1 Circles

1.1 Elementary Properties

Our ambient space will be the Euclidean plane $\mathbb{R}^2$. We will use the inner product (or dot product)

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} \cdot \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = x_1 x_2 + y_1 y_2$$

and the *norm* given by

$$|X|^2 = X \cdot X$$

to compute geometric quantities like distances and angles.

Definition 1.1 The *distance* between two points A and B is given by $|A - B|$.

Definition 1.2 Then *angle* $\phi \in [0, \pi)$ between two vectors v and w is given by

$$\cos(\phi) = \frac{v \cdot w}{|v||w|} \ .$$

The angle between two lines is the angle between their directions vectors. Depending on the choice of the direction vectors, there are two angles possible, namely ϕ and $\pi - \phi$.

In particular, two vectors v and w are *orthogonal* if $v \cdot w = 0$. We also write $v \perp w$ to indicate orthogonality.

Example 1.1 Let's see how to prove Pythagoras' theorem using these definitions. Let A and B be two edge vectors of a triangle with a vertex as their common foot point. Then the third edge has edge vector $A - B$, and we have $|A - B|^2 = |A|^2 + |B|^2$ if and only if $A \cdot B = 0$.

We begin our discussion of the circle with Thales' theorem and Apollonius characterization of circles.

Definition 1.3 A circle with center P and radius $r > 0$ is the set of points

$$C_r(P) = \{Q : |P - Q|^2 = r^2\} .$$

Two points $A, B \in C_r(P)$ form a *diameter* of the circle if A, B, and P are collinear.

This definition makes use of the distance between points, which can be computed using the inner product. It is an amusing question whether we really need to know distances to define what a circle is. As the dot product can also be used to define angles, we could for instance wonder whether we can define what a circle is in terms of just angles. This is indeed the case:

Theorem 1.1 (Thales) *Let A and B be two distinct points. Then a point P lies on the circle with diameter AB if and only if either $P = A$, $P = B$, or $\angle_P(A, B) = 90^\circ$.*

Proof: Choose the center of the circle as the origin 0 of our coordinate system so that $B = -A$. With $r^2 = |A|^2$ we have

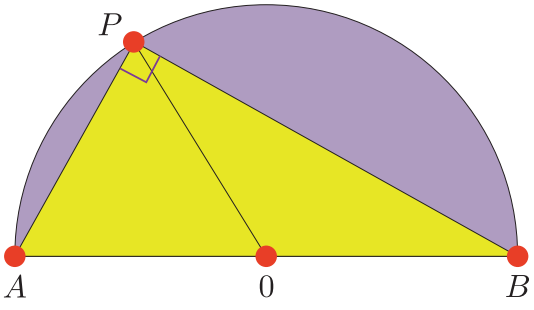


Figure 1 Theorem of Thales

$$|P|^2 = r^2 \Leftrightarrow (P-A)\bullet(P+A) = 0$$
$$\Leftrightarrow P = \pm A \quad \text{or} \quad P - A \perp P + A \ .$$

□

Another characterization of circles is due to Apollonius. Instead of knowing the distance to a single point, it suffices to know the distance *ratio* to two different points:

Theorem 1.2 *Let A and B be points, and $0 < \lambda < \infty$ be a real number. Then the set*

$$C_\lambda(A, B) = \{P : |PA| = \lambda |PB|\}$$

is a line for $\lambda = 1$ and a circle in general.

Proof: We leave the case $\lambda = 1$ as an exercise. Let C (resp. D) divide AB in the proportion λ (resp. $-\lambda$). Thus

$$C = \frac{1}{1+\lambda}(A + \lambda B) \quad \text{and} \quad D = \frac{1}{1-\lambda}(A - \lambda B) \ .$$

These points will serve as the diameter of the circle. Now take a point P with distance ratio λ to A and B, and make it the center of the coordinate system, so that $|A|^2 = \lambda^2 |B|^2$ by assumption. This is equivalent to $(A - \lambda B) \cdot (A + \lambda B) = 0$, or $C \cdot D = 0$. By Thales' theorem, this is equivalent to P being on a circle with diameter CD. □

Exercise 1.1 Show that the *bisector* $C_1(a, b)$ of a and b is a line.

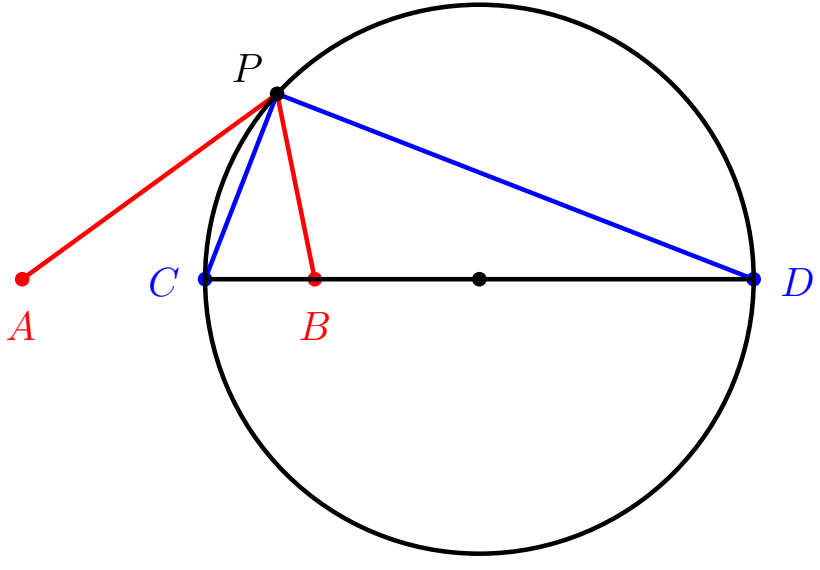


Figure 2 Apollonius circle

Solution: A point x has equal distance to a and b if and only if $(x-a)\cdot(x-a) = (x-b)\cdot(x-b)$. Expanding and simplifying gives

$$(b-a)\cdot x = \frac{1}{2}(b\cdot b - a\cdot a)$$

which is the equation of a line.

We can also define the angle of intersection between curves, provided they have tangent lines.

Definition 1.4 The tangent line to a circle through a point p is the unique line that intersects the circle only in p.

Proof: Let $v \neq 0$ be the direction vector of a line through a point p on a circle with center a. A point $p + tv$ on this line lies also on the circle if and only if

$$|p + tv - a|^2 = |p - a|^2 \ .$$

This is equivalent to $2t(p-a)\cdot v + t^2|v|^2 = 0$. This equation has only one solution in t if and only if $p - a$ is perpendicular to v. □

Definition 1.5 The angle of intersection of two circles at a point p is the angle of intersection of their tangent lines at p. Two circles are *tangent* at a point if they both contain p and have the same tangent line at p.

Example 1.2 Two circles with center A and B and radii r and s intersect orthogonally if and only if $|A - B|^2 = r^2 + s^2$. To see this, note that the radius to a point is perpendicular to the tangent line at that point, so two circles intersect at P orthogonally if and only if the radii AP and BP are orthogonal. Then the claim follows from Pythagoras.

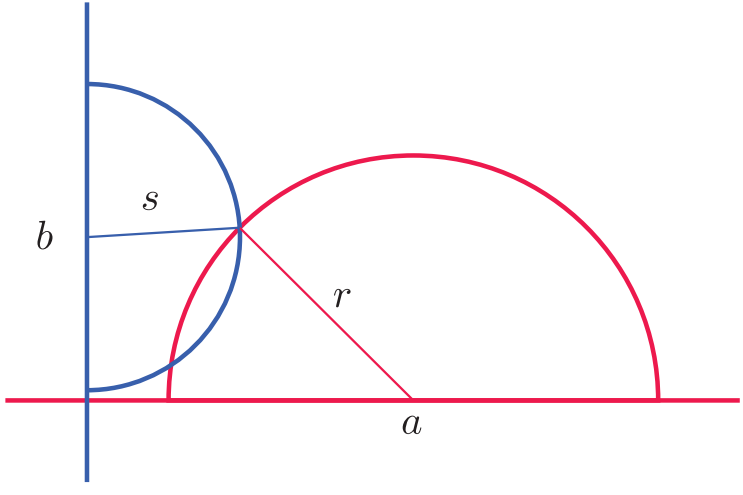


Figure 3 Never orthogonal

Exercise 1.2 Consider two circles as in **Figure 3** with centers $A = (a, 0)$ and $B = (0, b)$ and radii $r < a$ and $s < b$. Show that they do *not* intersect orthogonally.

Solution: Because

$$|A - B|^2 = |(a, -b)|^2 = a^2 + b^2 > r^2 + s^2$$

the condition from **Example 1.2** can never be satisfied.

Definition 1.6 A linear map A is called an *orthogonal map* if

$$Av \bullet Aw = v \bullet w$$

for all vectors $v, w \in \mathbb{R}^n$.
A linear map A is called a *similarity* if there is a positive number c such that

$$Av \bullet Aw = c \cdot v \bullet w$$

for all vectors $v, w \in \mathbb{R}^n$.

By definition, orthogonal maps preserve distances and angles between lines, while similarities preserve angles.

Lemma 1.1 *If two circles intersect at two different points, the angles of intersection at these two points are equal.*

Proof: Let l be a line through the centers of the two circles. The reflection at this line preserves both circles and exchanges the two intersections. As reflections preserve angles, the claim follows. □

Exercise 1.3 In **Figure 4**, you see a square with edge length one. Everything that seems to touch does touch, and all apparent concurrencies are true concurrencies. Determine the radius of the middle circle.

Exercise 1.4 In **Figure 5**, the large circle has radius 1 and the small circles have all the same radius. The small circles touch the large circle and their neighbors as shown. What is the radius of the small circles?

Solution: It helps to add a seventh circle at the center, touching the six small circles.

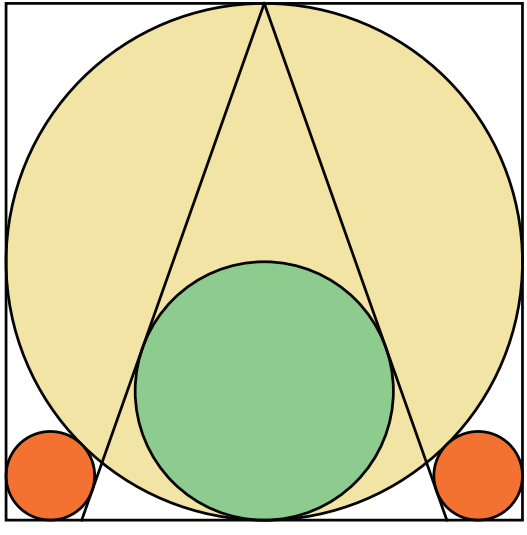


Figure 4 Circles and Square

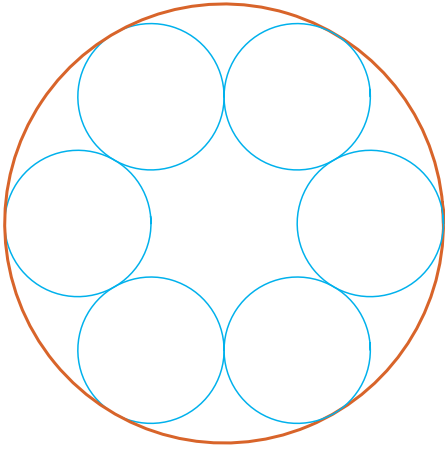


Figure 5 Circular chain

1.2 Inversion at the Circle

In this section, we will introduce a self-map of the plane that we can think of as a reflection at the circle:

Definition 1.7 The *inversion* at the circle $\mathbb{S} = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$ is defined as the map

$$I : \mathbb{R}^2 - \{(0,0)\} \to \mathbb{R}^2 - \{(0,0)\}$$

$$p \mapsto \frac{p}{|p|^2} .$$

By convention, we define $I(0) = \infty$ and $I(\infty) = 0$.

Thus the inversion fixes the circle $\mathbb{S}$ pointwise, and exchanges points inside with points outside of the circle. More precisely, a point p is mapped to a point $I(p)$ on the same ray through 0 and p so that $|p| \cdot |I(p)| = 1$.
It will be convenient to introduce a notion for the domain of the inversion:

Definition 1.8 The *extended plane* $\hat{\mathbb{E}}$ is the Euclidean plane $\mathbb{R}^2$ together with a point ∞. And *extended line* is a line together with ∞. A *generalized circle* is either a circle or an extended line.

Lemma 1.2 *Given three distinct points in $\hat{C}$, there is a unique generalized circle incident with the three points.*

Proof: If one of the points is ∞, the generalized circle is the (unique) line through the two other points. Similarly, if the three points are collinear, we take the line through them. Finally, let's assume that the three points a, b, and c are not collinear. Then the bisectors of ab and bc are not parallel and hence intersect in a unique point p that has equal distance to a, b, and c. □

The inversion has two fundamentally important properties that justify its introduction: it maps a generalized circle to a generalized circle, and it preserves the angles between curves.

Theorem 1.3 *The inversion maps a generalized circle to a generalized circle. More precisely, the inversion maps*

1. *lines through* 0 *to themselves,*
2. *lines not through 0 to circles through 0,*
3. *circles through 0 to lines not through 0,*
4. *circles not through 0 to circles not through 0.*

Proof: Using complex numbers $z = x + iy$, the inversion can be written as

$$I(z) = \frac{z}{|z|^2} = \frac{1}{\overline{z}} \, .$$

Suppose a circle (or line) is given as the set

$$C_\lambda(a, b) = \{z : |z - a| = \lambda |z - b|\}$$

using Apollonius' characterization of circles. Then we have

$$I(C_\lambda(a, b)) = C_{\lambda \frac{|b|}{|a|}}(1/\overline{a}, 1/\overline{b})$$

because

$$\left|\frac{1}{\overline{z}} - \frac{1}{\overline{a}}\right| = \frac{|z-a|}{|z\|a|}$$
$$= \lambda \frac{|z-b|}{|z\|a|}$$
$$= \lambda \frac{|b|}{|a|}\left|\frac{1}{\overline{z}} - \frac{1}{\overline{b}}\right| .$$

This proves the general claim that lines or circles are mapped to lines or circles. The more precise version follows by considering whether a circle or line passes through 0: If this is the case, the image must pass through infinity, and hence be a line, otherwise it must be a circle. □

Exercise 1.5 Centers of circles are not mapped to the centers of the image circles. For the circle with center $(a, 0)$ and radius $r < |a|$, find the center $(a', 0)$ and radius r' of the inverted circle. Show that never $a' = \frac{1}{a}$.

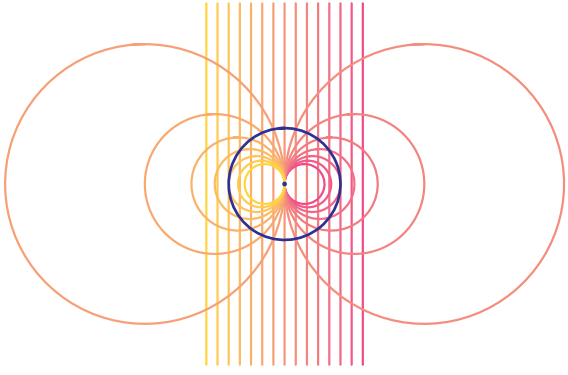


Figure 6 Parallel lines to touching circles

We note

Corollary 1.1 *For a given line λ through the origin, the inversion at the unit circle maps the family of lines parallel to λ to the family of generalized circles that touch λ at the origin.*

Proof: A line λ' parallel to λ but not equal to λ is inverted to a circle through 0. As λ and λ' have only ∞ in common, $I(\lambda)$ and $I(\lambda')$ have only 0 in common. Hence the circle $I(\lambda')$ touches the line $I(\lambda) = \lambda$ at 0. □

Much of the usefulness of the inversion at circles comes from the fact that we can often determine the image of a shape by geometric features instead of by computation. This is further being helped by the following theorem.

Theorem 1.4 *The inversion preserves the angle between any two curves.*

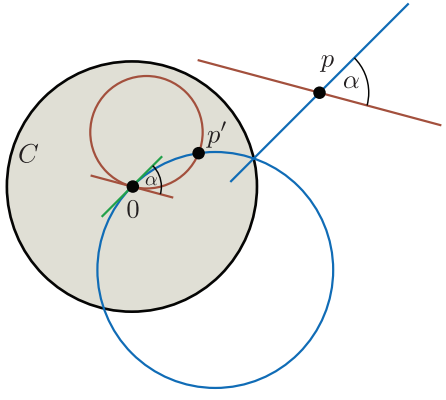


Figure 7 Inversion preserves angles

Proof: We will first give a geometric proof. Take two lines that intersect at some angle α at a point p. Invert them at a circle c with center 0. The result will be two generalized circles that intersect at 0 and a second point p'.
By **Corollary 1.1**, these circles intersect at 0 at the angle α, because they are tangent to lines that are parallel to the original lines and pass through 0.
Thus the two circles will also intersect at the other point where they meet at the angle α.

The second proof uses calculus. Let $x(t)$ be a curve. Then its tangent vector at $x(t)$ is $x'(t)$.
We compute the tangent vector to $I(x(t))$:

$$\begin{aligned}\frac{d}{dt}I(x(t)) &= \frac{d}{dt}\frac{x}{x\bullet x}\\ &= \frac{x'}{x\bullet x} - 2\frac{(x'\bullet x)}{(x\bullet x)^2}x\ .\end{aligned}$$

Thus for two curves $x(t)$ and $y(t)$ that both pass through the same point p at $t=0$, we compute at $t=0$

$$\begin{aligned}\frac{d}{dt}I(x(t))\bullet\frac{d}{dt}I(y(t)) &= \left(\frac{x'}{x\bullet x} - 2\frac{(x'\bullet x)}{(x\bullet x)^2}x\right)\bullet\left(\frac{y'}{y\bullet y} - 2\frac{(y'\bullet y)}{(y\bullet y)^2}y\right)\\ &= \left(\frac{x'}{p\bullet p} - 2\frac{(x'\bullet p)}{(p\bullet p)^2}p\right)\bullet\left(\frac{y'}{p\bullet p} - 2\frac{(y'\bullet p)}{(p\bullet p)^2}p\right)\\ &= \frac{x'\bullet y'}{(p\bullet p)^2}\ ,\end{aligned}$$

as the other terms cancel.
Therefore

$$\begin{aligned}\cos\angle\left(\frac{d}{dt}I(x(t)),\frac{d}{dt}I(y(t))\right) &= \frac{\frac{x'\bullet y'}{(p\bullet p)(p\bullet p)}}{\frac{|x'|}{|p|^2}\frac{|y'|}{|p|^2}}\\ &= \frac{x'\bullet y'}{|x'\|y'|}\\ &= \cos\angle(x',y')\ .\end{aligned}$$

□

Instead of inverting at the unit circle $\mathbb{S}^2$, we can use any circle as the 'mirror' for an inversion. To do so, we translate everything so that the mirror has its center at the origin, scale so that it has radius 1, apply the standard inversion at the unit circle, scale back to original size, and translate back:

Definition 1.9 The inversion at an arbitrary circle $\mathbb{S}^1_r(p)$ is defined

$$I_{p,r}(q) = p + r^2 I(q-p)\ .$$

Geometrically, this means that a point q is inverted to a point q' on the same ray pq so that the product of the distances $|pq|$ and $|pq'|$ is equal to r^2.

Clearly, this more general inversion will still map generalized circles to generalized circles, and preserve angles.
Sometimes it is useful to apply inversions to transform several circles at once into a convenient position. For instance, inverting two circles at a mirror circle whose center is at an intersection of the two circles results in two lines that either intersect, or are parallel (in case the two circles touch). If two circles are disjoint, we have the following:

Theorem 1.5 *Two disjoint circles can be mapped by inversions to two concentric circles.*

Proof: We begin with a motivation. If we have two concentric circles, an inversion will map their shared diameters to circles that are orthogonal to both circles and pass through two common points, namely the images of the center of the circles and of ∞. To prove the theorem, it hence suffices to find two intersecting generalized circles that are orthogonal to both of the two given circles.
The first of these orthogonal circles is easy to find: We can assume that the two given circles C_1 and C_2 are not already concentric, and take the line D through their centers. Now we apply a first inversion at a circle C whose center P lies on C_1 but not D. The images of C_1 and C_2 become a line C_1' and a circle C_2', disjoint, a met orthogonally by the image circle D' of the line D.

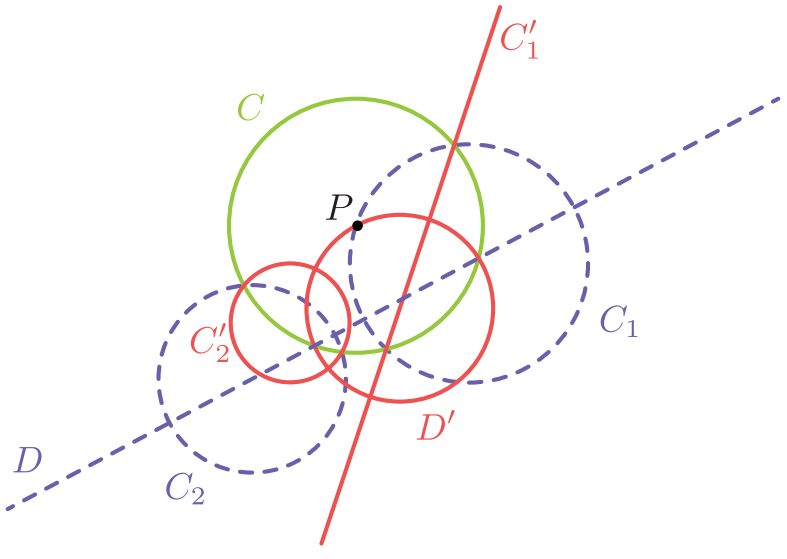


Figure 8 First inversion

Now it is easy to find a second generalized circle orthogonal to C_1' and C_2', namely the line E through the center of C_2' and orthogonal to C_1'. We will show below that D' and E intersect in a point Q. Applying a second inversion at a circle C' centered at Q will map C_1' and C_2' to circles ${C_1}''$ and ${C_2}''$. The circle D' becomes

a line D'', and E is mapped to itself. Both D'' and E are lines simultaneously orthogonal to ${C_1}''$ and ${C_2}''$, so these two circles are finally concentric as desired.

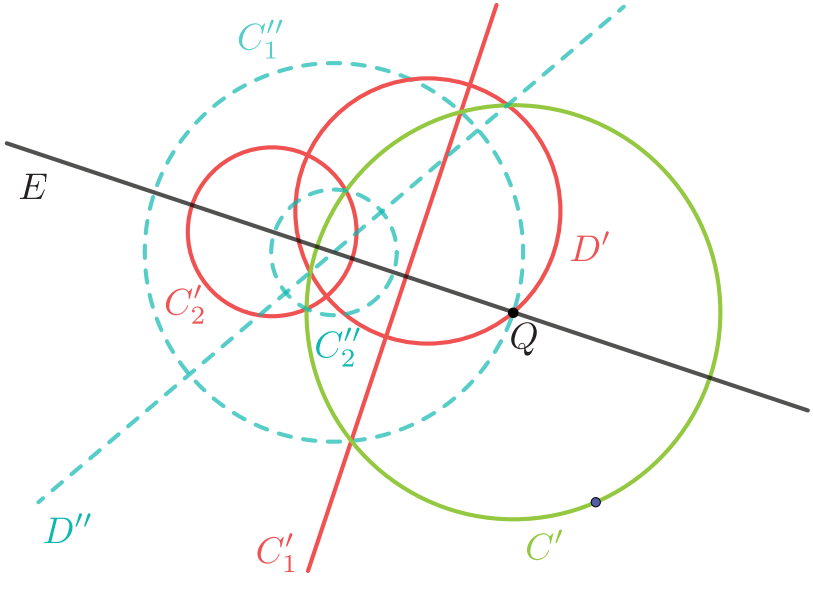


Figure 9 Second inversion

We still have to show that D' and E intersect. If this were not the case, D' would lie on one side of E, and we would be in the situation of **Exercise 1.2** that shows that C_2' and D' could not be orthogonal, contradicting our construction. □

The proof above shows that we can make two disjoint circles concentric using two consecutive inversions. We now give a computational proof that this can also be done with a single inversion. After a translation and rotation, we can assume that the centers of the circles are on the real axis, say at $(a+t, 0)$ and $(b+t, 0)$, and that their radii r and s satisfy either $b - a > r + s$ (when the circles are separated) or $a + r > b + s$ and $a < b$ (when the circles are nested). Here t is a free parameter that shifts both circles along the real axis. After applying an inversion at the unit circle $\mathbb{S}^1$, the centers of the circles are at

$$\frac{a+t}{(a+t)^2 - r^2} \quad \text{and} \quad \frac{b+t}{(b+t)^2 - s^2} \ .$$

By taking reciprocals, we see they are equal if and only if

$$f(t) = a - b + \frac{s^2}{b+t} - \frac{r^2}{a+t} \ .$$

We will show that there is a value of t so that $f(t) = 0$. In the first case that $b - a > r + s$, we evaluate f at a local minimum:

$$f\left(-\frac{as+br}{r+s}\right) = \frac{(a-b)^2-(r+s)^2}{a-b} < 0 \ .$$

As a weighted average, we have

$$-b < -\frac{as+br}{r+s} < -a \ .$$

On the other hand,

$$\lim_{t\to -b+} f(t) = +\infty = \lim_{t\to -a-} f(t) \ ,$$

so that f has at least two zeroes in the interval $(-b,-a)$. Either one will translate the circles so that their inverted images are concentric. In the second case, we evaluate f at another local minimum:

$$f\left(\frac{as-br}{r-s}\right) = \frac{(a-b)^2-(r-s)^2}{a-b} > 0 \ .$$

Now

$$\frac{as-br}{r-s} < -b \ ,$$

and

$$\lim_{t\to -b-} f(t) = -\infty \quad \text{and} \quad \lim_{t\to -\infty} f(t) = a - b < 0 \ ,$$

so that f has at least two zeroes for $t < -b$.

Example 1.3 Two circles both have radius 2 and their centers are distance 5 apart. Find inversions that make them concentric, and compute the ratio of the radii of the concentric image circles.

This problem does not appear to be well defined, because the answer could depend on how we arrive at the two concentric circles. We will see later when we talk about cross ratios that the answer is independent of our solution.

After a translation and rotation, we can assume that the two circles have centers at $(t, 0)$ and $(t + 5, 0)$. Then the function $f(t)$ in the above proof becomes

$$f(t) = \frac{4}{t+5} - \frac{4}{t} - 5$$

and we find the two solutions $t = -4$ and $t = -1$. For $t = -1$, inverting the two circles at the unit circle gives two circles with center at $1/3$ and radii $2/3$ and $1/6$, so their ratio is 4.

1.3 Applications of the Inversion

In many cases these properties are powerful enough so that one can argue with them instead of doing computations. Here are three examples:

Theorem 1.6 *Given three disjoint disks, there are two unique circles that touch the three disks from the inside and from the outside.*

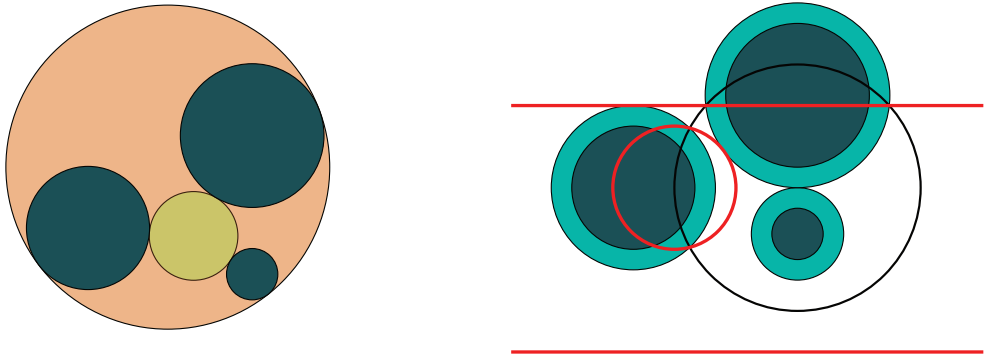


Figure 10 Three Circle Problem: Setting up the Inversion

Proof: Add a positive constant to each radius so that (at least) two of the three circles touch at a point p. We first find a solution for the new circles. Invert them at a circle with center p. This will take the two touching circles to parallel lines so that the region between the lines corresponds to the outside of the two disks. The third disk will be mapped to a disk inside that parallel strip.

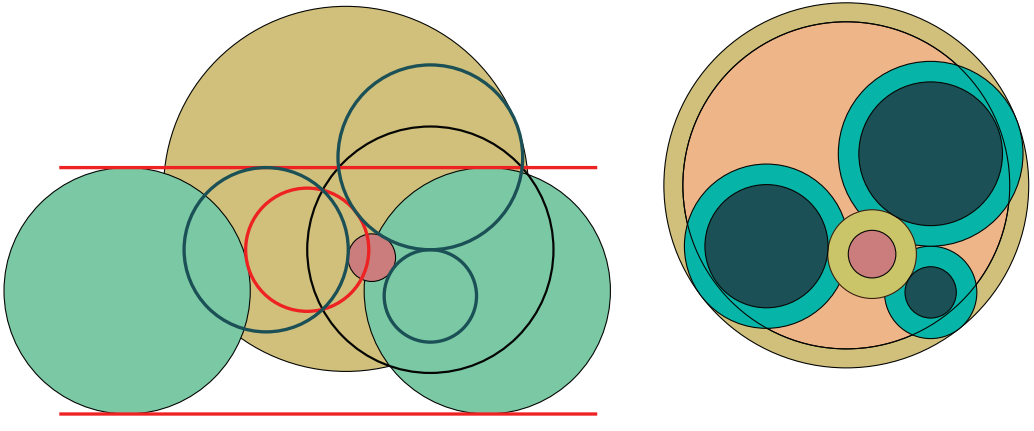


Figure 11 Three Circle Problem: Applying the Inversion

Again it will be sufficient to solve the problem for the two parallel lines and the circle. This is easy, because we can just slide two circles inside the strip until

they touch the third new disk. Applying an inversion at the mirror circle gives a solution to the modified problem. In the final step, we increase (resp. decrease) the radii of the two circles we found by the same amount by which we increased the radii in the first step. □

The second example is a theorem of Jakob Steiner.

Theorem 1.7 (Steiner) *Consider two disjoint circles, one contained within the other. Place a circle so that it touches the outer circle from the inside and the inner circle from the outside. Continue placing circles like this so that each new circle also touches the previously placed circle. If this chain of circles closes up in the sense that the n^{th} circle also touches the first circle, then this happens for any choice of circle that begins the chain.*

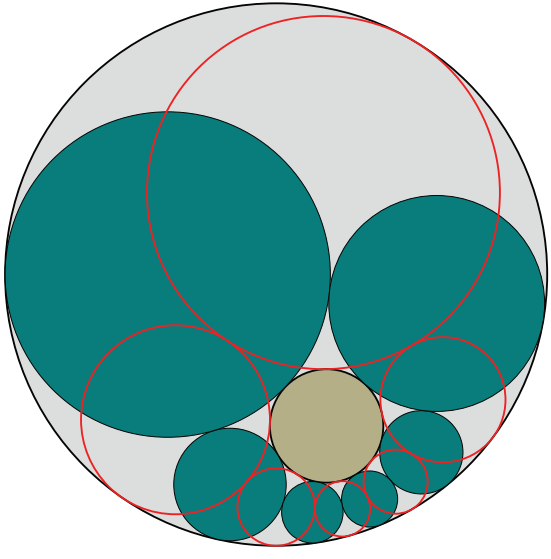
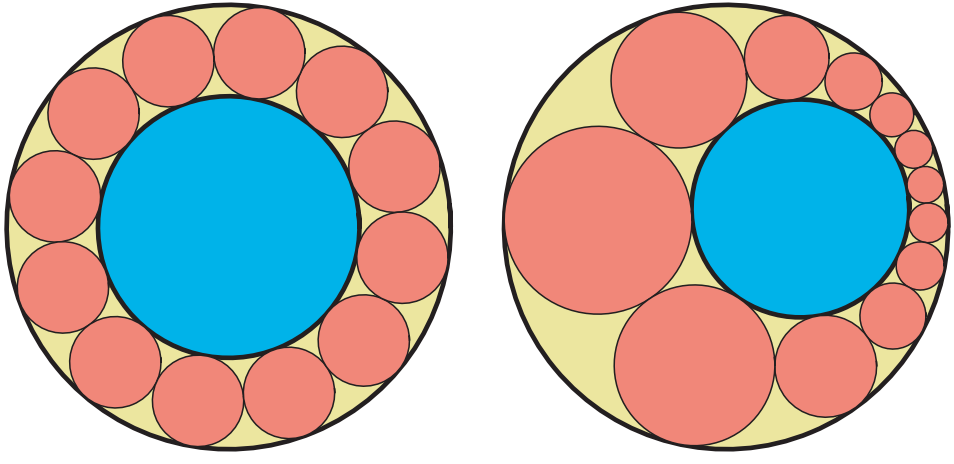


Figure 12 Chains of touching circles



Trivial Steiner Chain

Image under Möbius transformation

Figure 13 Proof by Inversion

Proof: Apply a Möbius transformation to the two given circles that maps them to two concentric circles. For these, the claim is trivial, because any two choices of chains are mere rotations of each other. But the Möbius transformation maps circles to circles and preserves tangencies. □

Exercise 1.6 Show that two circles with centers at $p, q \in \mathbb{R}^2$ and radii $r, s > 0$ intersect at an angle ϕ with

$$\cos\phi = \frac{r^2 + s^2 - |p-q|^2}{2rs}$$

This angle of intersection is defined as the angle between the vectors $x - p$ and $x - q$ for a point x that lies on both circles. Hint: Expand $|(x-p) - (x-q)|^2$.

The third example is about a shape called the *arbelos*, greek for 'shoemaker's knife'. It is bounded by three half circles with diameters $[0, 1]$, $[0, a]$ and $[a, 1]$ on the real axis as shown in **Figure 14**.
We consecutively add circles to the arbelos. The first touches all three half circles, and each next subsequent circle touches the previous circle and the two larger of the half circles.
Now apply an inversion to this configuration at a circle with center at 0. This will map the two larger half circles to parallel vertical lines. We can choose the radius of the mirror circle so that one (say the n^{th}) of the added circle is intersected orthogonally. This results in this circle being mapped to itself by the inversion. All added circles will be mapped to a chain of vertical circles between the two parallel lines, and hence have all the same diameter d as the n^{th} added circle. This shows:

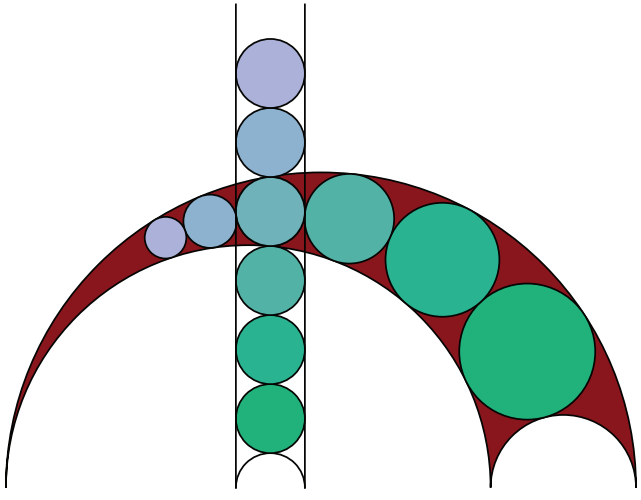


Figure 14 The Arbelos

Theorem 1.8 *Suppose the n^{th} added circle of the arbelos has diameter d. Then its center has as y-coordinate $y = d \cdot n$.*

2 Spherical Geometry

2.1 Background: Distances

In Euclidean vector spaces, the dot product can be used to define distances and angles.

Definition 2.1 The *length* or *norm* of a vector v is defined by

$$|v|^2 = v \bullet v \ .$$

The *distance* $\mathbf{d}(p, q)$ between to points p and q is the length of the vector $p - q$:

$$\mathbf{d}(p, q) = |p - q| \ .$$

Norm and dot product satisfy the important *Cauchy-Schwarz inequality*, on which virtually every geometric inequality is ultimately based:

Theorem 2.1 (Cauchy-Schwarz inequality) *For all $v, w \in \mathbb{R}^n$, we have*

$$|v \bullet w| \leq |v|\,|w|$$

with equality if and only if v and w are linearly dependent.

Proof: Write $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$ and $w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$. Then

$$\begin{pmatrix} v_1 & v_2 \\ w_1 & w_2 \end{pmatrix} \bullet \begin{pmatrix} v_1 & v_2 \\ w_1 & w_2 \end{pmatrix}^{\top} = \begin{pmatrix} |v|^2 & v \bullet w \\ v \bullet w & |w|^2 \end{pmatrix} .$$

Taking determinants and obeying applicable laws shows

$$0 \leq \det \begin{pmatrix} v_1 & v_2 \\ w_1 & w_2 \end{pmatrix}^2 = |v|^2|w|^2 - (v \bullet w)^2 .$$

This implies the Cauchy-Schwarz inequality. In case of equality, the determinant on the left hand side is zero, meaning that v and w are linearly dependent. □

The Euclidean distance satisfies the properties of a *metric*:

Definition 2.2 Let X be a non-empty set. A function $d : X \times X \to \mathbb{R}$ is called a *metric* if

1. (Symmetry) For all $x, y \in X$ we have $d(x, y) = d(y, x)$.
2. (Positivity) For all $x \in X$ we have $d(x, x) \geq 0$ with equality if and only if $x = y$.
3. (Triangle Inequality) For all $x, y, z \in X$ we have $d(x, z) \leq d(x, y) + d(y, z)$.

The set X together with a metric d is called a *metric space*.

Theorem 2.2 *The Euclidean space* $(\mathbb{R}^n, \mathbf{d}(p, q))$ *is a metric space.*

Proof: The only non-trivial thing to prove is the triangle inequality, which follows from the Cauchy-Schwarz inequality, because

$$|v + w|^2 = |v|^2 + 2v \cdot w + |w|^2 \leq (|v| + |w|)^2 \ .$$

□

This proof of the Euclidean triangle inequality via norm, dot product, and Cauchy-Schwarz inequality is a good example how the introduction and usage of concepts can reduce the amount of computation.

Definition 2.3 Let (X, d_X) and (Y, d_Y) be metric spaces. A bijective map $f : X \to Y$ is called an *isometry* if

$$d_Y(f(p), f(q)) = d_X(p, q)$$

for all points $p, q \in X$.

Example 2.1 Let A be an orthogonal $n \times n$ matrix, and $b \in \mathbb{R}^n$ a vector. Then the map $f(p) = Ap + b$ is an isometry from $\mathbb{R}^n$ to itself. In Euclidean geometry, an isometry is also called a *congruence*.

2.2 The Sphere

In Euclidean geometry, the two basic shapes are lines and circles, and everything else is based upon them. This is for the simple reason that the only accurate tool for measurement and construction that was available in ancient times was the rope, serving both as a ruler and compass. In general metric spaces, it might be difficult to define what a line is, but spheres are easy to define:

Definition 2.4 In a metric space X, the *sphere* with center p and radius $r > 0$ is the set $S_r(p) = \{x \in X : |x - p| = r\}$. Similarly, the *ball* of radius r is the (open) set $B_r(p) = \{x \in X : |x - p| < r\}$.

All spheres are metric spaces in their own right, but (at least) in Euclidean spaces, the spheres have a better metric that measures the distance between points along the surface of the sphere. To make this concrete, let's consider first the unit circle $\mathbb{S}^1 = S_1^1(0) \subset \mathbb{R}^2$. Let $p, q \in \mathbb{S}^1$ be two points. Their distance along the circle is just the length of the (shorter) arc between the two points, and this is the angle α between the vectors p and q. From Linear Algebra we know that this angle is given by $\cos(\alpha) = p \cdot q$. Now, more generally, consider the unit sphere $\mathbb{S}^n = S_1^n(0) \subset \mathbb{R}^{n+1}$. Let $p, q \in \mathbb{S}^n$ be two points. The 2-dimensional plane through the origin and p and q

cuts the sphere in a *great circle*. To measure the distance between p and q along that great circle we argue as in the case of S^1, the angle is again given by $\arccos(p \bullet q)$.

Definition 2.5 Thus the spherical distance between p and q in $\mathbb{S}^n$ is defined as

$$d(p,q) = \arccos(p \bullet q) \ ,$$

where the inverse of the cos-function is chosen to have its value in $[0, \pi]$.

Lemma 2.1 *The spherical distance is invariant under all orthogonal transformations.*

The spherical distance has the characteristic property that it is additive along a great circle: If we have three points a, b, c on the same great half circle occurring in that order, than $d(a,c) = d(a,b) + d(b,c)$. This is not the case between the usual Euclidean distance between points on a sphere, for which the triangle inequality is always strict.

Theorem 2.3 *The spherical distance is a metric on $\mathbb{S}^n$.*

Proof: We only need to verify the triangle inequality. As three points of $\mathbb{S}^n$ will always lie in a $\mathbb{S}^2 \subset \mathbb{S}^n$, we will assume that $n = 2$. Moreover, after applying a rotation, we can assume that a and c lie also on the half plane $x_1 = 0$ and $x_2 > 0$ so that $a_1 = c_1 = 0$ and $a_2, c_2 > 0$, and that b lies in the right hemisphere, i.e. that $b_1 > 0$. Let p be a point with $p_1 = 0$ and $p_3 = b_3$. Note that this implies $p_2 \geq b_2$. Thus

$$a \cdot b = a \cdot p + a \cdot (b - p) = a \cdot p + a_2(b_2 - p_2) \leq a \cdot p \ .$$

By the monotonicity of $\cos(x)$, we obtain $d(a, b) \geq d(a, p)$, and similarly $d(c, b) \geq d(c, p)$. By the additivity of the spherical distance along great circles we get

$$d(a, b) + d(b, c) \geq d(a, p) + d(c, p) = d(a, c)$$

as claimed. □

Thus great circles naturally take over the role of lines, and we obtain an incidence geometry on each Euclidean sphere.

We will mostly be concerned here with the 2- and 3-dimensional spheres $\mathbb{S}^2$ and $\mathbb{S}^3$. It will turn out the the group of rotations of $\mathbb{S}^2$ is as a space in itself closely related to $\mathbb{S}^3$.
In the next section we will need an elementary property of spheres:
The sphere $S_r(p)$ is invariant under the point reflection $x \mapsto 2p - x$ at its center. This characterizes the center:

Lemma 2.2 *If a sphere is invariant under a point reflection $x \mapsto 2p - x$, then p is the center of the sphere.*

Proof: It suffices to show that if a sphere is invariant under both $x \mapsto 2p - x$ and $x \mapsto 2q - x$, then $p = q$. To see this, we compose the two reflections and obtain $x \mapsto 2q - (2p - x) = 2(q - p) + x$. This maps is a translation that leaves the sphere still invariant. This can only be if $p = q$. □

Analogously to the inversion at a circle one can define the inversion at spheres:

Definition 2.6 Let $x \neq 0$ be a point in $\mathbb{R}^n$. The inversion at the unit sphere $\mathbb{S}^n$ is defined as

$$I : \mathbb{R}^n - 0 \to \mathbb{R}^n - 0$$

$$x \mapsto I(x) = \frac{x}{|x|^2} \ .$$

We think of 0 being mapped to a single point ∞.

The main properties of the inversion follow from the 2-dimensional case:

Theorem 2.4 *The inversion is an angle preserving involution. It maps planes through* 0 *to itself, and all other planes to spheres through* 0*. Correspondingly, spheres through* 0 *are mapped to planes through* 0*, and all other spheres are mapped to spheres again.*

Soddy's *Hexlet* is a chain of six spheres constructed as follows: Start by choosing three mutually touching spheres. In **Figure 15**, one of them is touching the others

externally. Begin the chain with a sphere touching all three spheres. Then consecutively add spheres to the chain so that they touch all three initial spheres and the previously added sphere.

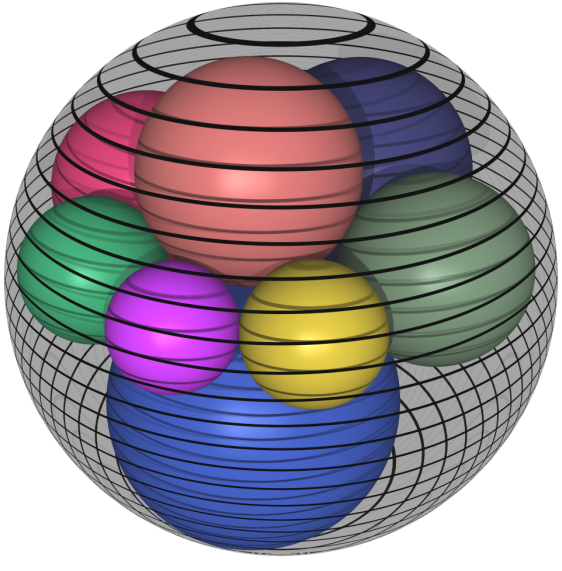


Figure 15 Soddy's Hexlet

Theorem 2.5 *The chain of spheres always closes after six spheres.*

Proof: Invert the sphere configuration at a sphere centered at the touching point of two of the initial spheres. These will be mapped to parallel planes, while the third initial sphere becomes a sphere between these two planes, touching them both. The chain of spheres will become a chain of spheres surrounding this sphere and of equal size. As this inverted chain clearly closes after six spheres, so does the original chain. □

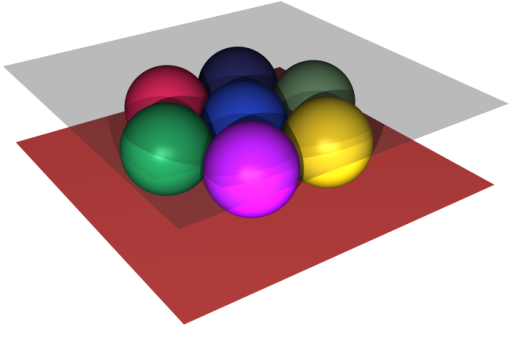


Figure 16 Inverted Soddy's Hexlet

2.3 The Stereographic Projection

The stereographic projection establishes a bijection between $\mathbb{S}^n$ and the extended Euclidean space $\hat{\mathbb{R}}^n$ that is both algebraically simple and geometrically useful. We explain it for $n = 2$, the general case is completely analogous.

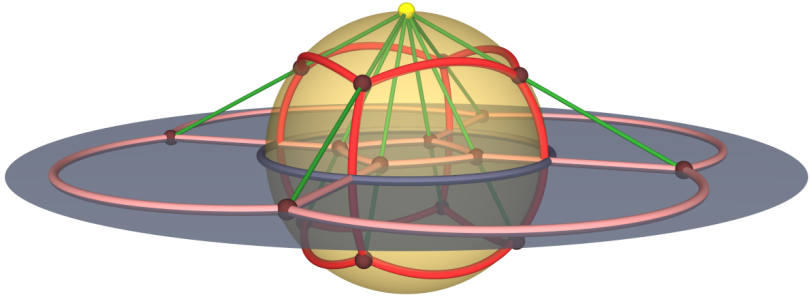


Figure 17 The Stereographic Projection

Let $P_0 = (0, 0, 1)$ be the north pole of $\mathbb{S}^2$. Draw a straight line through a given point $P \in S^2 - P_0$ and the point P_0. The line is not horizontal and thus intersects the coordinate (x, y)-plane in one point, which we denote by $p = \sigma(P)$. This map is called the *stereographic projection*. Here is an explicit formula.

Lemma 2.3 *The stereographic projection is given by*

$$\sigma(x, y, z) = \frac{1}{1-z}(x, y)$$

Proof: The line segment from $P_0 = (0, 0, +1)$ to $P = (x, y, z)$ is given by

$$s \mapsto sP + (1-s)P_0$$

This segment intersects the plane $z = 0$ for the value of s for which

$$sz + 1 - s = 0$$

that is, for

$$s = \frac{1}{1-z}$$

The claim follows. □

Occasionally it is useful to also have the inverse map:

Lemma 2.4 *The inverse map of the stereographic projection is given by*

$$s : (u, v) \mapsto \frac{2}{1 + u^2 + v^2} \left(u, v, \frac{1}{2}(u^2 + v^2 - 1) \right)$$

Proof: Let $(u, v) = \sigma(x, y, z)$. Then, using $x^2 + y^2 + z^2 = 1$,

$$\begin{aligned} u^2 + v^2 &= \frac{x^2 + y^2}{(1 - z)^2} \\ &= \frac{1 - z^2}{(1 - z)^2} \\ &= \frac{1 + z}{1 - z} \end{aligned}$$

so that

$$z = \frac{u^2 + v^2 - 1}{u^2 + v^2 + 1}$$

Therefore

$$x = (1-z)u = \frac{2}{1+u^2+v^2}u$$
$$y = (1-z)v = \frac{2}{1+u^2+v^2}v$$

□

One can use the stereographic projection to find points on $\mathbb{S}^2$ with rational coordinates. For instance, we have

$$s(1,1) = \frac{2}{3}(1,1,\frac{1}{2}) = \frac{1}{3}(2,2,1) \ .$$

Exercise 2.1 Find all points on $\mathbb{S}^2$ that have positive rational numbers as coordinates whose denominator does not exceed 7, up to permutations of the coordinates.

Exercise 2.2 Find all points on $\mathbb{S}^3$ that have positive rational numbers as coordinates whose denominator does not exceed 7, up to permutations of the coordinates.

More generally, the stereographic projection from $\mathbb{S}^n \subset \mathbb{R}^{n+1} \to \mathbb{R}^n$ is given by

$$\sigma : S^n \to \mathbb{R}^n$$

$$(p_1, ...p_{n+1}) \mapsto \frac{1}{1 - p_{n+1}}(p_1, ...p_n)$$

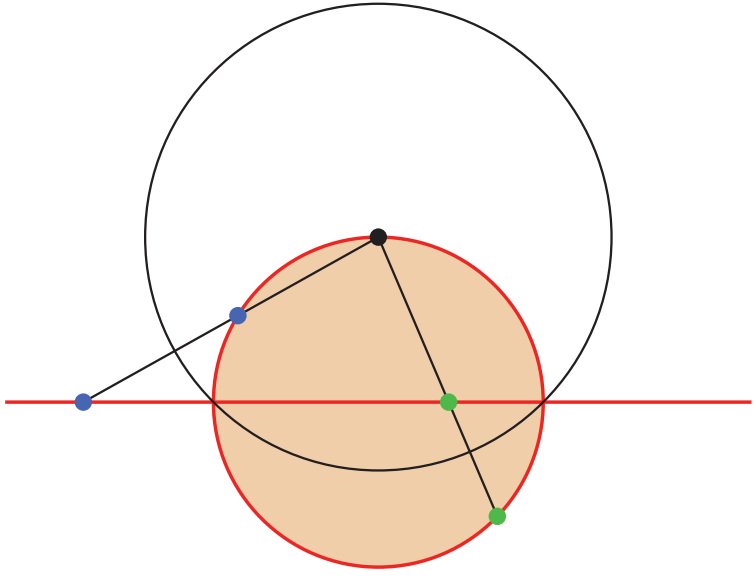


Figure 18 Stereographic projection and inversion

The geometric properties of the stereographic projection can be proven by interpreting it as an inversion at a sphere:

Lemma 2.5 *The stereographic projection of $\mathbb{S}^n$ is the inversion at the sphere which has its center at the north pole of $\mathbb{S}^n$ and passes through the equator of $\mathbb{S}^n$.*

Proof: The image of $\mathbb{S}^n$ under this inversion is the equator hyperplane of $\mathbb{S}^n$, and two points that are mirror images of each other must lie on the same radius through the center of the inversion sphere. □

This implies that the stereographic projection inherits the two main features of the inversion.

Corollary 2.1 *The stereographic projection maps circles, spheres, hyperspheres of $\mathbb{S}^n$ to circles, spheres, hyperspheres or lines, planes, hyperplanes of $\mathbb{R}^n$.*

2.4 The Area of Spherical Polygons

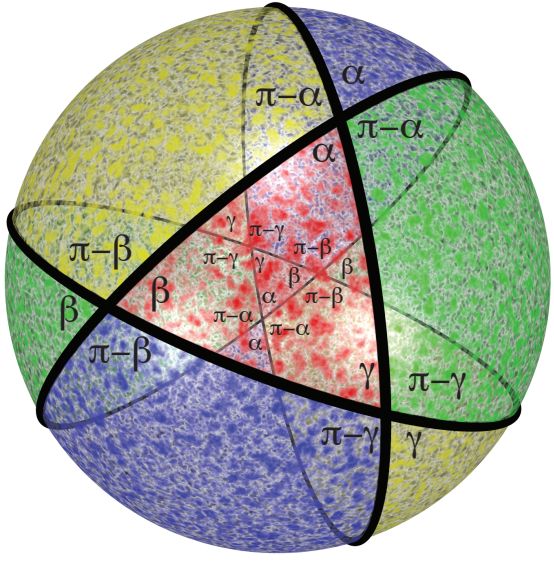


Figure 19 The Area of spherical triangles

A *spherical polygon* is a polygon, drawn on the sphere, whose edges are parts of great circles. On the sphere, these great circles play the role of straight lines in the

plane. They can be obtained by intersecting the sphere with a plane through the center of the sphere.
The *angle* of a spherical polygon at a vertex is the intersection angle of the two planes defining the great circles that meet at the vertex.
One unusual feature of the sphere is that there are 2-gons (or *bigons*): They always have a pair of antipodal points as vertices, and the angles at the two vertices are equal. What is the area of a bigon $B(\alpha)$ with angle α? Let's assume that $\alpha = \frac{\pi}{2}$. Then the bigon fills out one quarter of the sphere, so its area is $\frac{1}{4} \cdot 4\pi$. Similarly, if $\alpha = \frac{1}{3}\pi$, it fills out one sixth of the sphere, and its area is $\frac{1}{6} \cdot 4\pi$. As the area of disjoint domains is additive, the area of bigons is proportional to the angle. In general, the area of a bigon is therefore given by

$$\text{area}(B(\alpha)) = 2\alpha \ .$$

Surprisingly, this can be used to also find the area of triangles. Take a spherical triangle with angles α, β, γ, and draw the planes that define its edges. These planes divide the sphere into eight triangles.
They come in congruent antipodal pairs. We color the pairs red, green, blue, and yellow, and denote their areas by R, G, B and Y. Two adjacent triangles fit together to a bigon, and by using the notation introduced in **Figure 19**, we deduce

$$R + B = 2\alpha$$
$$R + G = 2\beta$$
$$R + Y = 2\gamma$$

Adding these equations gives

$$3R + B + G + Y = 2(\alpha + \beta + \gamma)$$

On the other hand,

$$R + B + G + Y = 2\pi$$

as the four triangles of one color each cover exactly a half sphere. Subtracting and dividing by two gives

$$R = \pi - \alpha - \beta - \gamma$$

We have proven

Proposition 2.1 *The area of a spherical triangle T with angles α, β, γ is*

$$\operatorname{area} T = \alpha + \beta + \gamma - \pi$$

This has as a consequence that the sum of the angles of a spherical triangle is always greater than π.
More generally, by subdivision we obtain a formula for the area of spherical polygons:

Theorem 2.6 *Let P be a spherical n-gon with interior angles $\alpha_1, ..., \alpha_n$. Then*

$$\operatorname{area} P = \sum_{i=1}^{n} \alpha_i - (n-2) \cdot \pi$$

Proof: If the formula is correct for two polygons P_1 and P_2 that share an edge, the formula also holds for the polygon $P = P_1 \cup P_2$. To see this, note that the sum interior angles of P is just the sum of the interior angles of P_1 and P_2 together, and the number of vertices of P is the sum of the vertices of P_1 and P_2 minus 2, because the two polygons share an edge.
As we know, the formula holds for triangles. To prove the general formula, we only need to divide the polygon under consideration into spherical triangles. That this is possible is not completely obvious but true. □

Definition 2.7 A regular spherical n-gon is a spherical n-gon that is invariant under a rotation around a line in space through the origin by angle $2\pi/n$.

As a consequence, all its angles are equal.
Another non-Euclidean feature of the sphere is that there are no homotheties: Regular n-gons of different size have different angles. We investigate this for a regular pentagon.
If the pentagon is very small, it doesn't notice very much the curved nature of the sphere, and its interior angles are close to the interior angles of a Euclidean regular pentagon, which are $\pi - 2\pi/5 = 3\pi/5$. If the pentagon gets larger, its approaches a hemisphere, so eventually its interior angles come close to π. Therefore, there is a regular spherical pentagon with interior angles equal to $2\pi/3$ which is strictly between the limits $3\pi/5$ and π. This pentagon serves as a nice tile on the sphere, because we now can fit three tiles around a vertex.
Can we possibly tile the whole sphere with copies of this pentagon? If so, how many tiles do we need? By our area formula, the single tile has area

$$\operatorname{area} T = 5 \cdot 2\frac{\pi}{3} - 3\pi = \frac{\pi}{3}$$

which is a twelfth of the area of the sphere. Therefore, we would need exactly 12 tiles. This lets us guess how the tiling will look like: Like a dodecahedron tiling!

2.5 Platonic Tilings of the 2-dimensional Sphere

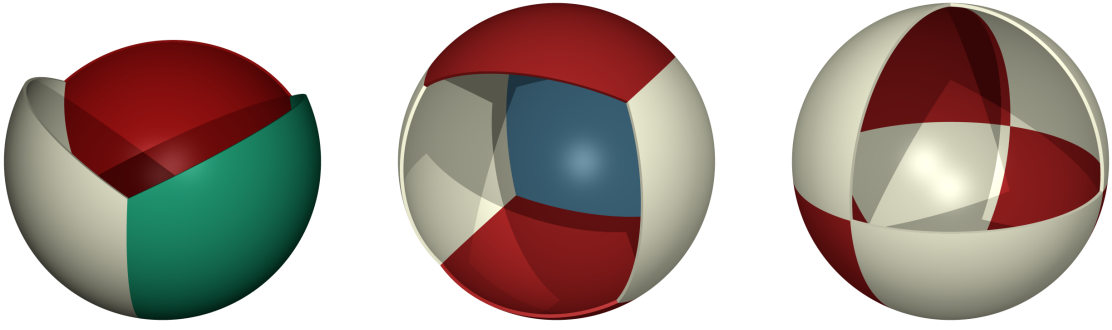
In this section, we will determine all regular spherical polygons that tile the sphere. The first question we need to address is which angles are possible for regular spherical polygons. We can construct such n-gons by first subdividing the equator great circle into n equal arcs, connecting the endpoints of the arcs with the north pole by great circular quarter arcs, and choosing points at equal distance from the north pole on these arcs. They mark the vertices of a regular spherical n-gon. The interior angle α of such an n-gon can be made arbitrarily close to 180° by choosing the vertices of the n-gon close to the equator. On the other hand, the closer the vertices are to the north pole, the closer the n-gon comes to a Euclidean regular n-gon. Thus the interior angle α of a spherical n-gon can vary between the interior angle α_0 of a Euclidean n-gon and π. To determine α_0, observe that the n *exterior angles* $\pi - \alpha_0$ add up to 2π, so that we have $n(\pi - \alpha_0) = 2\pi$ and thus $\alpha_0 = \left(1 - \frac{2}{n}\right)\pi$.

In order for a regular n-gon to tile the sphere, we must have $\alpha = \frac{2\pi}{a}$ for an integer $a \geq 3$ so that a polygons can fit around a vertex.

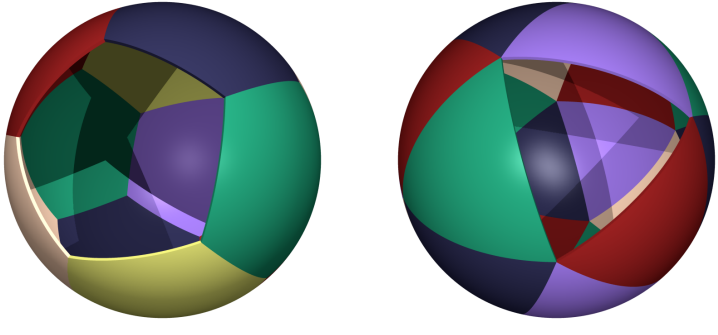
Moreover, we can determine the area of a regular n-gon with interior angle $\alpha = \frac{2\pi}{a}$ as $2\pi - n(1 - \frac{2}{a})\pi$.

We discuss the possible cases separately, using the inequality $\frac{1}{2} - \frac{1}{n} < \frac{1}{a} < \frac{1}{2}$:

1. If $n \geq 6$, $\frac{1}{3} = \frac{1}{2} - \frac{1}{6} < \frac{1}{2} - \frac{1}{n} < \frac{1}{a} < \frac{1}{2}$, so that this condition cannot be achieved for an integer a.
2. If $n = 5$, $\frac{3}{10} = \frac{1}{2} - \frac{1}{5} < \frac{1}{a} < \frac{1}{2}$, which allows for $a = 3$. The area of a 120° regular pentagon is $2\pi - 5(1 - \frac{2}{3})\pi = \frac{1}{3}\pi$, so that we need 12 pentagons to tile the sphere. This is accomplished by the regular dodecahedron.
3. If $n = 4$, $\frac{1}{4} < \frac{1}{a} < \frac{1}{2}$, which allows for $a = 3$. The area of a 120° square is $\frac{2}{3}\pi$, so that we need six such squares, and this is made possible by the cube.
4. If $n = 3$, $\frac{1}{6} < \frac{1}{a} < \frac{1}{2}$, so that $a \in \{3, 4, 5\}$. We need 20, 8, and 4 triangles, respectively, and the tilings are realized by the icosahedron, octahedron, and tetrahedron.



Tetrahedron Cube Octahedron



Dodecahedron Icosahedron

Figure 20 Platonic Tilings of S^2

2.6 Coxeter Triangles

In this section, we will look at tilings of the sphere and the Euclidean plane by special triangles.

Definition 2.8 A *Coxeter triangle* is a triangle with angles $\alpha = \pi/a$, $\beta = \pi/b$, and $\gamma = \pi/c$, where $a, b, c \geq 2$ are integers.

Theorem 2.7 *In the Euclidean plane, the only Coxeter triangles are given up to permuting the angles by*

$$(a, b, c) \in \{(2, 3, 6), (2, 4, 4), (3, 3, 3)\}$$

Proof: Because the sum of the angles in a Euclidean triangle is 180°, we need to find integral solutions of

$$\frac{1}{a} + \frac{1}{b} + \frac{1}{c} = 1$$

with $2 \leq a \leq b \leq c$. If $a \geq 3$, then

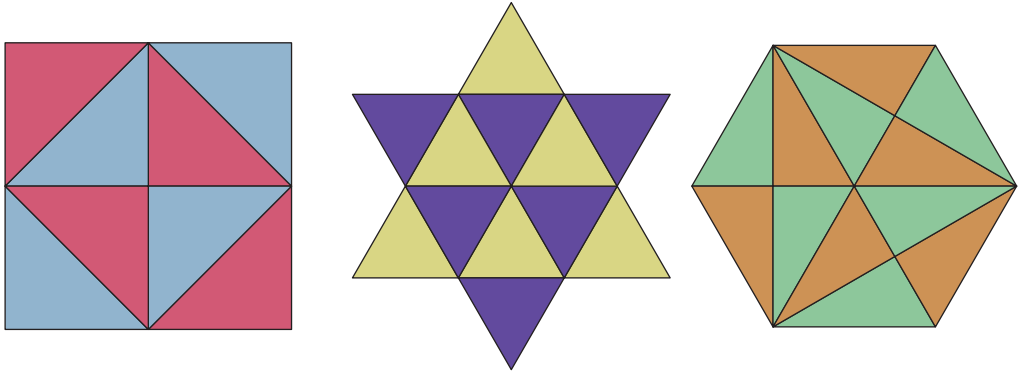


Figure 21 Euclidean Coxeter triangles

$$\frac{1}{a}+\frac{1}{b}+\frac{1}{c}\leq 1\ ,$$

so $a=b=c=3$, giving us the case $(3,3,3)$. Now we can assume that $a=2$. Then it remains to solve

$$\frac{1}{b}+\frac{1}{c}=\frac{1}{2}\ .$$

If $b \geq 4$, then $b = c = 4$, and we have the case $(2, 4, 4)$. As b cannot be 2, it remains to consider $b = 3$, leading to $(2, 3, 6)$. $\square$

Theorem 2.8 *In the sphere $\mathbb{S}^2$, all Coxeter triangles are given up to permuting the angles by*

$$(a, b, c) \in \{(2, 2, c), (2, 3, 3), (2, 3, 4), (2, 3, 5)\}$$

for $c \geq 2$.

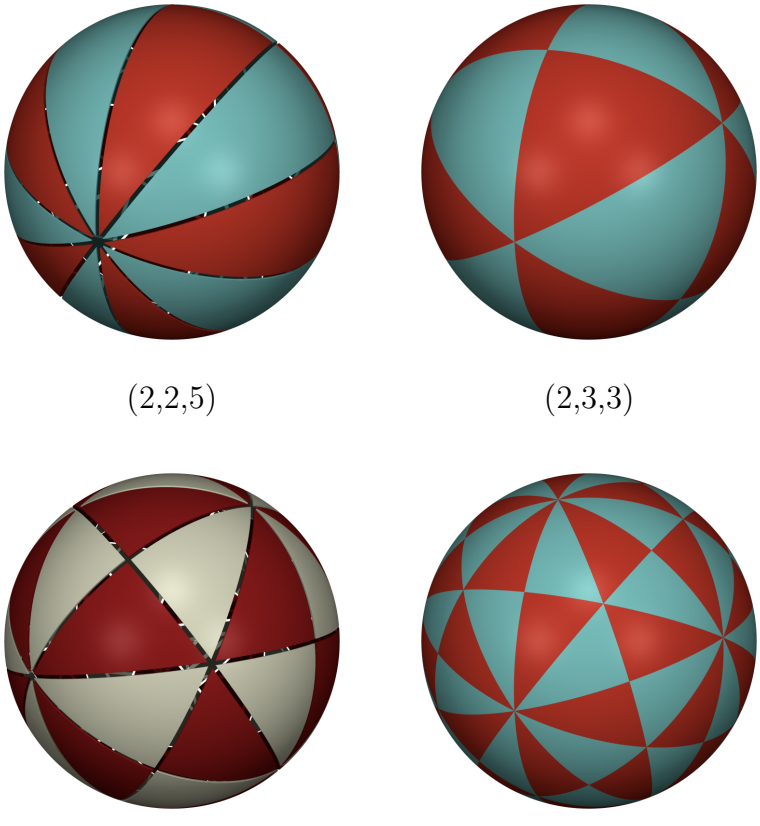


Figure 22 Coxeter triangles in $\mathbb{S}^2$

Proof: Because the sum of the angles in a spherical triangle is greater than 180°, we need to find integral solutions of

$$\frac{1}{a}+\frac{1}{b}+\frac{1}{c}>1$$

with $2 \le a \le b \le c$. Clearly $a \ge 3$ is impossible, so we have $a = 2$. It remains to solve

$$\frac{1}{b}+\frac{1}{c}>\frac{1}{2}\,.$$

Again, $b \ge 4$ is impossible. If $b = 3$, we are left with

$$\frac{1}{c}>\frac{1}{6}\,,$$

or $c \le 5$. This gives the cases $(2,3,3), (2,3,4), (2,3,5)$. If $b = 2$, then $c \ge 2$ is arbitrary. □

3 Möbius Transformations

3.1 The Geometry of Complex Numbers

Real numbers were originally invented to measure lengths, area, time and other quantities. In many mathematical problems, such numbers are often obtained not by direct measuring, but by solving equations. For instance, if we want to know where the line $y = 2x + 1$ intersects the unit circle $x^2 + y^2 = 1$, we need to solve these two equations *simultaneously* for x and y. Using the value of y from the first equation in the second gives the quadratic equation $x^2+(2x+1)^2 = 1$, or, simplified, $5x^2 + 4x = 0$, which is solved by $x = 0$ and $x = -4/5$, giving the two intersection points $(0, 1)$ and $(-4/5, -3/5)$. So far, so good. However, if the line is given by $y = 2x + 3$, the quadratic equation becomes $5x^2 + 12x + 8 = 0$, which doesn't have a solution. This reflects the fact that the line in question does not pass through the circle, so there are no intersection points. If we use the formula for the quadratic equation, we are nevertheless led to 'solutions', namely

$$(x, y) = -\left(\frac{6}{5} \pm \frac{2}{5}\sqrt{-1}, \frac{2}{5} \pm \frac{4}{5}\sqrt{-1}\right)$$

with the same choice of sign in both cases. This assumes that we can compute with $\sqrt{-1}$ as with any other number, which raises two questions: can we?, and if yes, is this something useful?
The answer to both questions is 'yes'.
We begin with the discussion with the introduction of a new symbol. We will use the letter i for a constant that we want to mean something that has the properties of $\sqrt{-1}$. It is, however, important to understand that we cannot just define $i = \sqrt{-1}$, because the latter has no meaning. Hence we will treat i as a symbol like the number π, which, however, has some unprecedented properties.
A *complex number* is by definition a composite number $a + bi$ where a and b are ordinary real numbers. We agree to add, subtract and multiply them according to

$$(a_1 + b_1 i) \pm (a_2 + b_2 i) = (a_1 + a_2) \pm (b_1 + b_2)i$$
$$(a_1 + b_1 i) \cdot (a_2 + b_2 i) = (a_1 a_2 - b_1 b_2) + (a_1 b_2 + a_2 b_1)i$$

The second equation is motivated by our desire that $i \cdot i = -1$ and that we can use the standard laws of computation, like the distributivity of multiplication.
These operations also have geometric significance. The addition of complex numbers is the same as the addition of vectors. To study the effect of multiplication, let's find out what the multiplication by i accomplishes:

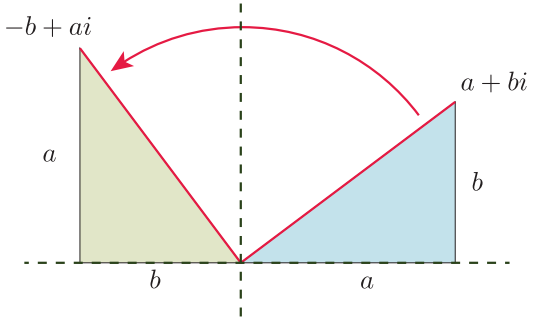


Figure 23 Multiplication by i

A complex number $z = a + bi$ determines a right triangle with legs of length a and b with one vertex at the origin. The complex number $iz = -b + ia$ determines a congruent triangle, and because the angles at the hypothenuse of a right triangle add up to 90° we see that iz is obtained from z by rotating z counter-clockwise around the origin by 90°. We will see below that complex multiplication in general has a lot to do with rotations. This is one of the reasons why complex numbers play an important role in planar geometry.

Now it is a tedious matter to verify that all the laws of computation we are using for real numbers remain valid for complex numbers. For instance, the commutativity of multiplication says that

$$(a_1 + b_1 i) \cdot (a_2 + b_2 i) = (a_2 + b_2 i) \cdot (a_1 + b_1 i)$$

which can now be checked using the definition of complex multiplication on both sides.

While addition and multiplication are very easily defined, this is not immediately clear for the reciprocal of a complex number: For any real number $x \neq 0$, we can compute a reciprocal $1/x$, characterized by $x \cdot 1/x = 1$. This allows to introduce the quotient of two real numbers.

There is an elegant way to define the reciprocal of a complex number which makes use of the *conjugate* of a complex number:

Definition 3.1 Let $z = a + bi$ be a complex number. Define its conjugate by

$$\bar{z} = \overline{a + bi} = a - bi$$

We also define the *norm* of a complex number as

$$|z| = \sqrt{a^2 + b^2}$$

The important fact about the conjugate is the following formula

$$(a + bi) \cdot \overline{(a + bi)} = a^2 + b^2$$

which is a real number. As a consequence, the norm is related to the conjugate by

$$|z|^2 = z \cdot \bar{z}$$

Thus, assuming that there is a reciprocal for a complex number, we can compute

$$\begin{aligned}\frac{1}{a+bi} &= \frac{a-bi}{(a+bi)\cdot\overline{(a+bi)}} \\ &= \frac{a-bi}{a^2+b^2} \\ &= \frac{a}{a^2+b^2} - \frac{b}{a^2+b^2}i\end{aligned}$$

This formula is now used as a definition of the reciprocal.
We finish the discussion of the elementary arithmetic of complex numbers with an innocent looking equation:

Theorem 3.1 *For any two complex numbers z_1 and z_2,*

$$|z_1 \cdot z_2| = |z_1| \cdot |z_2|$$

Proof: We use the commutativity of the complex multiplication:

$$\begin{aligned} |z_1 \cdot z_2|^2 &= z_1 z_2 \overline{z_1 z_2} \\ &= z_1 z_2 \overline{z_1}\,\overline{z_2} \\ &= z_1 \overline{z_1} z_2 \overline{z_2} \\ &= |z_1|^2 |z_2|^2 \ . \end{aligned}$$

□

Example 3.1 This equation has many consequences, as we will see. To give an idea how quickly one arrives at nontrivial facts, consider the two numbers 17 and 13. Both can be written as sums of two squares, namely as $17 = 1 + 16$ and $13 = 4 + 9$. I claim that their product, $221 = 17 \cdot 13$, can also be written as the sum of two squares. Using complex numbers, we just compute

$$(1 + 4i)(2 + 3i) = -10 + 11i$$

so that $221 = 10^2 + 11^2$.

Exercise 3.1 Write $1682 = 29 \cdot 58$ as the sum of two squares.

Instead of writing complex numbers in the Cartesian way as $z = x + iy$, it is often useful to write them using *polar coordinates* as $z = r(\cos t + i \sin t)$. Here $r = |z|$

is the modulus of z and $t = \arg z$ is called the argument of z. Unless $z = 0$, the argument is a well-defined number in the interval $[0, 2\pi)$.

Exercise 3.2 Compute the argument and modulus of the following complex numbers:

$$\frac{3}{5} + \frac{4}{5}i, \qquad 1 - i, \qquad \frac{1}{2} - \frac{\sqrt{3}}{2}i$$

We will now introduce a new function, the cis-function: For any real number t, write

$$\operatorname{cis}(t) = \cos t + \sin t \cdot i$$

This function is important because it satisfies an important addition law:

$$\operatorname{cis}(s + t) = \operatorname{cis}(s) \operatorname{cis}(t)$$

This formula is nothing but a compact version of the addition laws of the trigonometric functions:

$$\cos(s + t) = \cos s \cos t - \sin s \sin t$$
$$\sin(s + t) = \cos s \sin t + \sin s \cos t$$

It should also remind us of the exponential function which satisfies the same law

$$\exp(s+t) = \exp(s)\exp(t)$$

In fact, the exponential function can be defined for complex arguments $z = x + yi$ as

$$\exp(x+iy) = \exp(x)\operatorname{cis}(y)$$

in a meaningful and consistent way. Thus we can also write $\operatorname{cis}(t) = e^{it}$.

Exercise 3.3 Write $\left(\frac{1}{2} - \frac{\sqrt{3}}{2}i\right)^{17}$ in the form $a + bi$.

The next reason why complex numbers have been successful is that their arithmetic operations have simple geometric interpretations. Just as we imagine the set of real numbers as the *real line*, we can imagine the set of complex numbers as the plane, by identifying the number $a + bi \in \mathbb{C}$ with the point $(a, b) \in \mathbb{R}^2$.
Now choose a complex number $z_0 \in \mathbb{C}$ and consider the map

$$\begin{aligned} t_0 : \mathbb{C} &\to \mathbb{C} \\ z &\mapsto z + z_0 \end{aligned}$$

Geometrically, this is just a translation by the *vector* z_0.
Similarly, we would like to understand the geometric effect of multiplication by z_0, i.e. the map

$$r_0 : \mathbb{C} \to \mathbb{C}$$
$$z \mapsto z \cdot z_0$$

This can be achieved by writing z_0 as $z_0 = r_0 \operatorname{cis}(t_0)$, using polar coordinates. Similarly, we write $z = r \operatorname{cis}(t)$, and our multiplication map becomes

$$r \operatorname{cis}(t) \mapsto r \operatorname{cis}(t) r_0 \operatorname{cis}(t_0) = r r_0 \operatorname{cis}(t + t_0)$$

Hence two complex numbers can be multiplied by multiplying their moduli and adding their arguments. In particular, multiplication by a complex number $\operatorname{cis}(t)$ of modulus 1 has the effect of a rotation by angle t around the origin.

Exercise 3.4 Given a circle and two distinct points on it, show that there is a unique generalized circle that is orthogonal to the given circle and passes through the two chosen points.

Exercise 3.5 Given three pairwise disjoint circles, show that there is at least one generalized circle that is orthogonal to all three.

3.2 Möbius Transformations

Recall that for complex numbers $a, b, c, d \in \mathbb{C}$ such that $ad - bc \neq 0$, the map

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

is called a *Möbius transformation*. It is an invertible map from $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ to itself, if we declare that division by 0 (resp. ∞) has the value ∞ (resp. 0). The inverse is given by the Möbius transformation

$$f(z) = \frac{dz - b}{-cz + a} \ ,$$

which shows that the Möbius transformations form a group, where the group multiplication is the composition. This group is related to the general linear group $GL(2, \mathbb{C})$ of invertible 2×2 matrices:
Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ be an invertible 2×2 matrix, and define a Möbius transformation $f_A(z) = \frac{az+b}{cz+d}$. Note that if two matrices A and B are projectively equivalent, i.e. there is a number λ such that $B = \lambda A$, then $f_A = f_B$. Therefore the group of Möbius transformations is also called the *projective linear group* and denoted by $PSL(2, \mathbb{C})$.

For two invertible matrices A and B,

$$f_{AB} = f_A \circ f_B \ .$$

This means that the map $A \to f_A$ is a group homomorphism from $GL(2,\mathbb{C})$ to $PSL(2,\mathbb{C})$.
Our next objective will be to understand the geometry of Möbius transformations.

Lemma 3.1 *For any three different numbers $a, b, c \in \hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$, there is a Möbius transformation that maps $a \mapsto 0$, $b \mapsto 1$, and $c \mapsto \infty$.*

Proof: If $a, b, c \neq \infty$, then

$$f(z) = \frac{z-a}{z-c} \Big/ \frac{b-a}{b-c}$$

does the job. In case one of a, b, or c is equal to ∞, the transformation can be found by replacing the terms by 1 in the above formula that contain a, b, or c. For instance, if $a = \infty$, we have

$$f(z) = \frac{1}{z-c} \Big/ \frac{1}{b-c} = \frac{b-c}{z-c} \ .$$

□

Corollary 3.1 *Let a, b, c and a', b', c' be two triplets of pairwise distinct elements of $\hat{\mathbb{F}}$. Then there is a unique Möbius transformation that maps $a \to a'$, $b \to b'$ and $c \to c'$.*

Proof: Denote by f the transformation given by the Lemma, and by g the one that uses instead of a, b, c the elements a', b', c'. Then the composition $g^{-1} \circ f$ proves the existence of such a Möbius transformation. To show uniqueness, it suffices to show that any Möbius transformation $f(z) = \frac{az+b}{cz+d}$ that fixes 0, 1, and ∞ is the identity. The condition $f(0) = 0$ implies that $b = 0$, and $f(\infty) = \infty$ implies that $c = 0$. Thus $f(z) = \frac{a}{d}z$, and $f(1) = 1$ forces $\frac{a}{d} = 1$ so that indeed $f(z) = z$.

□

Exercise 3.6 Find all Möbius transformations that fix the points -1 and 1.

Exercise 3.7 Let S be a finite set with n elements. A *permutation* of S is any bijective (one-to-one and onto) map $\phi : S \to S$. The permutations of S form a group $S_n = \mathrm{Sym}(S)$, the *symmetric group*, by using the composition of maps as group multiplication. This group has $n!$ elements.

Now let $S = \{0, 1, \infty\} \subset \hat{\mathbb{C}}$. For each element $\phi \in \mathrm{Sym}(S)$, find the Möbius transformation that realizes ϕ, i.e. has the same effect on S as ϕ.

Exercise 3.8 Let $\zeta = e^{2\pi i/3} = \frac{1}{2}(-1 + \sqrt{3}i)$ be a third root of unity. Note that $\bar{\zeta} = 1/\zeta = \zeta^2$. Let $S = \{0, 1, \zeta, \bar{\zeta}\}$. Which elements of $\mathrm{Sym}(S)$ can be realized by Möbius transformations?

3.3 Geometry of Möbius Transformations

In this section, we will study the geometry of Möbius transformations.

Lemma 3.2 *Every Möbius transformation can be written as the composition of linear transformations and at most one reciprocal:*

$$\frac{az+b}{cz+d} = \frac{a}{c} - \frac{ad-bc}{c}\frac{1}{cz+d}$$

As a consequence, we can prove:

Theorem 3.2 *A Möbius transformation maps a generalized circle to a generalized circle.*

Proof: We already know that linear Möbius transformations $z \mapsto az + b$ are similarities and hence map circles to circles and lines to lines. As every Möbius transformations is by **Lemma 3.2** a composition of the reciprocal $1/z$ with linear Möbius transformations, it suffices to show the claim for the reciprocal. But

$1/z = I(\bar{z})$, so the claim follows from the same statement for the inversion and because the complex conjugation is a reflection at the real axis. □

We can therefore introduce a new incidence geometry, the *Möbius geometry*. Its points are the points of $\hat{\mathbb{E}}$, its lines are generalized circles, and we have just seen that all Möbius transformation are incidence automorphisms.
At the core of our discussion of Möbius transformations and generalized circles lies the fact that three distinct points determine both a circle and a Möbius transformation uniquely. Hence, in principle, we could have defined circles as images of lines under Möbius transformations. We will explore this a bit further: How can we detect whether it is possible to map four given points to four other points?

Definition 3.2 For four distinct numbers a, b, c, d, the *cross ratio* is defined as

$$[a, b; c, d] = \frac{d-a}{d-c} \Big/ \frac{b-a}{b-c} \, .$$

In other words, it is the value of d under the Möbius transformation that takes $a \mapsto 0$, $b \mapsto 1$, and $c \mapsto \infty$.

Theorem 3.3 *A Möbius transformation $f(z)$ preserves the cross ratio, i.e. we have*

$$[f(a), f(b); f(c), f(d)] = [a, b; c, d]$$

Proof: Let $g(z) = [a, b; c, z]$ and $h(z) = [f(a), f(b); f(c), z])$. Then $h^{-1} \circ g = f$ because both sides are Möbius transformations that do the same to a, b, and c. The claim follows. □

Corollary 3.2 *Four points lie on a circle or line if and only if their cross ratio is real.*

Proof: Clearly, the cross ratio of four real numbers is real. As any generalized circle can be mapped to the real axis by a Möbius transformation, the cross ratio of four points on *any* generalized circle will be real as well. If on the other hand four points a, b, c, d have a real cross ratio, the Möbius transformation $f : z \mapsto [a, b; c, z]$ will map a, b, c to 0, 1, ∞, respectively. By assumption, the cross ratio $[a, b; c, d]$ is real, so d will be mapped to a real number as well. Thus the four points lie on the image of the extended real line $\hat{\mathbb{R}}$ under the Möbius transformation f^{-1}, which is a generalized circle. □

Exercise 3.9 Let a_1, a_2, b_1, b_2 be four distinct points in $\hat{\mathbb{C}}$. Show that there is a Möbius transformation that exchanges the a_i and exchanges the b_i.

Lemma 3.3 *A Möbius transformation that maps two concentric circles to two other concentric circles is linear.*

Proof: After precomposing and postcomposing the Möbius transformation with linear transformations (i.e. homotheties), we can assume that the two given circles have centers at 0 and radii 1 and r, while their images also have centers at 0 and radii 1 and s, respectively. A diameter of the original circles will be mapped to a generalized circle C orthogonal to the two concentric image circles C_1 and C_2. We claim that this is again a line, i.e. a diameter of the image circles. Take the common diameter D of C_1 and C_2 that passes through the intersection p of C_1 and C. Invert everything at a circle with center p. The images of D and C become lines orthogonal to a line that is the image of C_1, while the image of C_2 is still a circle orthogonal to the images of D and C. But if a line is orthogonal to a circle, it must be a diameter. Hence the images of D and C are equal, and hence $C = D$.

By applying a rotation we can assume that the normalized Möbius transformation maps one diameter of the two circles to itself. Thus it fixes four points, namely

the intersections of that diameter with the two circles, and hence is the identity. □

4 Hyperbolic Geometry

4.1 The Disk Model

The Möbius geometry of $\hat{\mathbb{C}}$ is very different from Euclidean geometry. For instance, through any two points there are infinitely many circles incident with these points. We now come to a geometry that is in may ways very similar to Euclidean geometry. The set of points of this *hyperbolic geometry* will be the open unit disk $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$, and the lines will be the intersections of generalized circles that are orthogonal to $\mathbb{S}^1$. We begin by looking at Möbius transformations that map $\mathbb{D}$ bijectively to itself.
We need the following inequality:

Lemma 4.1 *Let $|z| \leq 1$ and $|w| \leq 1$. Then*

$$\left|\frac{z-w}{1-\bar{w}z}\right| \leq 1$$

with equality if and only if $|z| = 1$ or $|w| = 1$.

Proof: We equivalently modify the claim until it becomes obvious:

$$\begin{aligned} |z-w|^2 &\leq |1-\bar{z}w|^2 \\ \Leftrightarrow z\bar{z} - z\bar{w} - \bar{z}w + w\bar{w} &\leq 1 - z\bar{w} - \bar{z}w + z\bar{z}w\bar{w} \\ \Leftrightarrow z\bar{z} + w\bar{w} &\leq 1 + z\bar{z}w\bar{w} \\ \Leftrightarrow 0 &\leq (1-|z|^2)(1-|w|^2) \end{aligned}$$

□

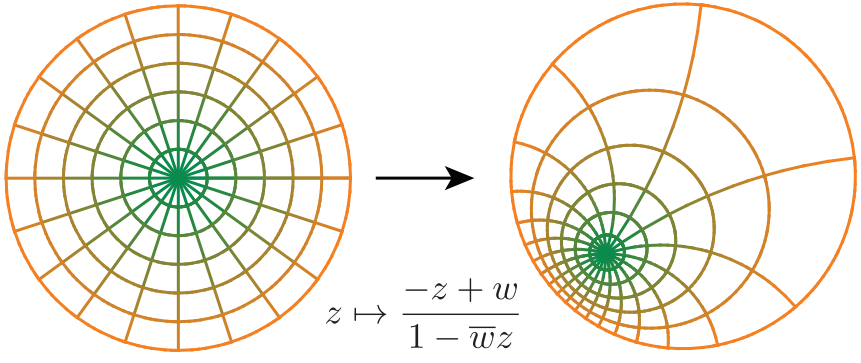


Figure 24 A Möbius automorphism of the disk

Corollary 4.1 *For $|w| < 1$, the Möbius transformation*

$$\mu_{w,\alpha} : z \mapsto e^{i\alpha} \frac{z - w}{1 - \overline{w}z}$$

is a bijective map of the unit disk $\mathbb{D}$ onto itself with $w \mapsto 0$ and inverse

$$z \mapsto \mu_{-e^{i\alpha}w,-\alpha}(z) = \frac{z + e^{i\alpha}w}{e^{i\alpha} + \overline{w}z} \ .$$

Proof: By **Lemma 4.1**, both Möbius transformations map $\mathbb{D}$ into $\mathbb{D}$, and it is easy to check that their composition is the identity. □

We use these Möbius transformations to prove again:

Corollary 4.2 *For any two disjoint circles, there is a Möbius transformation that maps these two circles to concentric circles.*

Proof: Without loss of generality, we can assume that the first circle is the unit circle. If the second circle lies outside the unit circle, we can send it inside using

$1/z$. Thus we can assume that it lies inside the unit circle. After a rotation, we can also assume that its diameter is on the real axis and intersects it in points $-1 < a < b < 1$.
Let's look at the Möbius transformations

$$g_t(z) = \mu_{-t,0}(z) = \frac{z+t}{1+tz}$$

for $t \in (-1,1)$. They fix the unit circle and the real axis. Moreover, for any $t \in (-1,1)$ and $a \in [-1,1]$, $g_t(a)$ maps $[-1,1]$ to itself such that $g_t(-1) = -1$ and $g_t(+1) = +1$. Thus, for $-1 < a < b < 1$, there is a $t \in (-1,1)$ such that $g_t(a) + g_t(b) = 0$. Thus for this t the Möbius transformation g_t will map the given circle to a circle centered within the unit disk, as claimed. □

Exercise 4.1 Find a Möbius transformation that maps the two circles with center at ± 2 and radius 1 to two concentric circles so that one of them is the unit circle.

Thus we have a new incidence geometry: The points are the points of the unit disk, and the lines are the intersections of generalized circles with the unit disk that are orthogonal to the unit circle. This incidence geometry is called *hyperbolic geometry*. It shares surprisingly many properties with Euclidean geometry but also is vastly different.

Definition 4.1 A Möbius transformation that maps the unit disk to itself is called a *hyperbolic automorphism*.

Euclid formulated five postulates for the Euclidean plane:

1. A straight line segment can be drawn joining any two points.
2. Any straight line segment can be extended indefinitely in a straight line.
3. Given any straight lines segment, a circle can be drawn having the segment as radius and one endpoint as center.
4. All Right Angles are congruent.
5. Given a line and a point not on it, at most one line parallel to the given line can be drawn through the point.

The first four are easy to verify in hyperbolic geometry. To prove the first one, we begin with a characterization of diameters. This will often allow us to reduce the case of a general hyperbolic line to that of a diameter.

Lemma 4.2 *Any hyperbolic line through the center of the disk is a diameter.*

Proof: Assume there is a hyperbolic line through the center that is not a diameter, and suppose it meets $\mathbb{S}$ in points p and q orthogonally. Take a diameter that passes through p, and invert everything at a circle with center at p. We obtain three lines: First the image of $\mathbb{S}$, then, orthogonal to it, the images of the diameter and the hyperbolic line. As the latter are parallel and are both incident with 0, they must be equal. □

Theorem 4.1 *For any two points there is a unique hyperbolic line incident with them.*

Proof: After applying a Möbius transformation, we can assume that one of the two points is the center of the disk. Then a hyperbolic line incident with both points is the diameter through them. The uniqueness follows from **Lemma 4.2**. □

Exercise 4.2 Show that for any two orthogonal hyperbolic lines there is a Möbius transformation that maps them to the horizontal and vertical diameter of $\mathbb{D}$.

However, the fifth postulate is incorrect in hyperbolic geometry:

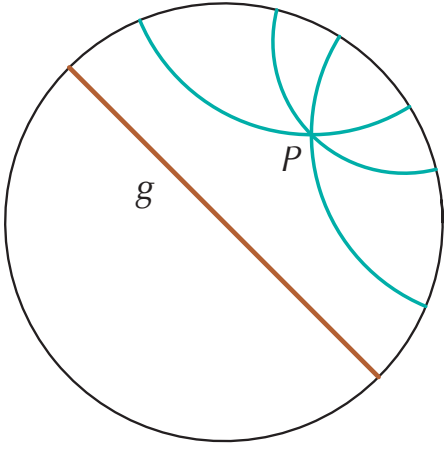


Figure 25

For over 2000 years, mathematicians had tried to prove the parallel postulate by only using Euclid's first four postulates. Around 1820, János Bolyai, Nicolai Lobachevsky and Carl Friedrich Gauss were all independently aware of a new geometry that satisfies the first four axioms but not the fifth axiom.
Farkas Bolyai was a friend of Gauss and had tried for many years to prove the parallel postulate. In an earlier letter from 1820, he warned his son János about the difficulty of the problem in emotional words:

I pity you. I see that my unhappy life repeats itself in you. I see you tossed around in the darkest storm among dangerous cliffs where so far everybody has suffered shipwreck. It is an eerie battle field where I was beat every time, an unconquerable fortress that withstands all inquiring minds. In this matter all life is nothing but a torch submerged into the sea.
The father continues like this for several paragraphs. At this time, János was already in the possession of the key ideas of hyperbolic geometry, and he responded : *My desire and energy to penetrate this problem grew immensely.*
Then, in 1823, János Bolyai wrote in a letter to his father Farkas: *I have created a new world out of nothing.*
He developed his theory into a long manuscript called *Appendix*, published in 1830, one year after Lobachevsky had published his own, similar results without knowing about each other. Gauss's comment about this was just that he had known all this for quite some time.
The Möbius transformations described in **Corollary 4.1** are incidence automorphisms of hyperbolic geometry. We now show that there are no other Möbius transformations with this property:

Theorem 4.2 *Every Möbius transformation that maps the unit disk bijectively into itself is of the form $\phi_{t,w}$ given in* **Corollary 4.1**.

Proof: Let ϕ be a Möbius transformation that maps the unit disk bijectively into itself. After composing it with one of the $\phi_{t,w}$, we can assume that ϕ maps 0 to 0 and the horizontal diameter$[-1, 1]$ onto itself so that -1 and 1 are also fixed. As a Möbius transformation is determined by its value at three points, ϕ is the identity. □

Exercise 4.3 Prove the two following statements in the hyperbolic plane:

1. Given a hyperbolic line and a point, there is a hyperbolic line incident with the point and orthogonal to the given line.
2. For any two given hyperbolic lines that do not intersect, there is a unique hyperbolic line that intersects the two given lines orthogonally.

You can use all standard facts from Euclidean geometry, and your knowledge of hyperbolic geometry.

Solution: For the first part, apply a hyperbolic automorphism that maps the given point to the center of the hyperbolic disk. The given hyperbolic line is mapped to another hyperbolic line, i.e. a circle that intersects $\mathbb{S}^1$ orthogonally in two points p_1 and p_2. The line through the center of the disk and the midpoint of p_1 and p_2 solves the problem for the transformed situation, and we just need to transform back.

For the second part, find a Möbius transformation that maps the two given hyperbolic lines onto arcs of two concentric circles. Then $\mathbb{S}^1$ will become a diameter of these two circles. The line through the common center of the two circles and orthogonal to the diameter solves the transformed problem, we again just need to transform back.

4.2 The Upper Half Plane

In this section we will introduce the upper half plane model $\mathbb{H}$ of hyperbolic geometry, and use it to establish a close connection between hyperbolic geometry and algebra.

Definition 4.2 The *upper half plane model* of the hyperbolic plane is the half plane $\mathbb{H} = \{(x, y) \in \mathbb{R}^2 : y > 0\}$. Lines in this geometry are circular arcs perpendicular to the real axis or vertical line segments. Hyperbolic automorphisms of $\mathbb{H}$ are Möbius transformations that map the upper half plane bijectively to itself.

We can construct an incidence isomorphism between the disk and upper half plane model by using an inversion at a circle with center at $(0, -1)$ and through $(\pm 1, 0)$. This inversion takes the upper half plane to the unit disk and thus sends hyperbolic lines in one model to hyperbolic lines in the other model.
An alternative incidence isomorphism is given in the next exercise.

Exercise 4.4 Let

$$f(z) = i\frac{1-z}{1+z}$$

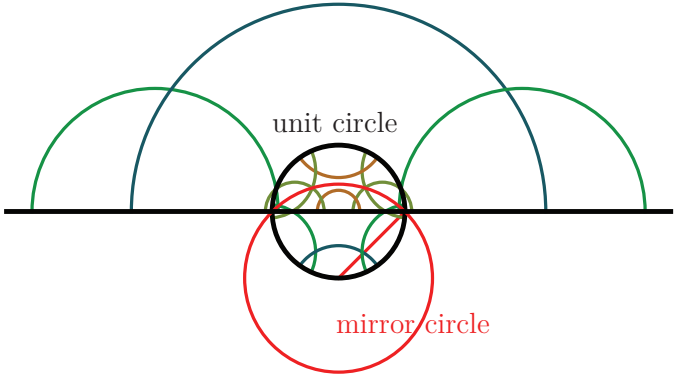


Figure 26 Mapping between disk and upper half plane

be the *C*ayley transformation. Show that it maps the unit disk bijectively onto the upper half plane. Hint: Write $f(z)$ as a composition of elementary Möbius transformations which we geometrically understand.

Solution: We have

$$f(z) = -i + \frac{2i}{1+z} \, .$$

The unit disk $\mathbb{D}$ is first translated to a disk of radius 1 centered at 1. The map $1/z$ is an inversion at the unit circle followed by conjugation, which takes this disk to a right half plane $x > \frac{1}{2}$. Multiplication by $2i$ maps this to the half plane $y > -1$, and the translation by $-i$ to $\mathbb{H}$. All these maps are bijective, and so is their composition.

We will now describe the incidence automorphisms of the upper half plane that are Möbius transformations. We begin with a few examples:

Example 4.1

- The translations $f(z) = z + d$ are Möbius transformations that can be represented by the 2×2 matrix $\begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix}$.
- The dilations $f(x) = r^2 z$ are Möbius transformations that can be represented by the 2×2 matrix $\begin{pmatrix} r & 0 \\ 0 & 1/r \end{pmatrix}$.
- The inversion is not a Möbius transformation, but if we compose it with a reflection at the y-axis, $f(z) = -1/z$ is one, which can be represented by the 2×2 matrix $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$.

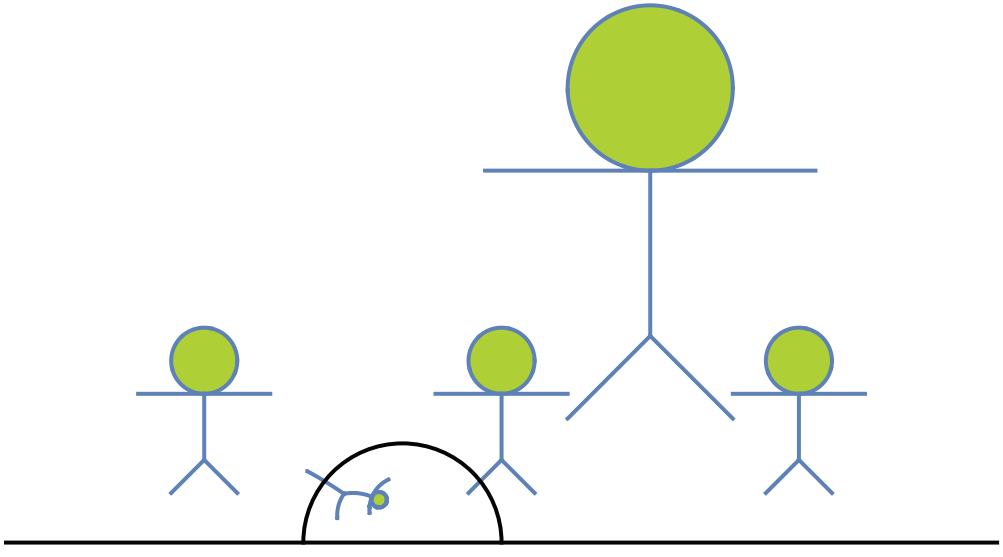


Figure 27 We are all congruent

Observe that the matrices we have chosen in this example to represent the Möbius transformations all have determinant 1.

Definition 4.3 The *special linear group* $\mathrm{SL}(2,\mathbb{R})$ is the group of real 2×2 matrices with determinant 1:

$$\mathrm{SL}(2,\mathbb{R}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : ad - bc = 1 \right\} .$$

Theorem 4.3 *Matrices in* $\mathrm{SL}(2,\mathbb{R})$ *represent hyperbolic automorphisms of the upper half plane.*

Proof: We show that every Möbius transformation with real coefficients is a composition of the ones in the example above.

If $c = 0$, then $ad = 1 \neq 0$, and hence $\frac{a}{d} > 0$. The claim follows from

$$\frac{az + b}{d} = \frac{a}{d} z + \frac{b}{d} .$$

If $c \neq 0$, the equation

$$\frac{az + b}{cz + d} = \frac{a}{c} - \frac{ad - bc}{c^2 z + cd}$$

shows the claim in this case.
Therefore, these are all incidence automorphisms. □

Exercise 4.5 Give an alternative proof of the above theorem by computing $\operatorname{Im} \frac{az+b}{cz+d}$.

Proposition 4.1 *Every Möbius transformation ϕ that maps the upper half plane bijectively onto itself is of the form*

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

with real a, b, c, d and $ad - bc > 0$.

Proof: By continuity, ϕ maps also the real line bijectively onto itself. We can compose it with a Möbius transformation with real coefficients so that three points on the real axis are fixed. This shows that ϕ can be written with real coefficients. We only need to verify the determinant condition. If $ad - bc = 0$, the Möbius transformation is constant. If $ad - bc < 0$, it maps the upper half plane to the lower half plane. □

Thus the properties of the group $\mathrm{SL}(2,\mathbb{R})$ can be studied by using hyperbolic geometry. In other words, if one wants to understand the multiplication of real 2×2 matrices, one has to understand hyperbolic geometry.

Exercise 4.6 Find an incidence automorphism of the upper half plane that fixes $(0,1)$ and that rotates every hyperbolic line through $(0,1)$ into a hyperbolic line perpendicular to the original one. Write the incidence automorphism as a Möbius transformation.

Exercise 4.7 The set $\mathrm{T} = \{\phi_t(z) = z+t\}$ is a subgroup of the group of hyperbolic automorphisms of the upper half plane. Using the Cayley transformation

$$\gamma(z) = i\frac{1-z}{1+z}\ ,$$

compute the *conjugate transformations* $\psi_t = \gamma^{-1} \circ \phi_t \circ \gamma$. These map the unit disk to itself and form a subgroup of the group of hyperbolic automorphisms of the disk. Find the common point on the unit circle that is fixed by all ψ_t. The *orbit* of 0 is the set $\{\psi_t(0) : t \in \hat{\mathbb{R}}\}$. Show that this is a circle, and find its center and radius.

4.3 Densities, Length, and Area

The Euclidean *length* of a curve $\gamma : [a, b] \to \mathbb{R}^2$ is defined to be

$$\text{length}(\gamma) = \int_a^b |\gamma'(t)|\, dt \ .$$

This definition is invariant under Euclidean motions because the Euclidean norm is invariant under orthogonal transformations. We would like to have a new definition for the length of curves in the hyperbolic plane that is invariant under hyperbolic automorphisms.
Similarly, the Euclidean area of a set $\Omega \subset \mathbb{R}^2$ is defined as

$$\text{area}(\Omega) = \int_\Omega 1\, dxdy \ .$$

In this case, the area is invariant under a Euclidean motion T, because

$$\begin{aligned}\operatorname{area}(T(\Omega)) &= \int\limits_{T(\Omega)} 1\, dxdy \\ &= \int\limits_{\Omega} |\det dT|\, dxdy \\ &= \operatorname{area}(\Omega)\end{aligned}$$

Here we have used the transformation formula and the fact that the Jacobi matrix of a Euclidean motion is an orthogonal matrix, which has determinant 1. Again, we would like to have an alternative definition for the area of sets in the hyperbolic plane that is invariant under hyperbolic automorphisms.
This will be accomplished by introducing the concept of a *density* in a set. This is a special case of the concept of a *Riemannian metric*, which is studied in differential geometry.

Definition 4.4 Let $\Omega \subset \mathbb{R}^2$ be a subset of $\mathbb{R}^2$. A *density* in Ω is a continuous function $\rho : \Omega \to \mathbb{R}^{>0}$. The ρ-length of a curve $\gamma : [a, b] \to \Omega$ is defined as

$$\text{length}_\rho(\gamma) = \int_a^b \rho(\gamma(t))|\gamma'(t)|\, dt \ ,$$

and the ρ-area of a subset $U \subset \Omega$ as

$$\text{area}_\rho(U) = \int_U \rho(x, y)^2\, dxdy \ .$$

We can think of the density in a set as a material property like viscosity. To explain the square $\rho(x, y)^2$ in the definition of the ρ-area, we consider the case of a constant density function $\rho(x, y) = \rho_0$. Compared to the Euclidean case where $\rho_0 = 1$, the length of curves is scaled by the factor ρ_0, so we expect all areas to be scaled by ρ_0^2. This is exactly what happens in that definition.
Our basic example will be the set $\Omega = \mathbb{H}$, the upper half plane model of the hyperbolic plane.

Definition 4.5 The *hyperbolic density* in the hyperbolic plane $\mathbb{H}$ is defined as

$$\rho(x, y) = \frac{1}{y} \ .$$

This definition ensures that the boundary (the real axis) becomes infinitely far away. We give two example computations:

Example 4.2 The hyperbolic length of the segment $[ai, bi]$ on the imaginary axis for $0 < a < b$ is obtained by parametrizing $\gamma : [a, b] \to \mathbb{H}$ as $\gamma(t) = it$ and computing

$$\text{length}_\rho(\gamma) = \int_a^b \frac{dt}{t} = \log(b/a) \ .$$

We clearly see that the length of this segment converges to ∞ when $a \to 0$.

Example 4.3 We compute the area of the hyperbolic triangle Δ given by $\cos(\alpha) < x < \cos(\beta)$ for $0 < \beta < \alpha < \pi$, and the half circle $x^2 + y^2 > 1$. This triangle has angles 0 at ∞, $\pi - \alpha$ at $\cos(\alpha)$, and β at $\cos(\beta)$.

$$\begin{aligned}
\text{area}_\rho(\Delta) &= \int_{\cos\alpha}^{\cos\beta} \int_{\sqrt{1-x^2}}^{\infty} \frac{1}{y^2}\, dydx \\
&= \int_{\cos\alpha}^{\cos\beta} \frac{1}{\sqrt{1-x^2}}\, dx \\
&= \arcsin(\cos(\beta)) - \arcsin(\cos\alpha) \\
&= \arcsin(\sin(-\beta + \frac{\pi}{2})) - \arcsin(\sin(-\alpha + \frac{\pi}{2})) \\
&= -\beta + \alpha \ .
\end{aligned}$$

Note the sign change in the last lines. We assume that $0 < \alpha < \pi$ so that $-\frac{\pi}{2} < \frac{\pi}{2} - \alpha < \frac{\pi}{2}$ where the sin function is invertible.
In terms of the interior angles $\alpha' = \pi - \alpha$, $\beta' = \beta$, and $\gamma' = 0$, the area of this particular triangle is given by

$$\text{area}_\rho\, \Delta = \pi - \alpha' - \beta' - \gamma' \ .$$

Our goal is to show that this formula is valid for any hyperbolic triangle. In contrast to Euclidean geometry, the edge lengths don't play a role!

We now show:

Theorem 4.4 *Let ϕ be a hyperbolic automorphism of $\mathbb{H}$, and $\Omega \subset \mathbb{H}$ a subset. Then* $\text{area}_\rho(\Omega) = \text{area}_\rho(\phi(\Omega))$.

Proof: This is clear for translations $\phi(z) = z+c$. We verify it for a dilation $\phi(z) = c\cdot z$:

$$\begin{aligned}\text{area}_\rho(\phi(\Omega)) &= \int\limits_{\phi(\Omega)} \frac{1}{y^2}\, dxdy \\ &= \int\limits_{\Omega} \frac{1}{(cy)^2} \cdot c^2\, dxdy \\ &= \text{area}_\rho(\Omega)\ .\end{aligned}$$

The factor c^2 appears because the Jacobian of ϕ has determinant c^2. Finally, we show that the inversion (and therefore $-1/z$) preserves area. As a preparation, we compute the Jacobian of the inversion

$$dI = \begin{pmatrix} \frac{y^2-x^2}{(x^2+y^2)^2} & -\frac{2xy}{(x^2+y^2)^2} \\ -\frac{2xy}{(x^2+y^2)^2} & \frac{x^2-y^2}{(x^2+y^2)^2} \end{pmatrix}$$

and

$$\det dI = -\frac{1}{(x^2+y^2)^2} \ .$$

Thus, by the change of variables formula,

$$\begin{aligned} \mathrm{area}_\rho(I(\Omega)) &= \int\limits_{I(\Omega)} \frac{1}{y^2}\, dxdy \\ &= \int\limits_{\Omega} \frac{1}{(y/(x^2+y^2))^2} \cdot |\det dI|\, dxdy \\ &= \mathrm{area}_\rho(\Omega) \ . \end{aligned}$$

□

We can use this to compute the hyperbolic area of any triangle:

Theorem 4.5 *Let Δ be a hyperbolic triangle with interior angles α, β, and γ. Then its hyperbolic area is*

$$\operatorname{area}_\rho \Delta = \pi - \alpha - \beta - \gamma$$

Proof: After applying a hyperbolic automorphism, and without changing the area or the angles, we can assume that the vertices A and B (at the angles α and β) lie on the imaginary axis. Add a vertical hyperbolic line through the vertex C (at the angle γ), making an angle δ with the segment BC. Then we have two triangles $AC\infty$ and $BC\infty$ with one ideal vertex, whose areas we computed in **example 4.3** as $\pi - \alpha - (\gamma + \delta)$ and $\pi - \beta - \gamma$. The difference of these areas is the area of the triangle ABC as claimed. □

Similarly, we have for the length of curves:

Theorem 4.6 *Let $\gamma : [a, b] \to \mathbb{H}$ be a curve, and $\phi : \mathbb{H} \to \mathbb{H}$ a hyperbolic automorphism. Then*

$$\operatorname{length}_\rho(\phi \circ \gamma) = \operatorname{length}_\rho(\gamma) \ .$$

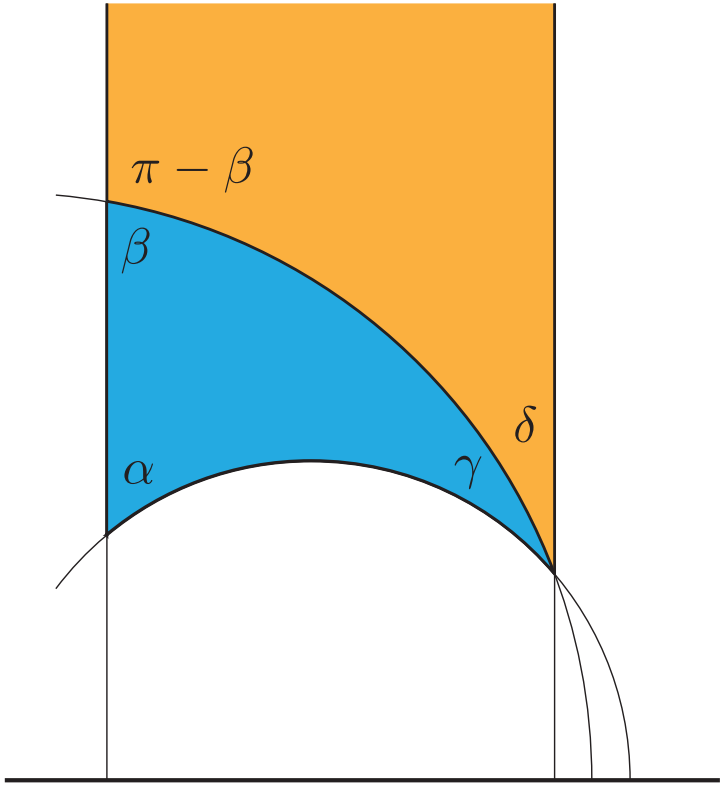


Figure 28 Hyperbolic area of triangles

Proof: It is again sufficient to verify this for translations, dilations, and the inversion.

We carry out the computation for the inversion. Write $\gamma(t) = (x(t), y(t))$. Then

$$(\phi \circ \gamma)'(t) = \frac{1}{(x(t)^2 + y(t)^2)^2} \begin{pmatrix} (y(t)^2 - x(t)^2)\, x'(t) - 2x(t)y(t)y'(t) \\ (x(t)^2 - y(t)^2)\, y'(t) - 2x(t)y(t)x'(t) \end{pmatrix}$$

so that

$$|(\phi \circ \gamma)'(t)|^2 = \frac{x'(t)^2 + y'(t)^2}{(x(t)^2 + y(t)^2)^2} \ .$$

Thus

$$\begin{aligned} \text{length}_{\rho}(\phi \circ \gamma) &= \int_a^b \frac{1}{y(t)/(x'(t)^2 + y'(t)^2)} \frac{|\gamma'(t)|}{x(t)^2 + y(t)^2}\, dt \\ &= \text{length}_{\rho}(\gamma) \end{aligned}$$

□

We can also define a density in the disk model of hyperbolic geometry:

Definition 4.6 The hyperbolic density of the disk is given by

$$\rho_{\mathbb{D}}(z) = \frac{2}{1-|z|^2} \, .$$

Proposition 4.2 *The Cayley transformation*

$$f(z) = i\frac{1-z}{1+z}$$

preserves the hyperbolic length of curves.

Proof: Let $z(t)$ be a curve in $\mathbb{D}$, and let $w(t) = f(z(t))$. Then

$$w' = -2i\frac{z'}{(1+z)^2}$$

and

$$\operatorname{Im} w = \frac{1-|z|^2}{|1+z|^2}$$

so that

$$\begin{aligned}\rho_{\mathbb{H}}(w)|w'| &= \frac{1}{\operatorname{Im} w}|w'| \\ &= \frac{(1+z)^2}{1-|z|^2}\frac{2|z'|}{|1+z|^2} \\ &= \frac{2|z'|}{1-|z|^2} \\ &= \rho_{\mathbb{D}}(z)|z'|\end{aligned}$$

□

Exercise 4.8 Define the spherical density on $\mathbb{R}^2$ by

$$\rho_{\mathbb{S}}(x,y) = \frac{2}{1+x^2+y^2}\ .$$

Show that the inverse stereographic projection

$$s : (x,y) \mapsto \frac{2}{1+x^2+y^2}\left(z, \frac{1}{2}(x^2+y^2-1)\right)$$

preserves the length of curves when the length of a curve $c(t) = (x(t), y(t))$ is measured using the spherical density $\rho_{\mathbb{S}}$ and the length of the image curve $q(t) = s(p(t))$ is measured using the constant Euclidean density 1 of $\mathbb{R}^3$.

Solution: Check that

$$|q'(t)| = \rho_{\mathbb{S}}(p(t))|p'(t)| \ .$$

4.4 The Hyperbolic Metric

Our next goal is to define a distance that is being preserved by all hyperbolic automorphisms. To this end we will use the upper half plane model. This section is slightly redundant in that it does not rely on the previous one.

Proposition 4.3 *Let $p = (p_x, p_y)$ and $q = (q_x, q_y)$ be points in $\mathbb{H}$. The function*

$$\delta(p, q) = \frac{|p - q|^2}{p_y q_y}$$

is invariant under hyperbolic automorphisms.

Proof: Clearly, this function is invariant under horizontal translations and dilations. Thus it suffices to check that it is also invariant under reflection at the unit circle. Let's compute

$$\begin{aligned}
\delta\left(I(p), I(q)\right) &= \delta\left(\frac{p}{|p|^2}, \frac{q}{|q|^2}\right) \\
&= \frac{\left|\frac{p}{|p|^2} - \frac{q}{|q|^2}\right|^2}{\frac{p_2}{|p|^2}\frac{q_2}{|q|^2}} \\
&= \frac{\left|p\frac{|q|}{|p|} - q\frac{|p|}{|q|}\right|^2}{p_y q_y} \\
&= \frac{|p-q|^2}{p_y q_y}
\end{aligned}$$

□

Any function of $\delta(p, q)$ will also be invariant under the hyperbolic automorphisms.

Definition 4.7 We define the hyperbolic distance between two points in $\mathbb{H}$ as

$$d(p, q) = \cosh^{-1}(1 + \frac{1}{2}\delta(p, q))$$

For the moment, we postpone the proof that this indeed satisfies the triangle inequality. This particular choice is motivated by the following:

Proposition 4.4 *The hyperbolic metric is additive along hyperbolic lines.*

Proof: It suffices to show this for a specific geodesic, and we choose the y-axis. Recall that

$$\cosh^{-1}(x) = \log(x + \sqrt{x^2 - 1}) \ .$$

For $p = (0, a)$ and $q = (0, b)$ with $a < b$, we compute

$$
\begin{aligned}
d(p,q) &= \cosh^{-1}\left(1+\frac{1}{2}\frac{(a-b)^2}{ab}\right) \\
&= \cosh^{-1}\left(\frac{a^2+b^2}{2ab}\right) \\
&= \log\left(\frac{a^2+b^2}{2ab}+\sqrt{\left(\frac{a^2+b^2}{2ab}\right)^2-1}\right) \\
&= \log\left(\frac{a^2+b^2}{2ab}+\frac{b^2-a^2}{2ab}\right) \\
&= \log(b)-\log(a)
\end{aligned}
$$

□

Corollary 4.3 *The distance between two points is equal to the length of the hyperbolic line connecting them.*

Proof: This follows from **Example 4.2**. □

The additivity of the distance allows us to prove the triangle inequality:

Proposition 4.5 *The hyperbolic metric satisfies the triangle inequality.*

Proof: Let p, q, r be three points in the upper half plane. We can find a hyperbolic automorphism that maps the hyperbolic line through p and r to the y-axis. Denote by q' the point $(0, q_2)$ on the y-axis at the same height as q. Then $\delta(p, q') \leq \delta(p, q)$, and thus $d(p, q') \leq d(p, q)$ as well. Likewise we have $d(r, q') \leq d(r, q)$. By the additivity of the hyperbolic distance along hyperbolic lines we obtain

$$\begin{aligned} d(p, r) &\leq d(p, q') + d(q', r) \\ &\leq d(p, q) + d(q, r) \end{aligned}$$

□

We now return to the disk model. We would like to define the metric on $\mathbb{D}$ so that the incidence homomorphism

$$\sigma : \mathbb{D} \to \mathbb{H}$$

$$\sigma : p \mapsto 2\frac{p + e_2}{|p + e_2|^2} - e_2$$

becomes an isometry.

Proposition 4.6 *The function*

$$\eta(p,q) = 4\frac{|p-q|^2}{(1-|p|^2)(1-|q|^2)}$$

satisfies $\eta(p,q) = \delta(\sigma(p),\sigma(q))$.

This follows from a similar computation as in the proof of **Proposition 4.3**. Consequently, the disk model with the metric $d(p,q) = \cosh^{-1}(1+\frac{1}{2}\eta(p,q))$ becomes *isometric* to the upper half disk model.

Example 4.4 Let's compute the edge length of the edge of a hyperbolic square with 60° angles. If we put the vertices of the square at the points $(\pm a, \pm a)$, then the circle through say (a,a) and $(a,-a)$ that is perpendicular to the boundary of the unit disk will have center at a point $(b,0)$ and radius r with $(b-a)^2+a^2 = r^2$ and $b^2 = 1+r^2$ (using Pythagoras and orthogonality). We also want the vectors (a,a) and $(b,0)-(a,a)$ to make a 60° angle, which implies that

$$\frac{b-2a}{\sqrt{2}\sqrt{a^2+(a-b)^2}} = \frac{1}{2}\ .$$

This determines $r = \sqrt{2}$, $b = \sqrt{3}$, and $a = \frac{1}{2}(\sqrt{3}-1)$. Using the formula for the hyperbolic distance, we get $d((a,0),(0,-a)) = \cosh^{-1}(3) \approx 1.76$.

Exercise 4.9 Determine the hyperbolic length of the edge of a regular hyperbolic hexagon with interior angle 90°.

Exercise 4.10 Find the center and radius of the circle through the points $p_1 = (2/5, 1/5)$ and $p_2 = (3/5, -2/5)$ that is perpendicular to the unit circle.

There is another, more conceptual way to introduce the hyperbolic distance, using the *cross ratio*. Recall that we defined

$$[a, b; c, d] = \frac{d-a}{d-c} \Big/ \frac{b-a}{b-c} \ .$$

For two points p and q in the disk model or the upper half plane plane model, denote by p_0 and q_0 the intersection points of the hyperbolic line through p and q with the boundary unit circle or the real axis, ordered such that p_0, p, q and q_0 occur in this order on the hyperbolic line. Then the cross ratio $\lambda(p,q) = [p_0, p; q, q_0]$ is evidently invariant under all Möbius transformations that preserve the disk or the upper half plane, and is also preserved by any isometry between the two models. Its relation to δ is given by the following formula:

Theorem 4.7 *For two points $p, q \in \mathbb{H}$, we have*

$$\delta(p,q) = \frac{\lambda(p,q)^2}{1-\lambda(p,q)}$$

Exercise 4.11 Verify this theorem. Using the invariance of both λ and δ under Möbius transformations, you can make simplifying assumptions. For instance, you can assume that p and q lie on the imaginary axis, and even that $q = i$.

5 Geometric Group Theory

Geometric group theory studies groups using geometric tools. In this section, we will see one of the earliest examples.

5.1 Generators and Relations

A purely algebraic way to define a group is through *generators* and *relations*.

Definition 5.1 Let G be a group. A subset $\Gamma \subset G$ *generates* G if every element of G can be written as a product of elements of Γ and their inverses, called the *generators* of G.

Exercise 5.1 The additive group $(\mathbb{Z}, +)$ of integers is generated by $\Gamma = \{+1\}$. The set $\Gamma = \{2\}$ only generates the subgroup of even integers. However, $\Gamma = \{2, 3\}$ generates $\mathbb{Z}$ again, because $1 = 3 - 2$. As an exercise, can you show that $\Gamma = \{10, 12, 15\}$ also generates $\mathbb{Z}$? Can you formulate a general condition when a set of integers generates $\mathbb{Z}$?

Example 5.1

1. The cyclic groups are those groups that can be generated by a single element.
2. The symmetric group S_n is generated by the transpositions $(1,2),(2,3),...(n-1,n)$. This can be seen by applying the bubble sort algorithm: Apply one of the transpositions to any pair of consecutive elements in a permutation that are not in order. Repeat the process until you reach the identity permutation.

Definition 5.2 Given a set $\Gamma = \{a_1, ..., a_n\}$, a *word* of length N in Γ is an expression of the form

$$w = a_{i_1}^{\pm} \cdot a_{i_2}^{\pm} \cdot ... \cdot a_{i_N}^{\pm} \ .$$

A word of length 0 is denoted by 1. If Γ consists of elements of a multiplicative group $(G, \cdot)$, then a word in Γ becomes a group element. Γ generates G if and only if every group element can be written as a word in Γ. We will abbreviate $a^n = a \cdot a \cdot ... \cdot a$ for n-fold products as usual.

Example 5.2 Given a set $\Gamma = \{a_1, ..., a_n\}$, we define the *free group* generated by Γ as the set G of all words in Γ, modulo the identification of words that differ

by cancellation of products of the form $a \cdot a^{-1} = 1 = a^{-1} \cdot a$. For instance, for $\Gamma = \{a, b\}$, we consider the words

$$a \cdot b^{-1} \cdot b \cdot a^{-1} = a \cdot 1 \cdot a^{-1} = a \cdot a^{-1} = 1$$

all equal. We define the multiplication in G by juxtaposing words. For instance,

$$(a \cdot b) \cdot (b^{-1} \cdot a \cdot b) = a \cdot a \cdot b = a^2 \cdot b \ .$$

A set of generators does not describe a group completely. For instance, the additive group $\mathbb{Z} \oplus \mathbb{Z} = \{(a, b) : a, b \in \mathbb{Z}\}$ is generated by the set $\{(1, 0), (0, 1)\}$. Likewise, the free group of two generators a, b that consist of all words in a and b and their inverses are both generated by two elements, but very different as groups. In general, the elements of a generating set will satisfy *relations*: For instance, in a the cyclic group of oder n, a generator a will satisfy $a^n = 1$. Similarly, in the group $\mathbb{Z} \oplus \mathbb{Z}$, the generators $(1, 0)$ and $(0, 1)$ commute: $(1, 0) + (0, 1) = (1, 1) = (0, 1) + (1, 0)$.
In general, a relation is a word in group elements that reduces to the neutral element.

Proposition 5.1 *Let G be a group, $\Gamma = \{g_1, ...\}$ a set of generators, and R the set of all relations consisting of elements of Γ. Then R is a normal subgroup of the free group $\tilde{G}$ generated by Γ. Moreover, G is isomorphic to the quotient group $\tilde{G}/\Gamma$*

Proof: The subgroup property is trivial: If two words in elements of R reduce to the neutral element in G, so does there product and inverse. Now let $g \in \tilde{G}$ and $r \in R$. We have to show that $g^{-1}rg \in R$. But this is trivial, as r reduces to the neutral element in G and hence so does $g^{-1}rg$. The group homomorphism from $\tilde{G}$ to G, defined by sending a word in R to its product in G, is well defined on $\tilde{G}/R$, as all elements of R are sent to the 1 by definition. Thus we obtain a group homomorphism $\tilde{G}/R \to G$. It is onto, because R is a generating set of G. Finally, it is one-to-one, because whenever a word in Γ has product 1 in G it is in R. □

Example 5.3 Let

$$C_6 = \langle r \mid r^6 = 1 \rangle .$$

This is the *cyclic group* with six elements $\{1, r, r^2, ..., r^5\}$.

Example 5.4 Here is a description of the *dihedral group* D_6 in terms of generators and relations:

$$D_6 = \langle r, s \mid r^2 = s^2 = 1 \quad \text{and} \quad (rs)^6 = 1 \rangle .$$

The elements of the group D_6 are words in the letters r and s subject to the condition that the equations $rr = 1$, $ss = 1$, and $(rs)^6 = 1$ can be used to

simplify words. For instance, the word $(sr)^6$ can be reduced to 1. Here is a full list of words in D_6:

$$D_6 = \{1, r, rs, rsr, rsrs, rsrsr, rsrsrs$$
$$s, sr, srs, srsr, srsrs\} \ .$$

5.2 Cayley Graphs

A good way to visualize a group given by generators and relations is through its *Cayley graph*. We begin by reviewing the basic concepts of graph theory:

Definition 5.3 A *graph* is given by a set of vertices, together with a relation on that set that tells whether two vertices are adjacent or not. An (unordered) pair of adjacent vertices is called an *edge*. We assume that the relation is *symmetric* (so that the graph will be undirected) and that no vertex is adjacent to itself. Thus our graphs have no multiple edges between points, and all edges have two distinct end points.

This is not the only legitimate definition of a graph, but it will do for our purposes. We will now introduce the notion of a graph distance that counts how many edges lie between two vertices.

Definition 5.4 A *path* in a graph is a sequence of vertices so that consecutive vertices are adjacent. A graph is called *connected* if there is a path connecting any two vertices. The *length* of a path in graph is the number of edges in the path. The *graph distance* between two vertices is the minimum of the lengths of all paths that connect the two vertices.

Proposition 5.2 *The graph distance is a metric on the set of vertices of a connected graph.*

Proof: Symmetry and positivity of the graph distance follow immediately from the definition. To show the triangle inequality, we just note that the set of all paths connecting x to y to $z$is a subset of the set of all paths connecting x with z. □

A class of graphs that is important both in geometry and in algebra is that of the *Cayley graph* of a group.

Definition 5.5 Let $(G, \cdot, 1)$ be a group and Γ be a finite subset of G such that with $g \in \Gamma$, also $g^{-1} \in \Gamma$. We also assume that $1 \notin \Gamma$. Define a relation on G by calling $a, b \in G$ *adjacent* if there is an element $g \in \Gamma$ such that $a = gb$. Because of $\Gamma^{-1} = \Gamma$, this relation is symmetric, and the resulting graph is called the Cayley graph of G with respect to the set Γ.

Exercise 5.2 Show that the Cayley graph of a group G with respect to Γ is connected if and only if Γ generates G.

Notice that in the definition of adjacency for the Cayley graph, we multiply by g from the left. This is arbitrary, but forces us to use right multiplication in the next construction:

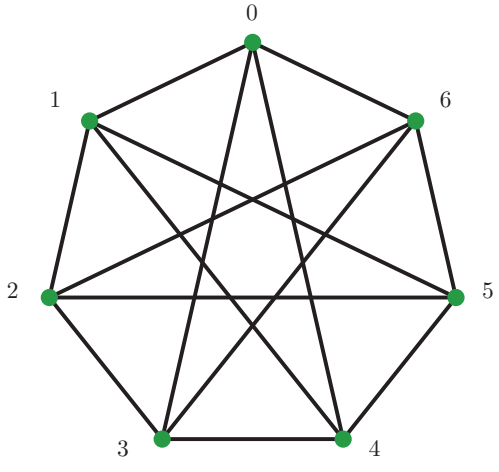
Proposition 5.3 *Let G be a group with generating set Γ The map $R_g : G \to G$, defined by $a \mapsto ag$, is an isometry of the Cayley graph of G with respect to Γ.*

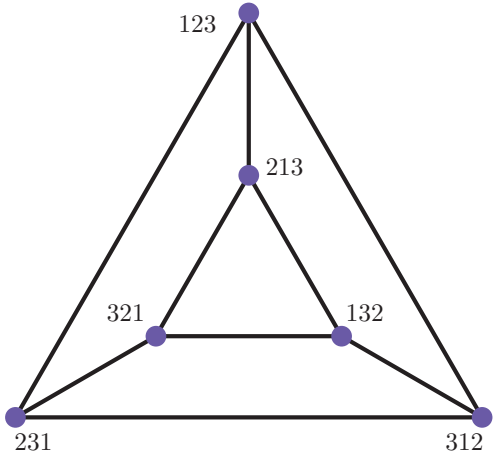
Proof: Suppose that a and b are adjacent in the Cayley graph. Then there is $\gamma \in \Gamma$ such that $a = \gamma b$, and thus $ag = \gamma bg$, so that also $R_g(a)$ and $R_g(b)$ are adjacent. Thus the lengths of paths are preserved by R_g, and thus all distances. □

This shows that the Cayley graph of a group is very symmetric. There is another source of more symmetries:

Proposition 5.4 *Let G be a group and Γ be a symmetric subset. Let $\phi : G \to G$ be a group automorphism such that $\phi(\Gamma) = \Gamma$. Then ϕ is an isometry of the Cayley graph of G with respect to Γ.*

Proof: Suppose that a and b are adjacent in the Cayley graph. Then there is $\gamma \in \Gamma$ such that $a = \gamma b$, and thus $\phi(a) = \phi(\gamma)\phi(b)$, as ϕ is a homomorphism. Because $\phi(\Gamma) = \Gamma$, in particular $\phi(\gamma) \in \Gamma$, so that $\phi(a)$ and $\phi(b)$ are adjacent. □





$G = \mathbb{Z}/7\mathbb{Z}$ and $\Gamma = \{\pm 1, \pm 3\}$ $\quad$ $G = S_3$ and $\Gamma = \{(12), (123), (321)\}$

Figure 29

Example 5.5 Let $G = S_3$ be the symmetric group of the numbers 1,2, and 3, and $\Gamma = \{(12), (123), (321)\}$. Here we use the cycle notation. Note that the inverse of a cycle is the reverse cycle, and that $(12)^{-1} = (21) = (12)$.

Exercise 5.3 We could as well have defined the Cayley graph of a group by declaring $a, b \in G$ *adjacent* if there is an element $g \in \Gamma$ such that $a = bg$. How is this graph related to our definition of the Cayley graph?

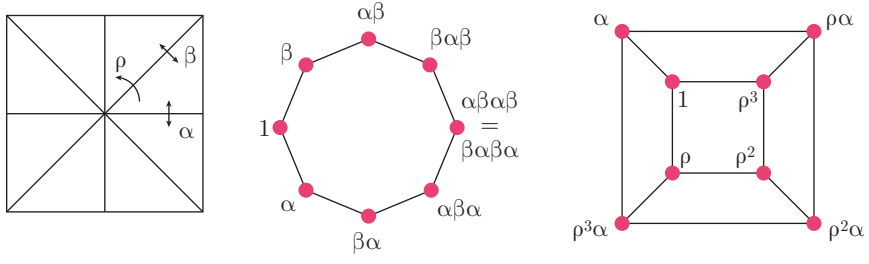


Figure 30 Two Cayley graphs for D_4

Example 5.6 Let $G = D_4$ be the *dihedral group* with four elements, i.e. the symmetry group of the square. Denote by α the reflection at the horizontal symmetry axis, by β the reflection at the main diagonal, and by ρ the 90° rotation of the square. We show that Cayley graphs of G with respect to $\Gamma_1 = \{\alpha, \beta\}$ and $\Gamma_2 = \{\alpha, \rho, \rho^{-1}\}$. Note that $\beta\alpha = \rho$, so that both $\alpha\beta\alpha\beta$ and $\beta\alpha\beta\alpha$ are 180° rotations of the square, and hence identical.

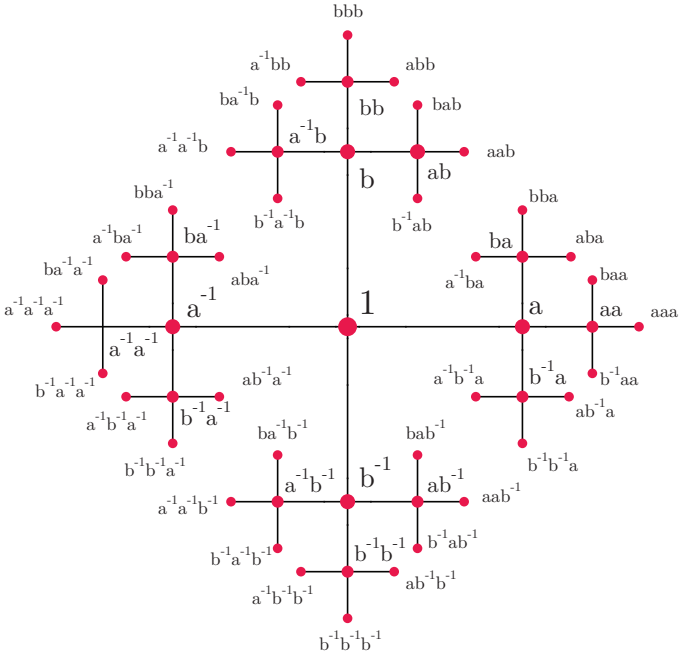


Figure 31 A Cayley graph for the free group of two elements

Example 5.7 Let G be the free group of two elements. This is the set of all words in the alphabet a and b, allowing inverses and the empty word 1. with the convention that adjacent pairs of a letter with its inverse are automatically cancelled. Examples of words are $ab^{-1}a$ or $aabba^{-1}$. We reduce $a^{-1}bb^{-1}a$ to the empty word. Words are multiplies by concatenation, i.e. $ab \cdot b^{-}1b^{-1}a = ab^{-1}a$ after cancellation. We take as $\Gamma = \{a, a^{-1}, b, b^{-1}\}$.

Example 5.8 Let's look at the infinite *grid graph*. We can define it as the Cayley graph of the abelian group $\mathbb{Z}^2$ with respect the set $\{(\pm 1, 0), (0, \pm 1)\}$. This graph can be visualized by drawing all horizontal lines $y = a$ and all vertical lines $x = a$ with $a \in \mathbb{Z}$. The vertices are then the intersection points of the lines, and edges are horizontal and vertical segments of length 1 connecting the vertices. The graph distance of this Cayley graph is also called the *taxicab metric* on $\mathbb{Z}^2$, because the grid graph resembles the street map of large cities like Manhattan.

An important feature of Cayley graphs is that each vertex has the same *valency*, i.e. the same number of edges footing at that vertex. In fact, the valency is just the number of elements in the set Γ.

Exercise 5.4 Let

$$\mathrm{SL}_2(\mathbb{F}_3) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbb{F}_3, ad - bc \equiv 1 \pmod 3 \right\}$$

be the set of 2×2 matrices with entries in the field of 3 elements and determinant 1 (modulo 3). For each such matrix, we associate the Möbius transformation

$$f(z) = \frac{az + b}{cz + d} \, .$$

You obtain 12 different Möbius transformation this way; they form a group under composition, using modular arithmetic for the basic operations involving the coefficients. They can be considered as maps of the set $\mathbb{F}_3 \cup \{\infty\}$ with the proviso that any attempted division by 0 has the result ∞, and $f(\infty) = a/c$. Draw the Cayley graph of this group with respect to the set

$$f(z) = f^{-1}(z) = \frac{2}{z}, \qquad g(z) = \frac{1}{2z + 1} \qquad \text{and} \qquad g^{-1}(z) = \frac{2z + 1}{2z} \, .$$

Exercise 5.5 Let's look at the group $G_{2,3,3}$ defined by

$$G_{2,3,3} = \langle r, s, t \mid r^2 = s^2 = t^2 = 1 \quad \text{and} \quad (rs)^2 = (st)^3 = (tr)^3 = 1 \rangle.$$

Show that its Cayley graph is given by **Figure 32**. It is clear that all vertices of the Cayley graph are trivalent. What you have to check is that all displayed

edges are indeed caused by applying generators, and for this you will have to use the relations of the group.

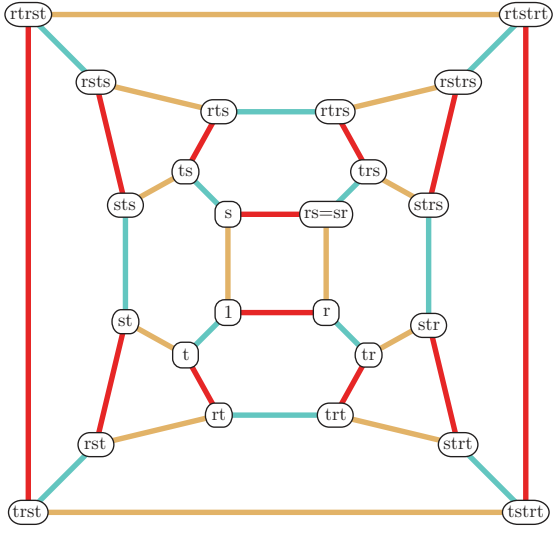


Figure 32 The Cayley graph of $G_{2,3,3}$

5.3 Dyck's Theorem

The following theorem by W. Dyck from 1882 is, according to H.S.M. Coxeter, "surely one of the most remarkable contributions of geometry to algebra."

Theorem 5.1 *Let $a, b, c \geq 2$ be integers. Let*

$$G = \langle r, s, t \mid r^2 = s^2 = t^2 = 1 \quad \text{and} \quad (rs)^a = (st)^b = (tr)^c = 1 \rangle.$$

Then G is finite if and only if $\frac{1}{a} + \frac{1}{b} + \frac{1}{c} > 1$.

To explain this, let's look at a simpler example first. Recall that we defined the *dihedral group* D_6 as

$$D_6 = \langle r, s \mid r^2 = s^2 = 1 \quad \text{and} \quad (rs)^6 = 1 \rangle .$$

This abstract group can be recognized as the symmetry group Δ_6 of a regular Euclidean hexagon. To see this, let ρ and σ be the reflections about a diagonal and a neighboring bisector of the hexagon as indicated in **Figure 33**

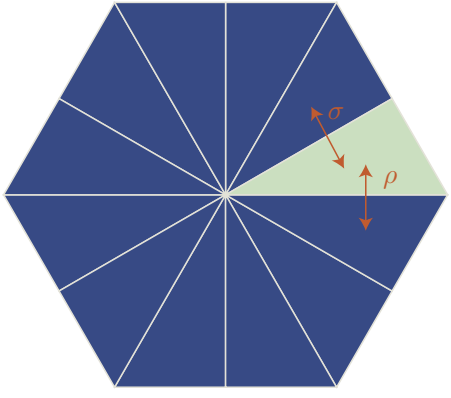


Figure 33 The Dihedral group D_6

Then ρ and σ satisfy the same relations as r and s, whence we have a group homomorphism $\delta : D_6 \to \Delta_6$. To see why δ is onto, observe that the green triangle is a *fundamental domain* for Δ_6: Any point of the hexagon can be moved by a group element into the closed triangle, and no two points in the open triangle can be moved to each other, except by the identity element of the group. This implies that Δ_6 is generated by ρ and σ, so that δ is onto. Because both Δ_6 and D_6 have the same number of elements, δ is in fact an isomorphism.

This argument is at the heart of the proof of the theorem. The first step is to construct a reflection group $\Gamma_{a,b,c}$. To this end, we will need all three two-dimensional geometries: the Euclidean plane, the sphere, and the hyperbolic plane.
The reason for this is that we will need triangles where the angles do not add up to 180°. They are supplied by the following elementary fact:

Proposition 5.5 *Let $\alpha, \beta, \gamma > 0$ be angles such that $\alpha + \beta + \gamma < 2\pi$. Then there is a triangle with these angles in the Euclidean plane, the sphere, or the hyperbolic plane, depending on whether $\alpha + \beta + \gamma$ is $= \pi$, $> \pi$, or $< \pi$. The triangle is unique up to a similarity in the Euclidean case and up to a congruence in the spherical and hyperbolic case.*

However, just having triangles with the correct angles does not immediately prove the theorem. We pause for a minute to see what difficulties we still face:

Example 5.9 First, it is not clear that a given triangle will tile the sphere, Euclidean plane, or hyperbolic plane. In fact, how should it know? For example, we could take a curved rectangle whose edges are circular arcs and try to tile plane with it. This is easy near a given tile, but the tiles can interfere in the large, as can be seen in **Figure 35**. Here, the curved rectangles have as edges the radii and arcs of concentric circles.

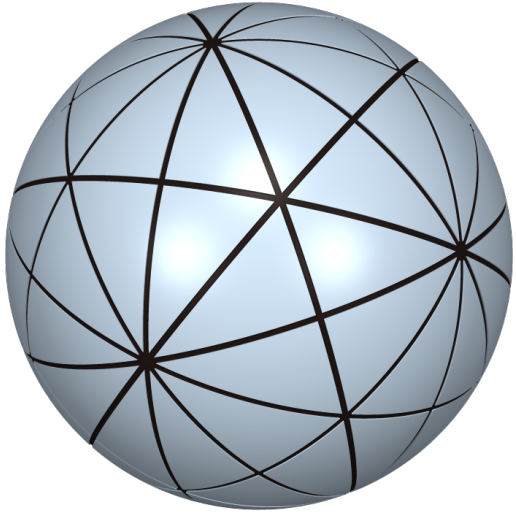


Figure 34 The $(2, 3, 5)$-tiling of the sphere

For triangles however we have the intuitively clear

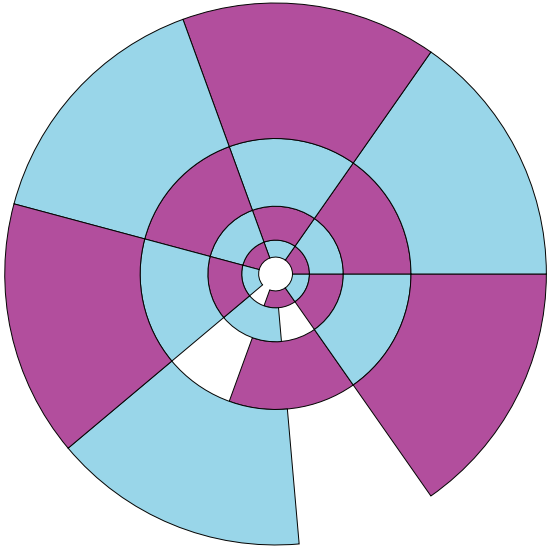


Figure 35 Curved rectangles don't always tile

Theorem 5.2 *The unique triangle with angles $\alpha, \beta, \gamma > 0$ and $\alpha+\beta+\gamma < 2\pi$ tiles the Euclidean plane, the sphere, or the hyperbolic plane, depending on whether $\alpha + \beta + \gamma$ is $= \pi$, $> \pi$, or $< \pi$.*

Example 5.10 The second issue is that even if we can use a triangle to tile, this alone does not establish a group isomorphism. For instance, the $(2, 2, 4)$ triangle does not only tile the Euclidean plane, but also (for instance) the Euclidean torus obtained by identifying opposite edges in the rectangle of **Figure 37**:
What makes the sphere, the Euclidean plane, and the hyperbolic plane special is that they are *simply connected*, which means that any closed curve can be shrunk to a point.

Now let a, b, c be as in our theorem, and define $\alpha = \frac{\pi}{a}$, $\beta = \frac{\pi}{b}$, $\gamma = \frac{\pi}{c}$. Reflecting the corresponding Coxeter triangle about its edges will create a so-called (a, b, c)-tiling of the Euclidean plane, the sphere, or the hyperbolic plane. Denote by $\Gamma_{a,b,c}$ the group generated by the reflections at the three edges of the triangle. If we can show that $G_{a,b,c}$ is isomorphic to $\Gamma_{a,b,c}$, we are done with proving the theorem, because the number of elements of $\Gamma_{a,b,c}$ is equal to the number of tiles used. This number is finite if and only if we are tiling the sphere.
Let's fix one (α, β, γ)-triangle within its tiling, and let ρ, σ, τ be the reflections about the edges opposite to the angles β, γ and α. Again these group elements

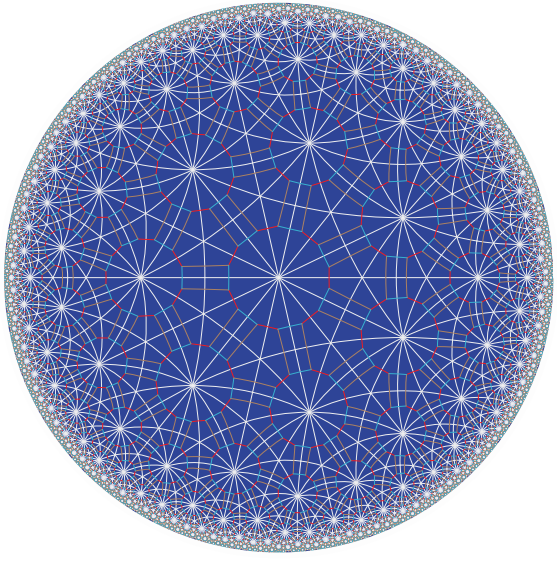


Figure 36 The $(2, 3, 7)$-tiling of the hyperbolic plane together with the Cayley graph of $\Gamma_{2,3,7}$

of $\Gamma_{a,b,c}$ satisfy the same relations as r, s, t. Thus we get a group homomorphism $\delta : G_{a,b,c} \to \Gamma_{a,b,c}$. Any of the triangles of the tiling is a fundamental domain for $\Gamma_{a,b,c}$; this implies that δ is onto.

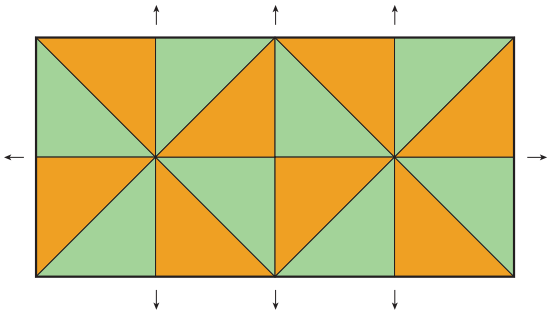


Figure 37 Tiling a torus with $(2, 4, 4)$-triangles

To see that δ is one-to-one, we have to show that any word in ρ, σ, τ that represents the identity in $\Gamma_{a,b,c}$ can be reduced to the identity by using the relations in $G_{a,b,c}$. Let $v_1 v_2 \cdots v_n$ be such a word. There are two ways to understand its action on the plane, sphere or disk: By reading it from right to left, we consecutively apply the reflections about the corresponding edges of the triangle we have fixed.
But we can also use the following recipe to determine the image of the initial triangle: Color the edges of all triangles in the tiling red, sienna, and turquoise, so that red corresponds to r, sienna to s, and turquoise to t. Choose a point P_1 in the initial triangle.

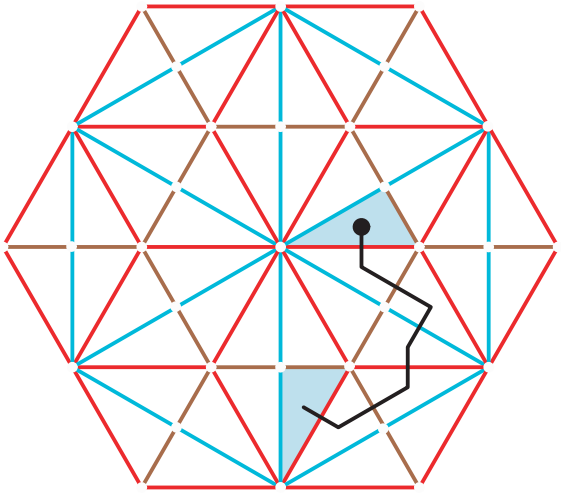


Figure 38 Reflection path of the word *rstrsr* in $\Gamma_{2,3,6}$

Obtain P_2 by reflecting P_1 about the edge corresponding to the letter v_1, then P_3 by reflecting P_2 about the edge corresponding to the letter v_2 of the triangle that contains P_2, etc., until we reach P_{n+1}. For instance, in the figure above, we see the path corresponding to the word *rstrsr*.

To see why we get the same point P_{n+1} both ways, let ϕ_k be the congruence that maps the initial triangle Δ_0 to the triangle Δ_k at the endpoint of the path corresponding to the word $v_1...v_k$. We will prove by induction that $\phi_k = \delta(v_1...v_k)$. Clearly $\phi_1 = \delta(v_1)$, and we claim that $\phi_{k+1} = \phi_k \circ \delta(v_{k+1})$. By definition, $\phi_{k+1}(\Delta_0) = \Delta_{k+1}$. On the other hand, $\phi_k(\Delta_0) = \Delta_k$ implies that ϕ_k maps $\delta(v_{k+1})(\Delta_0)$ to Δ_{k+1}. Thus ϕ_{k+1} and $\phi_k \circ \delta(v_{k+1})$ have the same effect on Δ_0.
In other words: A word in the generators of a group determines a path in the *dual graph* of the tiling, which is the *Cayley graph* of the group with respect to the generators, so that the group element represented by the word maps the initial point of the path to its end point.
Now we are assuming that $v_1 v_2 \cdots v_n = 1$ in $\Gamma_{a,b,c}$. This means that the path constructed above must close up at the initial triangle. As all our model spaces are *simply connected*, this path can be shrunk to the point P_0. Such a *homotopy* can be realized as a combination of finitely many elementary homotopies, which are:

i) replace an edge together with a backtracking edge by the constant path;

ii) replace a cycle of edges around one vertex by the constant path.

Both these elementary homotopies correspond to reductions of the word by group relations. Thus our word can be reduced to 1 using the relations, showing that ϕ is one-to-one. That one can do this indeed with finitely many elementary homotopies hinges on the fact that a homotopy will have support in a ball of some finite radius.

This ball has finite area and can therefore be covered with finitely many of the triangles of the tiling. Note that this is exactly what goes wrong in **Example 5.9**. In case the group is finite, the number of triangles in the tiling is equal to the number of group elements, as for any two labeled triangles there is a unique isometry mapping one triangle to the other. Thus we have

Corollary 5.1 *If* $d = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} > 1$, *then* $G_{a,b,c}$ *has order* $\frac{4}{d-1}$.

Proof: The area of a spherical triangle with angles π/a, π/b, π/c is $(d-1)\pi$. On the other hand, the entire sphere has area 4π. □

5.4 Ford Circles

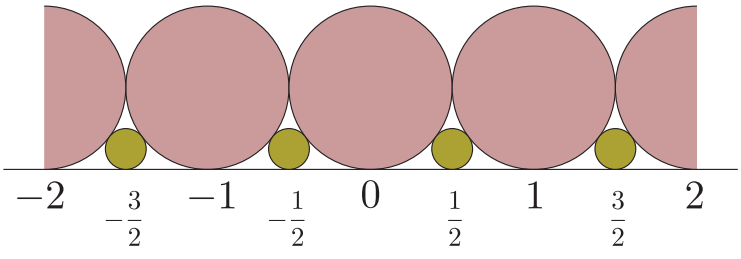


Figure 39 Placing Circles on the real axis

For every reduced rational number $\frac{p}{q}$, we define the *Ford circle* $F_{p/q}$ as the circle with center at $\left(\frac{p}{q}, \frac{1}{2q^2}\right)$ and radius $\frac{1}{2q^2}$.
To explore this collection of circles, let's start with the circles $F_{p/1}$ where the denominator of the rational number is 1, meaning it is actually an integer. We get circles of radius $\frac{1}{2}$ that touch the real axis and neighboring circles. Adding the circles of the second generation where p/q has a 2 in the denominator, suggests that these circles touch the circles from the first generation. We can verify quickly that $F_{0/1}$ touches $F_{1/2}$. This is because the centers have distance

$$\left|\left(0,\frac{1}{2}\right)-\left(\frac{1}{2},\frac{1}{8}\right)\right|=\sqrt{\frac{1}{4}+\frac{9}{64}}=\frac{5}{8}$$

which equals the sum of their radii. This is no coincidence. When adding more circles $C_{p/q}$ each new generation touches circles from previous generations.



Figure 40 More Ford circles

To see this we first determine when two general Ford circles touch:

Lemma 5.1 *Two different Ford circles $F_{a/b}$ and $F_{c/d}$ touch if and only if*

$$|ad - bc| = 1 \ ,$$

and are disjoint otherwise.

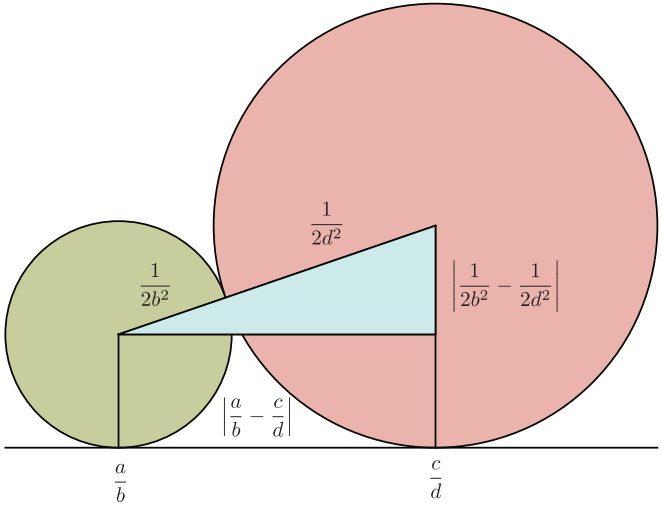


Figure 41 When do Ford circles touch?

Proof: For two circles $F_{r_1}(p_1)$ and $F_{r_2}(p_2)$ to touch, the sum of their radii must equal the distance of their centers. Using Pythagoras, we measure the gap between two Ford circles $F_{a/b}$ and $F_{c/d}$ by

$$|p_1 - p_2|^2 - (r_1 + r_2)^2 = \left(\frac{a}{b} - \frac{c}{d}\right)^2 + \left(\frac{1}{2b^2} - \frac{1}{2d^2}\right)^2 - \left(\frac{1}{2b^2} + \frac{1}{2d^2}\right)^2$$
$$= \frac{(ad - bc)^2 - 1}{(bd)^2} \ .$$

For different Ford circles, the right hand side is always non-negative, and zero if and only if $(ad - bc)^2 = 1$. □

This allows us to find the Ford circle sitting between two touching Ford circles:

Lemma 5.2 *If the two Ford circles $F_{a/b}$ and $F_{c/d}$ touch, then $F_{(a+c)/(b+d)}$ touches them as well.*

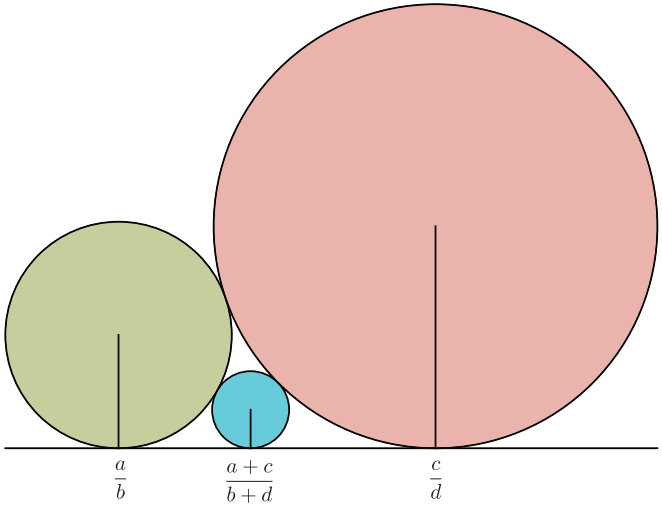


Figure 42 Fitting Ford circles in between

Proof: $F_{a/b}$ and $F_{(a+c)/(b+d)}$ touch because

$$|a(b+d) - b(a+c)| = |ad - bc| = 1 \ .$$

The same argument applies to $F_{c/d}$ and $F_{(a+c)/(b+d)}$. □

The addition of fractions has never been easier.

We can also locate the point where two Ford circles touch:

Lemma 5.3 *Let $F_{a/b}$ and $F_{c/d}$ be two touching Ford circles. Then the semicircle that passes through $\frac{a}{b}$ and $\frac{c}{d}$ passes through the point where the Ford circles touch.*

Proof: Draw the tangent to the two Ford circles at the point p where they touch. This tangent will intersect the real axis in a point q. Then $|\frac{a}{b} - q| = |p - q| = |\frac{c}{d} - q|$, so that the semicircle centered at q with radius $p - q$ will pass through $\frac{a}{b}$ and $\frac{c}{d}$. □

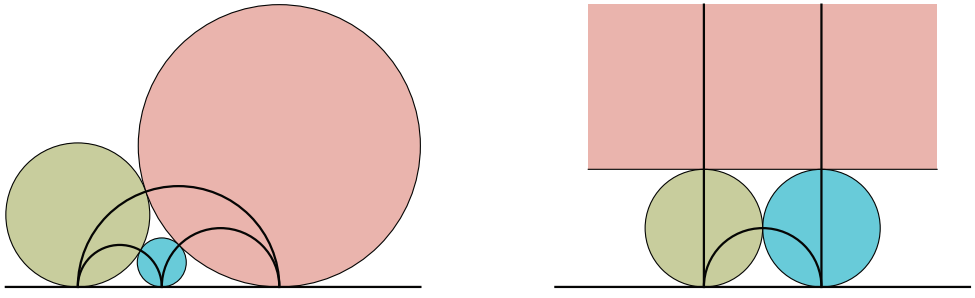


Figure 43 Ford circles and ideal triangles

Definition 5.6 We say that $\frac{p}{q}$ and $\frac{p'}{q'}$ form a *Farey pair* if $|pq' - p'q| = 1$. Three rational numbers form a *Farey Triangle* if any two of them form a Farey pair.

Definition 5.7 An *ideal triangle* is given by three real numbers and the semi-circles orthogonal to the real axis connecting them in the upper half plane.

Clearly, each Farey triangle defines an ideal triangle by using the rational numbers as vertices. We can reconstruct a Farey triangle from its ideal triangle. More generally, we have:

Lemma 5.4 *For any ideal triangle, there are three unique circles touching each other and the real axis at the vertices of the ideal triangle.*

Proof: Given an ideal triangle Δ, apply to it a Möbius transformation that sends its vertices to 0, 1, and ∞. Its edges will be two vertical half lines at 0 and 1, and the semicircle connecting 0 with 1. We consider this an ideal triangle Δ' with one vertex at ∞. If we had three circles touching each other and the real axis in the vertices of the original triangle Δ, these circles would be mapped by the Möbius transformation to the horizontal line $\operatorname{Im} z = 1$ and the two circles that touch this line and the real axis in 0 and 1, respectively. Therefore we can find the circles belonging to Δ by applying the inverse of the Möbius transformation to the circles belonging to the triangle Δ'. □

We can apply this to show:

Lemma 5.5 *Let $\phi \in PSL_2(\mathbb{Z})$ be a Möbius transformation. Then ϕ takes Ford circles to Ford circles.*

Proof: We first observe that ϕ takes Farey pairs to Farey pairs: If

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

then

$$\phi\left(\frac{p}{q}\right) = \frac{ap+bq}{cp+dq} \ .$$

If $\frac{p}{q}$ and $\frac{p'}{q'}$ are a Farey pair, then so are $\phi\left(\frac{p}{q}\right)$ and $\phi\left(\frac{p'}{q'}\right)$, because

$$\begin{aligned}(ap+bq)(cp'+dq') - (ap+bq)(cp'+dq') &= (ad-bc)pq' - (ad-bc)p'q \\ &= (ad-bc)(pq'-p'q) \\ &= \pm 1 \ .\end{aligned}$$

Now take any Ford circle and complete it to a Farey triangle Φ by adding two more Ford circles. Its corresponding ideal triangle Δ is mapped by ϕ to another

ideal triangle Δ', and the Ford circles belonging to Δ are sent to the circles belonging to Δ'. The vertices of Δ' are rational numbers forming Farey pairs, because ϕ preserves Farey pairs. Thus they form a Farey triangle Φ' whose corresponding ideal triangle is equal to Δ'. As the circles of an ideal triangle determine the touching circles, the circles of the Farey triangle Φ' are the same as the circles belonging to Δ'. $\square$