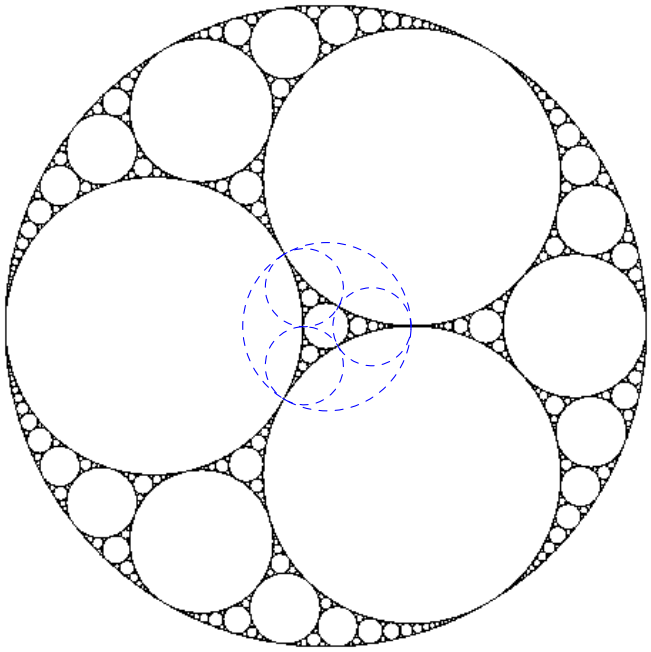
> Given any 3 mutually tangent circles, there are 2 further circles each of which touches each of the original 3. Start with 3 mutually tangent circles and use this process to add further circles recursively.

Exercise:

33. Show that we can choose 4 circles each pair of which touch so that they have tetrahedral symmetry. Show that the limit set, the Apollonian gasket also has tetrahedral symmetry.

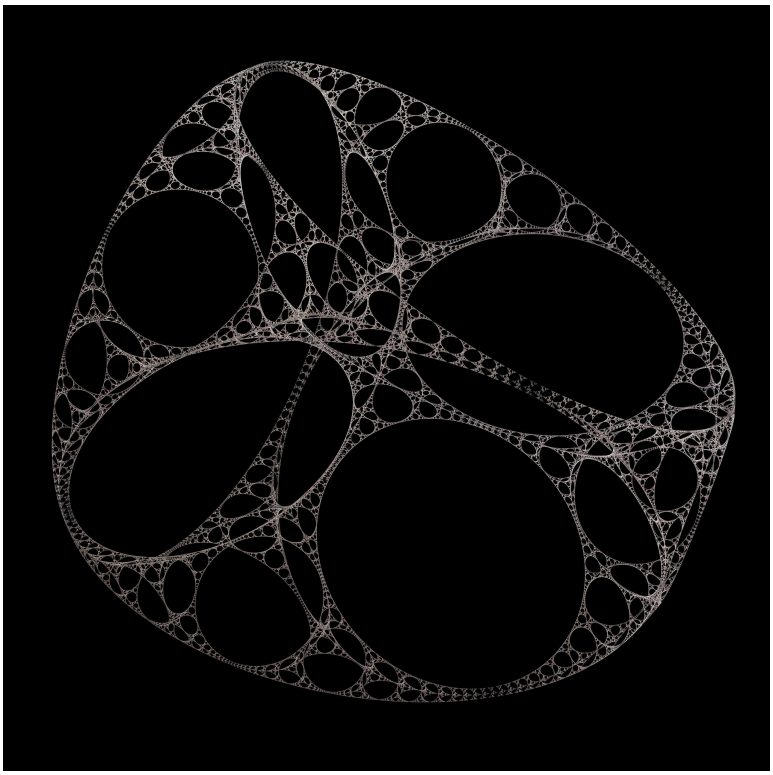


Four circles with each pair tangent.

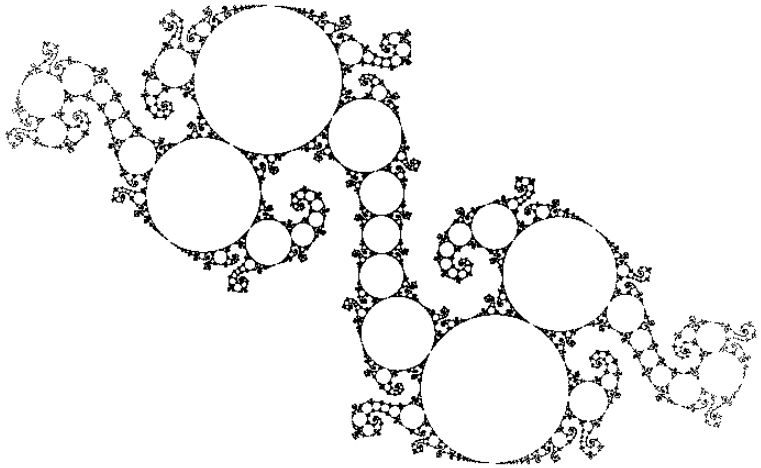


Limit set — The Apollonian Gasket

We ought really to draw this picture on the Riemann sphere, as illustrated below.



By allowing these sorts of degenration for Schottky groups we can create very complicated, fractal limit sets. An example is given below.



*24.2 Riemann Surfaces *

Let G be a group of Möbius transformations that acts discontinuously on the unit disc $\mathbb{D}$. Then the quotient $\mathbb{D}/G$ is an orientable surface. Each of the Möbius transformations in G is analytic and has an analytic inverse. Hence we can do complex analysis on the quotient $\mathbb{D}/G$. We call such a surface which looks locally like a piece of the complex plane a *Riemann surface*.

Similar arguments apply to groups acting discontinuously on the complex plane. For example, consider the group of translations $G = \{z \mapsto z + \lambda : \lambda \in \Lambda\}$. When $\Lambda = \{0\}$, the quotient $\mathbb{C}/G$ is just $\mathbb{C}$; when $\Lambda = \mathbb{Z}\omega_1$, the quotient $\mathbb{C}/G = \mathbb{C}/\mathbb{Z}\omega_1$ is a cylinder; when $G = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$ (with ω_1 and ω_2 linearly independent over $\mathbb{R}$), the quotient $\mathbb{C}/G = \mathbb{C}/(\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2)$ is a torus.

The Riemann Mapping Theorem (and the Uniformization Theorem) states that every Riemann surface is either the quotient of the unit disc $\mathbb{D}$ by a discrete group or else on of the special surfaces: $\mathbb{P}$, $\mathbb{C}$, a cylinder or a torus. This means that almost all Riemann surfaces are quotients of the unit disc. We can study them by studying the discrete groups of Möbius transformations that maps the unit disc to itself, that is, by studying Fuchsian groups.

Compact Riemann surfaces are compact orientable surfaces. It can be shown that these are homeomorphic to spheres with K handles for some $K \geqslant 0$. Hence every such compact orientable surface is the quotient of the Riemann sphere by a Schottky group. Indeed, more is true, by choosing the discs $\Delta_k^{\pm}$ appropriately, we can obtain every compact Riemann surface from a Schottky group, including the complex structure on the surface. This leads us to think about classifying all compact Riemann surfaces by using the Schottky groups. This is a very active area of research known as the study of Teichmüller space.

For further study I recommend the book *Indra's Pearls: The Vision of Felix Klein* by David Mumford, Caroline Series, and David Wright. Cambridge University Press, 2002 (ISBN 0-521-35253-3). This has much more information, beautifully presented.