Geometry in two dimensions

by Ian Biringer

Contents

Introduction

This course is intended for students who have a background in multivariable calculus and some experience in proof-based mathematics. Throughout the course, we will assume a basic framework of Euclidean geometry, e.g. that there are notions of distance, angle, etc... that satisfy the usual properties. In the first few sections, we will spend some time building up a coordinate-dependent geometry on top of this.

Some of the material is presented at an intuitive rather than rigorous level, but for the most part the arguments given in class and required on the homework will be rigorous. In particular, our discussion of the topology of surfaces is necessarily intuitive, but most of the time we will consider only surfaces built from 'gluing diagrams', which can be treated rigorously. Students with a background in analysis or point set topology are encouraged to fill in the details where appropriate.

Image credits: The current version of this book is for personal use, so I haven't given appropriate credit for some of the images. I made most of them using Ipe, and handdrew a couple more, but the rest of the images were taken from the Internet. I'll add image citations in a later version, but for the moment if you see an image that you would like removed or cited, please just let me know.

CHAPTER 1

Euclidean geometry

1.1. Distance in $\mathbb{R}^n$

We work in n-dimensional Euclidean space, $\mathbb{R}^n$. Points in $\mathbb{R}^n$ are represented in coordinates as $x = (x_1, \ldots, x_n)$, where $x_1, \ldots, x_n$ are real numbers, and adding subscripts to a point in $\mathbb{R}^n$ will always represent its coordinates. Although we will only really care about $n = 2, 3$, it makes sense to develop the theory in general.

One of the major themes of this course will be the notion of 'distance'. In $\mathbb{R}^2$, a formula for the distance between two points comes from the Pythagorean theorem.

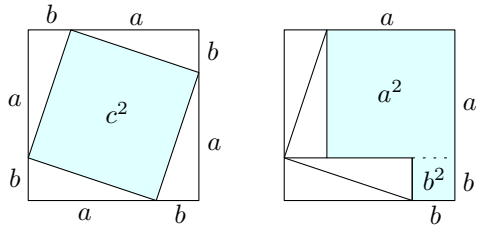
THEOREM 1.1.1 (Pythagoras). *If a right triangle has side lengths a, b, c, where c is the hypotenuse, then $a^2 + b^2 = c^2$.*

Accordingly, the distance $d(p, q)$ between points $p, q \in \mathbb{R}^2$ should be

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2},$$

applying the Pythagorean theorem to the right triangle with vertices p, q and (q_1, p_2). There are a huge number of proofs of the Pythagorean theorem; here is a beautiful geometric proof that is usually attributed to Chinese antiquity.

PROOF. Draw a square with side lengths $a + b$, and four enclosed triangles:



The area of the blue region is c^2. Rotating two of the triangles changes the blue region into a union of two rectangles, with areas a^2 and b^2. Thus $c^2 = a^2 + b^2$. □

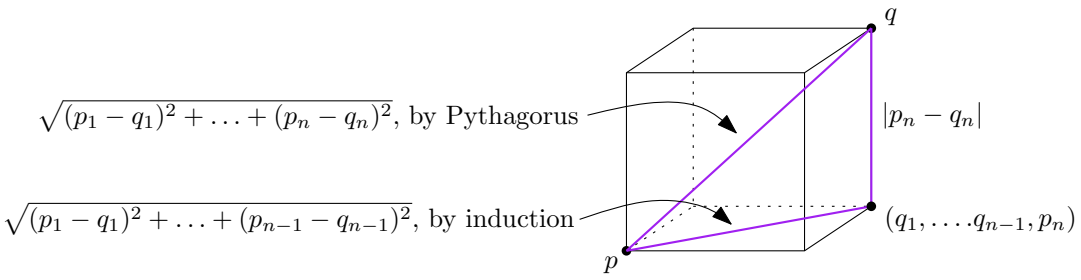
The distance between two points in $\mathbb{R}^n$ can be calculated inductively by the same means. We claim that if $p, q \in \mathbb{R}^n$, then

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + \cdots + (p_n - q_n)^2}.$$

Assuming that the analogous formula holds in $\mathbb{R}^{n-1}$, suppose that $p, q \in \mathbb{R}^n$. The plane defined by fixing the last coordinate in $\mathbb{R}^n$ to be p_n is a copy of $\mathbb{R}^{n-1}$, so

$$d\big(p, (q_1, \ldots, q_{n-1}, p_n)\big) = \sqrt{(p_1 - q_1)^2 + \cdots + (p_{n-1} - q_{n-1})^2},$$

There is a right triangle with vertices p, q and $(q_1, \ldots, q_{n-1}, p_n)$, as pictured below, and the Pythagorean theorem gives the formula for $d(p, q)$ described above.



The distance formula is best understood with the assistance of a tool from multivariable calculus. Recall that if $v, w \in \mathbb{R}^n$, their *dot product* is the real number

$$v \cdot w = v_1 w_2 + \cdots + v_n w_n$$

PROPOSITION 1.1.2. *For $v, w, u \in \mathbb{R}^n$, the dot product satisfies the following properties:*

- (Commutativity) $v \cdot w = w \cdot v$,
- (Distributivity) $v \cdot (w + u) = v \cdot w + v \cdot u$,
- (Scalars come out) $v \cdot (rw) = r(v \cdot w)$, *for $r \in \mathbb{R}$.*

PROOF. Exercise, using the analogous properties of arithmetic of real numbers. □

Note that using commutativity, one can also distribute the dot product over an addition or extract a scalar from the first input, not just the second.

Using the dot product, we may define the *length* of a vector $v \in \mathbb{R}^n$ by

$$|v| = \sqrt{v \cdot v}$$

Note that $|v|$ is exactly the distance from the origin to the head of v, so length is compatible with our definition of distance from before. Furthermore, we have

$$d(v, w) = |v - w|, \ \forall v, w \in \mathbb{R}^n.$$

The dot product has an important geometric interpretation.

THEOREM 1.1.3. *If the angle between two vectors $v, w \in \mathbb{R}^n$ is θ, then $v \cdot w = |v||w| \cos\theta$.*

In particular, $v \cdot w = 0$ if and only if v, w are perpendicular.

PROOF. Let us temporarily write $v \star w = |v||w| \cos\theta$, so that the theorem claims that $\star$ is the same as the dot product. If $e_1 = (1, 0, \ldots, 0), \ldots, e_n = (0, \ldots, 0, 1)$, then

$$e_i \star e_j = e_i \cdot e_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{1}$$

For any two distinct such vectors are perpendicular, so the cosines of their angles vanish, while a quick computation shows that their dot products also vanish.

Now, observe that $\star$ satisfies the three properties of Proposition 1.1.2. Commutativity is easy, since as cos is an even function,

$$|v|\,|w| \cos\theta = |w|\,|v| \cos(-\theta).$$

Scalars come out of $\star$ since

$$|v|\,|rw| \cos(\theta) = |v|\sqrt{(rw) \cdot (rw)} \cos(\theta) = r|v|\sqrt{w \cdot w} \cos(\theta) = r|v|\,|w| \cos(\theta),$$

so the point is to prove the distributive law, which we leave as an exercise. Assuming this, we then compute

$$\begin{aligned} v \star w &= \left(\sum_i v_i e_i\right) \star \left(\sum_j w_j e_j\right) \\ &= \sum_{i,j} (v_i e_i) \star (w_j e_j), \quad \text{by distributivity,} \\ &= \sum_{i,j} v_i w_j e_i \star e_j, \quad \text{by pulling out scalars,} \\ &= \sum_i v_i w_i, \quad \text{by Equation (1),} \\ &= v \cdot w. \end{aligned}$$

□

EXERCISE 1.1.4. If $v \in \mathbb{R}^n$, define a map $\text{proj}_v : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ by

$$\text{proj}_v(w) = w \cdot v \frac{v}{|v|^2}.$$

Show that $w - \text{proj}_v(w)$ is perpendicular to v. Then show that $\text{proj}_v(w)$ is the closest point to w on the line $\{tv \mid t \in \mathbb{R}\}$. *Hint: use the Pythagorean theorem.*

COROLLARY 1.1.5 (Law of cosines). *Suppose that a triangle has side lengths a, b, c and that the angle opposite c is θ. Then $c^2 = a^2 + b^2 - 2ab\cos\theta$.*

Note: 'side length' here just means the distance between the endpoints.

PROOF. Regard the sides of the triangle as vectors a, b, c as pictured in Figure 1, so that what we want to prove is $|c|^2 = |a|^2 + |b|^2 - 2|a|\,|b| \cos\theta$. Then $c = a - b$, so

$$|c|^2 = c \cdot c = (a - b) \cdot (a - b) = a \cdot a - 2a \cdot b + b \cdot b = |a|^2 + |b|^2 - 2\,|a|\,|b| \cos\theta. \quad \square$$

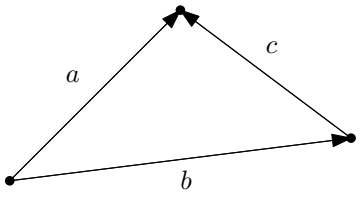


FIGURE 1. The triangle in the law of cosines.

COROLLARY 1.1.6 (The triangle inequality). *Suppose that $p, q, z \in \mathbb{R}^n$. Then*

$$d(p, z) \leqslant d(p, q) + d(q, z),$$

with equality if and only if p, q, z are collinear, with q between p, z.

PROOF. Let θ be the angle at q in the triangle pqz. By the law of cosines,

$$\begin{aligned} d(p,z)^2 &= d(p,q)^2 + d(q,z)^2 - 2d(p,q)d(q,z)\cos\theta \\ &\leqslant d(p,q)^2 + d(q,z)^2 + 2d(p,q)d(q,z), \text{ since } \cos\theta \geqslant -1 \\ &= (d(p,q) + d(q,z))^2, \end{aligned}$$

so taking square roots gives the triangle inequality. Furthermore, we have equality above if and only if $\cos\theta = -1$, in which case $\theta = \pi$. This means exactly that q lies on the line segment between p and z. □

DEFINITION 1.1.7. If X is a set, a *metric* on X is a function $d : X \times X \longrightarrow \mathbb{R}$ that satisfies, for all $p, q, z \in X$, the following three properties:

(a) $d(p, q) \geqslant 0$, with $d(p, q) = 0$ if and only if $p = q$,
(b) $d(p, q) = d(q, p)$,
(c) $d(p, z) \leqslant d(p, q) + d(q, z)$.

PROPOSITION 1.1.8. *$d(p, q) = \sqrt{(p_1 - q_1)^2 + \cdots + (p_n - q_n)^2}$ defines a metric on $\mathbb{R}^n$.*

PROOF. The triangle inequality is proved above. For the first property,

$$d(p, q) = (p_1 - q_1)^2 + \ldots + (p_n - q_n)^2$$

is a sum of nonnegative numbers. It is therefore nonnegative, and is zero exactly when all of the terms are zero, i.e. when $p = q$. The second property is obvious and the third is the triangle inequality, which we proved above. □

The three properties defining a metric are a minimum that one might require in order to make d behave like our usual notion of distance. There are a number of other metrics on $\mathbb{R}^n$, for instance

$$\begin{aligned} d_{max}(p, q) &= \max\{|p_1 - q_1|, \cdots, |p_n - q_n|\}, \\ d_{sum}(p, q) &= |p_1 - q_1| + \cdots + |p_n - q_n|. \end{aligned}$$

EXERCISE 1.1.9. Verify the three metric properties for d_{max}.

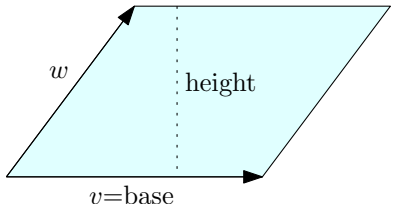
Some of these can really be viewed as physical distance functions, in some sense. For example, in $\mathbb{R}^2$ the metric $d_{sum}(p,q) = |p_1 - q_1| + |p_2 - q_2|$ is called *Manhattan distance* – the distance between two points is the sum of the horizontal and vertical distance, which is how far one must travel if one is constrained to roads and cannot cut diagonally. A more abstract example of a metric is the *discrete metric* on (any) set X, in which the distance between two distinct points $p, q \in X$ is always 1.

EXERCISE 1.1.10. If $v, w \in \mathbb{R}^2$, the *determinant* of the 2×2 matrix $(v\ w)$ is

$$\det \begin{pmatrix} v_1 & w_1 \\ v_2 & w_2 \end{pmatrix} = v_1 w_2 - w_1 v_2.$$

Note that $v_1 w_2 - w_1 v_2 = v^\perp \cdot w$, where $v^\perp = (-v_2, v_1)$. The vector $v^\perp$ is obtained from v by rotating $\pi/2$ counterclockwise; to see this, note that $v \cdot v^\perp = 0$, so they make a right angle, and by inspection one can just check that the angle from v to $v^\perp$ is $\pi/2$ counterclockwise rather than clockwise.

(a) Mention why the counterclockwise angle from v to w is between 0 and π if and only if $\det(v\ w) \geqslant 0$.
(b) The area of a parallelogram is the length of its base times its height. Show that if $v, w \in \mathbb{R}^2$, the area of the parallelogram spanned by v, w is $|\det(v\ w)|$.



EXERCISE 1.1.11 (SSS). Show that if two triangles have all the same side lengths, they have all the same interior angles as well.

EXERCISE 1.1.12 (SSA). Show that if two triangles both have two sides of lengths a and b that meet angle θ, then the remaining sides have the same length as well.

1.2. Isometries, especially of $\mathbb{R}^2$

A recurring theme in this course will be the study of 'rigid motions'. Formally, an *isometry* of a metric space X is a bijection $f : X \longrightarrow X$ that preserves distances:

$$d(f(x), f(y)) = d(x, y), \quad \forall x, y \in X.$$

EXERCISE 1.2.1 (Isometries form a group). Show that the identity map $id : X \longrightarrow X$ is always an isometry, that the composition of two isometries is an isometry, and that the inverse of an isometry is an isometry.

In this section, we will mostly be interested in isometries of $\mathbb{R}^2$.

DEFINITION 1.2.2 (Translations). If $v \in \mathbb{R}^2$, the *translation* by v is the map

$$T_v : \mathbb{R}^2 \longrightarrow \mathbb{R}^2, \quad T_v(x) = x + v.$$

Translations are bijections, as $T_v^{-1} = T_{-v}$. They also preserve distances, as

$$d(T_v(x), T_v(y)) = |x + v - (y + v)| = |x - y| = d(x, y),$$

and therefore are isometries. One should imagine a translation as rigidly shifting the plane $\mathbb{R}^2$ in the direction indicated by v.

DEFINITION 1.2.3 (Rotations). If $p \in \mathbb{R}^2$ and $\theta \in [0, 2\pi)$, the *rotation* around p by angle θ is defined to be the map

$$O_{p,\theta} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2,$$

such that $O_{p,\theta}(p) = p$, and otherwise $O_{p,\theta}(x)$ is the unique point such that

$$d(p, x) = d(p, O_{p,\theta}(x))$$

and the angle from px to $pO_{p,\theta}(x)$ is θ.

EXERCISE 1.2.4. Using the law of cosines, show that rotations are isometries.

Rotations around the origin can be expressed in coordinates as follows: representing points in $\mathbb{R}^2$ as column vectors we have

$$O_{0,\theta} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1\cos\theta - x_2\sin\theta \\ x_1\sin\theta + x_2\cos\theta \end{pmatrix}.$$

To see why this is true, first note that

$$\begin{aligned} d\left(0, O_{0,\theta}(x)\right) &= \left| \begin{pmatrix} x_1\cos\theta - x_2\sin\theta \\ x_1\sin\theta + x_2\cos\theta \end{pmatrix} \right| \\ &= \sqrt{(x_1\cos\theta - x_2\sin\theta)^2 + (x_1\sin\theta + x_2\cos\theta)^2} \\ &= \sqrt{(x_1^2 + x_2^2)\cos^2\theta + (x_1^2 + x_2^2)\sin^2\theta} \\ &= \sqrt{x_1^2 + x_2^2} \\ &= d\left(0, x\right). \end{aligned}$$

Next, we compute the angle ψ between the vectors x and $O_{p,\theta}(x)$ using Theorem 1.1.3:

$$\begin{aligned} \cos\psi &= \frac{x \cdot O_{p,\theta}(x)}{|x|\,|O_{p,\theta}(x)|} \\ &= \frac{x_1(x_1\cos\theta - x_2\sin\theta) + x_2(x_1\sin\theta + x_2\cos\theta)}{\sqrt{x_1^2 + x_2^2}\sqrt{x_1^2 + x_2^2}} \\ &= \cos\theta. \end{aligned}$$

Therefore, $\psi = \pm\theta$.

EXERCISE 1.2.5. Show that $\psi = \theta$. *(Use Exercise 1.1.10).*

Here is a cool application of this description in coordinates of rotations. Geometrically, it is clear that rotating around 0 first by angle θ, then by angle ψ gives a rotation by angle $\theta + \psi$. However, we can also compute the composition in coordinates. As function composition is just matrix multiplication, we compute:

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{pmatrix} = \begin{pmatrix} \cos\theta\cos\psi - \sin\theta\sin\psi & -\cos\theta\sin\psi - \sin\theta\cos\psi \\ \sin\theta\cos\psi + \cos\theta\sin\psi & \cos\theta\cos\psi - \sin\theta\sin\psi \end{pmatrix}$$

As this must be equal to the rotation matrix by angle $\theta + \psi$, we have:

$$\begin{aligned} \sin(\theta + \psi) &= \sin\theta\cos\psi + \cos\theta\sin\psi \\ \cos(\theta + \psi) &= \cos\theta\cos\psi - \sin\theta\sin\psi, \end{aligned}$$

so this gives a proof of the angle sum formulas for cos and sin!

So, how can we find coordinate descriptions for rotations that are *not* around the origin? For this, we use the convenient identity

$$O_{p,\theta} = T_p \circ O_{0,\theta} \circ T_{-p}, \tag{2}$$

which implies that

$$O_{p,\theta}(x) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} (x - p) + p.$$

EXERCISE 1.2.6. Draw a picture, and prove that (2) is true.

The identity above is an example of a general philosophy. When we have isometries f and g, the isometry $f \circ g \circ f^{-1}$ is called the *conjugate* of g by f. Imagine you're playing a videogame where the screen is a view of the ground from above, and your character is in the center of the screen. Suppose g represents how the terrain moves when you press the up arrow. For example, maybe the terrain moves down a few pixels, indicating that your character has moved up. Now imagine that you pick up your monitor and rotate it counterclockwise by 90 degrees. If you now press the up arrow, the terrain will seem to move to the right rather than down. If f is the 90 degree rotation, then this movement to the right is $f \circ g \circ f^{-1}$.

The general philosophy is that a conjugate $f \circ g \circ f^{-1}$ has the same type (e.g. translation, rotation) as g, but its defining data (e.g. direction of translation, point of rotation) has been moved by f. A precise statement along these lines is:

EXERCISE 1.2.7. Given $f : X \longrightarrow X$, let $\mathrm{Fix}(f) = \{x \in X \mid f(x) = x\}$. Show that

$$\mathrm{Fix}(f \circ g \circ f^{-1}) = f(\mathrm{Fix}(g)),$$

whenever f, g are both bijections $X \longrightarrow X$.

For another couple of example, try convincing yourself that

$$O_{0,\theta} \circ T_v \circ O_{0,\theta}^{-1} = T_{O_{p,\theta}(v)}, \text{ and } T_v \circ T_w \circ T_{-v} = T_w.$$

You should verify the equations using the coordinate descriptions given above, but also try to give a intuitive explanation as in the videogame example.

DEFINITION 1.2.8 (Reflections). If ℓ is a line in $\mathbb{R}^2$, the *reflection* through ℓ is

$$R_\ell : \mathbb{R}^2 \longrightarrow \mathbb{R}^2,$$

where $R_\ell(x) = x$ whenever $x \in \ell$ and otherwise $R_\ell(x)$ is the unique point such that the line segment from x to $R_\ell(x)$ is perpendicularly bisected by ℓ.

EXERCISE 1.2.9. Show that reflections are isometries.

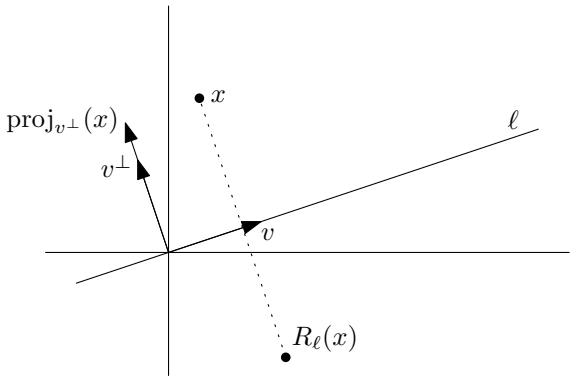
When ℓ goes through the origin, we can write it as $\ell = \{tv \mid t \in \mathbb{R}\}$ for some $v \in \mathbb{R}^2$. In this case, the reflection R_ℓ has the following description in coordinates:

$$R_\ell(x) = x - 2\,\mathrm{proj}_{v^\perp}(x), \text{ where } v^\perp = \begin{pmatrix} -v_2 \\ v_1 \end{pmatrix}.$$

Recall from Exercise 1.1.4 that $\mathrm{proj}_{v^\perp}(x) = x \cdot v^\perp \frac{v^\perp}{|v^\perp|^2}$ is the closest point to x along the line spanned by $v^\perp$. Also, $v^\perp$ is just v rotated by $\pi/2$ counterclockwise.

For general lines, we may write $\ell = \{p + tv \mid t \in \mathbb{R}\}$, in which case

$$R_\ell(x) = T_p \circ R_{\{tv \mid t\in\mathbb{R}\}} \circ T_{-p} = x - 2\,\mathrm{proj}_{v^\perp}(x - p).$$



EXERCISE 1.2.10. Show that reflections are isometries, but now using the coordinate description.

Here are a couple more exercises to help you get the feel for isometries. They are phrased in $\mathbb{R}^n$ rather than $\mathbb{R}^2$, simply because there is no difference in the proof.

EXERCISE 1.2.11 (Isometries send lines to lines). Let $x, y \in \mathbb{R}^n$. By the triangle inequality, a point z lies on the line through x and y, and between x and y, if and only if $d(x, z) + d(z, y) = d(x, y)$. Use this to show that if $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is an isometry, then $f(\ell)$ is also a line.

EXERCISE 1.2.12 (Isometries preserve angles). Suppose that $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is an isometry and $x, y, z \in \mathbb{R}^n$. Let θ be the angle from the segment xy to the segment xz, and let ψ be the angle from $f(x)f(y)$ to $f(x)f(z)$.

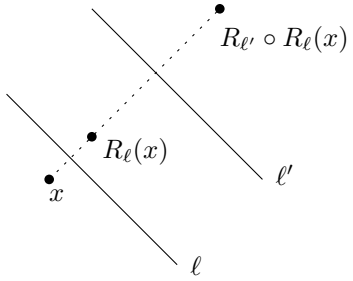
Show that $\theta = \pm\psi$, i.e. the angles have the same magnitude, but one may be counterclockwise while the other is clockwise. *Hint: use the law of cosines.*

1.2.1. Composing isometries of $\mathbb{R}^2$. We know that compositions of isometries are isometries. For instance,

$$T_v \circ T_w = T_{v+w}, \text{ and } O_{p,\theta} \circ O_{p,\psi} = O_{p,\theta+\psi}.$$

What happens if we compose other pairs of isometries?

EXAMPLE 1.2.13 (Composing reflections through parallel lines). Suppose ℓ, ℓ' are parallel lines in $\mathbb{R}^2$. What's $R_{\ell'} \circ R_\ell$? To get some evidence, let's pick some $x \in \mathbb{R}^2$ that lies on the far side of ℓ from ℓ', and very close to ℓ, and then compute $R_{\ell'} \circ R_\ell(x)$.



First, we reflect over ℓ to create $R_\ell(x)$ and then we reflect that over ℓ' to get $R_{\ell'} \circ R_\ell(x)$. So, how does the resulting point compare to x?

We claim that $R_{\ell'} \circ R_\ell(x)$ is obtained from x by translating x in the direction perpendicular to ℓ and ℓ', and 'from' ℓ to ℓ', by a distance that's twice the distance from ℓ to ell'. To see this, note that the line segments from x to $R_{\ell'}(x)$ and from $R_{\ell'}(x)$ to $R_\ell \circ R_{\ell'}(x)$ are perpendicular to ℓ' and ℓ, respectively. Since ℓ' and ℓ are parallel, this means that these segments union to the segment from x to $R_{\ell'} \circ R_\ell(x)$, which is therefore perpendicular to ℓ and ℓ'. The distance from x to $R_{\ell'} \circ R_\ell(x)$ is twice that from ℓ' to ℓ, since ℓ and ℓ' bisect the segments from x to $R_\ell(x)$ and from $R_\ell(x)$ to $R_{\ell'} \circ R_\ell(x)$, respectively. We then might expect that in general:

CLAIM 1.2.14. *If ℓ, ℓ' are parallel, we have $R_{\ell'} \circ R_\ell = T_{2v}$, where v is a vector with its tail on ℓ and its head on ℓ' that is perpendicular to both lines.*

To prove this, however, we have to show that $R_{\ell'} \circ R_\ell(x) = T_{2v}(x)$ for *any* $x \in \mathbb{R}^2$, not just the x in the picture above. For instance, what happens if x is on ℓ or ℓ', or on the far side of ℓ'? Of course, one way to get around this would be to just use coordinate descriptions of the reflections, and show computationally that when they are composed, you get a translation. However, instead of doing this, we will describe now a trick that allows us to geometrically prove that $R_{\ell'} \circ R_\ell = T_{2v}$, but without doing any additional cases.

The key is the following lemma and its corollary.

LEMMA 1.2.15. *If $p, q, x \in \mathbb{R}^2$ and $d(x,p) = d(x,q)$, then x lies on the line that perpendicularly bisects the segment pq. In particular, given p, q all such x are colinear.*

PROOF. Let m be the midpoint of the segment pq. Then the triangle with vertices m, p, x has the same side lengths as the triangle with vertices m, p, y. By Exercise 1.1.11 (SSS) the angles of these triangles at m are the same. Since the angles sum to π, both are $\pi/2$. Hence x lies on the perpendicular bisector as promised. □

COROLLARY 1.2.16. *Suppose that $f, g : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ are isometries and that $x_1, x_2, x_3 \in \mathbb{R}^2$ are non-collinear. If $f(x_i) = g(x_i)$ for $i = 1, 2, 3$, then $f(x) = g(x)$ for all $x \in \mathbb{R}^2$.*

PROOF. Suppose that $f, g : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ are isometries, that $x_1, x_2, x_3 \in \mathbb{R}^2$ are non-collinear and $y_i = f(x_i) = g(x_i)$ for $i = 1, 2, 3$. As both f, g are isometries,

$$d(f(x), y_i) = d(x, x_i) = d(g(x), y_i), \quad \text{for } i = 1, 2, 3.$$

Since x_1, x_2, x_3 are non-collinear, so are y_1, y_2, y_3, by Exercise 1.2.11. It follows from Lemma 3.2.3 that $f(x) = g(x)$. □

Corollary 1.2.16 says that if you want to know whether two isometries are the same, it suffices to check equality only on three points. For example, we can now finish the example above where we compose reflections through parallel lines.

PROOF OF CLAIM 1.2.14. We want to show that if ℓ, ℓ' are parallel, we have $R_{\ell'} \circ R_\ell = T_{2v}$, where v is a vector with its tail on ℓ and its head on ℓ' that is perpendicular to both lines. Above, we showed that $R_{\ell'} \circ R_\ell(x) = T_{2v}(x)$ whenever x is close to ℓ, and on the opposite side of ℓ from ℓ'. But we can certainly find three such x that are non-colinear, so $R_{\ell'} \circ R_\ell = T_{2v}$ on three non-colinear points. Since both sides are isometries, they are equal by Corollary 1.2.16. □

Here's an exercise that has a similar solution.

EXERCISE 1.2.17. Show that if ℓ' and ℓ intersect at a point p, the composition $R_{\ell'} \circ R_\ell$ is the rotation $O_{p,\theta}$, where θ is twice the angle from ℓ to ℓ'. *Hint: use Corollary 1.2.16 to reduce the proof to a single case, as in Claim 1.2.14.*

So far, we have only composed isometries of the same type. What happens if we compose a translation and a reflection?

DEFINITION 1.2.18 (Glide reflection). Suppose $0 \neq v \in \mathbb{R}^2$ and ℓ is a line parallel to v. The composition $T_v \circ R_\ell$ is called the *glide reflection* along ℓ by v.

Note that actually $T_v \circ R_\ell = R_\ell \circ T_v$, so it doesn't matter in which order we write the composition. Also, it is worth mentioning that glide reflections are not translations, rotations or reflections: not all points are translated by the same vector, and glide reflections do not have fixed points as do rotations or reflections.

What happens if ℓ is not parallel to v?

EXERCISE 1.2.19. Suppose that $v \in \mathbb{R}^2$ and ℓ is a line in $\mathbb{R}^2$.

- If v is perpendicular to ℓ, show that the composition $T_v \circ R_\ell$ is a reflection through some line parallel to ℓ.
- If v is not perpendicular to ℓ, show that $T_v \circ R_\ell$ is a glide reflection. *Note: to do this, you must show that $T_v \circ R_\ell = T_{v'} \circ R_{\ell'}$, where v' and ℓ' are parallel. As a hint, write $v = u + w$ where u is parallel to ℓ and w is perpendicular to ℓ, and note that $T_v = T_u \circ T_w$.*

Here's a complete table listing all compositions of translations, rotations, reflections and glide reflections. After doing Exercise 1.2.19, try to prove that some of the other assertions made in the table are correct! In the table, the angles θ and ψ are assumed to be in the interval $(0, 2\pi)$. And the order of multiplication doesn't matter: for example, the entry in the first column, third row describes both $T_v \circ R_{\ell'}$ and $R_{\ell'} \circ T_v$.

	T_v	$O_{p,\theta}$	R_ℓ	$T_v \circ R_\ell$ $v \| \ell$
T_w	id if $v = -w$ otherwise translation	rotation	reflection if $\ell \perp w$ otherwise glide	reflection if $v + w \perp \ell$ otherwise glide
$O_{q,\psi}$	rotation	id if $p = q$ and $\theta = 2\pi - \psi$ translation if $p \neq q$ and $\theta = 2\pi - \psi$, o.w. rotation	reflection if $q \in \ell$ otherwise glide	reflection or glide
$R_{\ell'}$	reflection if $v \perp \ell'$ otherwise glide	reflection if $p \in \ell$ otherwise glide	id if $\ell = \ell'$ translation if $\ell \| \ell'$, $\ell' \neq \ell$ otherwise rotation	translation if $\ell' \| \ell$ otherwise rotation
$T_w \circ R_{\ell'}$ $w \| \ell'$	reflection if $v + w \perp \ell'$ otherwise glide	reflection or glide	translation if $\ell' \| \ell$ otherwise rotation	id if $v = -w$ o.w. translation if $\ell \| \ell'$ o.w. rotation

In particular, any composition of reflections, rotations, translation or glide reflections is again an isometry of one of these types.

So, are there other isometries of $\mathbb{R}^2$ that we haven't discovered yet?

THEOREM 1.2.20 (Classification of Euclidean isometries). *Every isometry of $\mathbb{R}^2$ is either the identity, a translation, a rotation, a reflection or a glide reflection.*

We will work towards a proof of this theorem in steps. Here is step one.

CLAIM 1.2.21. *Suppose $f : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is an isometry and there are points $x \neq y \in \mathbb{R}^2$ such that $f(x) = x$ and $f(y) = y$. Then f is either the identity or a reflection.*

PROOF. Pick a point z that is not on the line ℓ through x, y. Then

$$d(f(z), x) = d(z, x) \text{ and } d(f(z), y) = d(z, y).$$

Therefore, either $f(z) = z$ or $f(z)$ is the other point of intersection of the circle around x with radius $d(z, x)$ and the circle around y with radius $d(z, y)$, which is $R_\ell(z)$. Corollary 1.2.16 then shows that either $f = id$ or $f = R_\ell$. □

Next, let's assume that f fixes only a single point.

CLAIM 1.2.22. *Suppose $f : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is an isometry and $f(x) = x$ for some $x \in \mathbb{R}^2$. Then f is either the identity, a rotation or a reflection.*

PROOF. Pick some $y \neq x$. Then $d(f(y), x) = d(y, x)$, so there is a rotation $O_{x,\theta}$ with $O_{x,\theta}(y) = f(y)$. Consequently, $O_{x,\theta}^{-1} \circ f$ fixes both x and y, so must be either the identity or a reflection (in a line through x) by the previous claim. So,

$$f = O_{x,\theta} \circ \left(O_{x,\theta}^{-1} \circ f\right)$$

is a composition of a rotation and either the identity or a reflection, and hence is either the identity, a reflection or rotation. □

Finally, we can prove the full theorem.

PROOF OF THEOREM 1.2.20. Let $f : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ be an isometry and pick some $x \in \mathbb{R}^2$. Then $T_{x-f(x)} \circ f(x) = x$, so by the previous claim $T_{x-f(x)} \circ f$ is either the identity, a rotation, or a reflection. Since we have

$$f = T_{f(x)-x} \circ (T_{x-f(x)} \circ f),$$

f is a composition of isometries of the given types, so also is one of the given types of isometries, from our work above. □

1.2.2. Exercises.

EXERCISE 1.2.23. Prove that every isometry of $\mathbb{R}^2$ is the composition of at most three reflections. Is three necessary, or would only two reflections suffice?

EXERCISE 1.2.24. Show that every isometry of $\mathbb{R}$, where $d(x, y) = |x - y|$, is either the identity, a translation, or a reflection through a point. Here, the *reflection through* $a \in \mathbb{R}$ is the map $R_a : \mathbb{R} \longrightarrow \mathbb{R}$ defined by $R_a(x) = 2a - x$.

EXERCISE 1.2.25. Describe at least 5 qualitatively different types of isometries of $\mathbb{R}^3$. *You can try to write them out in coordinates for a challenge, but it will be sufficient to just give a geometric description. You don't have to prove they are isometries.*

EXERCISE 1.2.26. Consider $\mathbb{R}^2$ with the metric $d_{max}(x, y) = \max_i\{|x_i - y_i|\}$. Show that all translations are d_{max}-isometries, and that a rotation $O_{p,\theta}$ is a d_{max}-isometry if and only if θ is a multiple of $\pi/2$.

EXERCISE 1.2.27. Here's a proof that the shortest path between two points is a line segment using the integral formula for path length, rather than Definition 1.5.1.

(a) Using the integral formula for path length, show that any path joining the points $(x, 0), (y, 0) \in \mathbb{R}^2$ has length at least $|y - x|$, with equality only if it stays on the line segment between them. *Hint: if $\gamma(t) = (\gamma_1(t), \gamma_2(t))$ is such a path, compare its length to that of the path $\alpha(t) = (\gamma_1(t), 0)$.*

(b) Using isometries and part (a), prove that a path joining two arbitrary points $p, q \in \mathbb{R}^2$ has length at least $d(p, q)$, with equality only if it stays on the line segment between them.

We say that a bijection $f : X \longrightarrow X$ of a metric space is a *similarity* if there is some constant $\lambda > 0$ such that $d(f(x), f(y)) = \lambda d(x, y),\ \forall x, y \in X$. Any isometry is a similarity, where $\lambda = 1$. Another example is the *dilation* around a point $p \in \mathbb{R}^n$:

$$D_{p,\lambda} : \mathbb{R}^n \longrightarrow \mathbb{R}^n, \quad D_{p,\lambda}(x) = \lambda(x - p) + p.$$

Geometrically, a dilation fixes p and stretches every vector based at p by the scaling factor λ. Dilations are similarities, since

$$\begin{aligned} d(D_{p,\lambda}(x), D_{p,\lambda}(y)) &= |\lambda(x - p) + p - \lambda(y - p) - p| \\ &= \lambda\, |x - y| \\ &= \lambda\, d(x, y). \end{aligned}$$

Exercise 1.2.28. Show that every similarity of $\mathbb{R}^n$ is a composition of a dilation and a isometry of $\mathbb{R}^n$.

Similarities send lines to lines – the proof is exactly the same as that for isometries, as described in Exercise 1.2.11. Moreover, similarities of $\mathbb{R}^2$ send circles to circles:

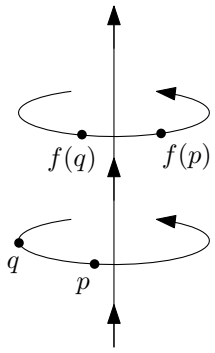
Exercise 1.2.29. Show that if C is a circle in $\mathbb{R}^2$ and $f : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is a similarity, then $f(C)$ is also a circle.

1.3. Isometries of $\mathbb{R}^3$

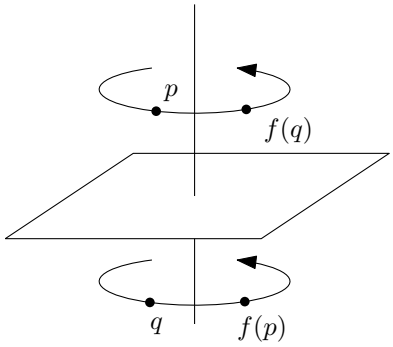
We discussed isometries of $\mathbb{R}^2$ at length in §1.2, and showed that there are only four types: translations, rotations, reflections and glide reflections.

Theorem 1.3.1. *There are seven types of isometries of $\mathbb{R}^3$: the identity, translations, rotations, screw motions, reflections, glide reflections, and twist reflections.*

Here, rotations are around lines and reflections are through planes, a *screw motion* is the composition of a rotation around a line ℓ and a translation parallel to ℓ, a *glide reflection* is the composition of a reflection through a plane P and a translation parallel to P, while a *twist reflection* is the composition of a rotation around a line and a reflection through a perpendicular plane, as pictured below.



screw motion



twist reflection

The proof of Theorem 1.3.1 similar to the classification of isometries of $\mathbb{R}^2$ presented in Theorem 1.2.20, although there are a couple more cases to consider. Here are some exercises that will guide you through the proof.

EXERCISE 1.3.2. Suppose $x_1, \ldots, x_4 \in \mathbb{R}^3$ are not coplanar. If $p, q \in \mathbb{R}^3$, show that

$$d(p, x_i) = d(q, x_i) \ \forall i = 1, \ldots, 4 \implies p = q.$$

Conclude that whenever $f, g : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ are isometries such that $f(x_i) = g(x_i)$ for all i, then $f(p) = g(p)$ for all $p \in \mathbb{R}^3$.

EXERCISE 1.3.3. Suppose P and P' are planes in $\mathbb{R}^3$. Show that the composition of the reflections through P and P' is a rotation around the line $\ell = P \cap P'$ if the planes intersect, and is a translation otherwise.

EXERCISE 1.3.4. Show that a composition of two rotations around lines passing through $p \in \mathbb{R}^3$ is another rotation around a line passing through p.

EXERCISE 1.3.5. Suppose that a line ℓ and a plane P pass through a point $p \in \mathbb{R}^3$. Show that the composition of a rotation around ℓ and a reflection through P is a reflection if $\ell \subset P$ and a twist reflection otherwise.

EXERCISE 1.3.6. Suppose that $f : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ is an isometry.

(a) If $f(p) = p$ for all p in some plane P, show that f is either a reflection through P or is the identity. *Hint: use 1.3.2.*
(b) If $f(p) = p$ for all p in some line ℓ, show that f is either the identity, a reflection through plane containing ℓ, or a rotation around ℓ. *Hint: use (a) and 1.3.3.*
(c) If $f(p) = p$ for some $p \in \mathbb{R}^3$, show that f is either the identity, a reflection, a rotation, or a twist reflection. *Hint: use (b) and 1.3.3.*

EXERCISE 1.3.7. Show that the composition of a translation and rotation is a screw motion unless the direction of translation is perpendicular to the axis of rotation, in which case the composition is a rotation.

EXERCISE 1.3.8. Show that the composition of a translation and a twist reflection is a twist reflection.

EXERCISE 1.3.9. Show that the composition of a translation and a reflection is a reflection if the direction of translation is perpendicular to the plane of reflection, and a glide reflection otherwise.

EXERCISE 1.3.10. Prove Theorem 1.3.1, using the previous 3 exercises and 1.3.6 (c).

1.4. Paths and lines

The shortest path between two points is a line segment.
– Ms. Tuttino

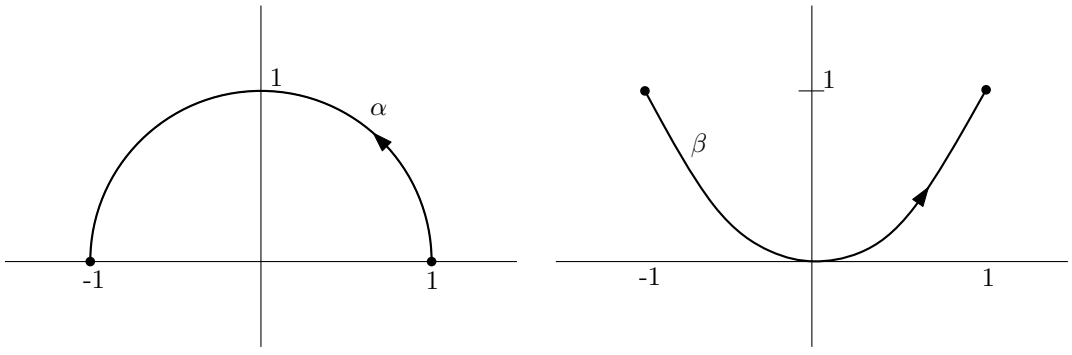
The quote above is repeated in nearly every high school geometry class. In the next two sections, we introduce the definitions needed to make the sentence precise.

DEFINITION 1.4.1. A *path* in $\mathbb{R}^n$ is a continuous map

$$\gamma : [a, b] \longrightarrow \mathbb{R}^n,$$

where $a, b \in \mathbb{R}$. Sometimes, we will let either a or b be infinite, so that the path γ is defined on an interval of the form $[a, \infty)$, $(-\infty, b]$ or $(-\infty, \infty) = \mathbb{R}$.

For example, we have the following $\alpha : [0, \pi] \longrightarrow \mathbb{R}^2$, $\alpha(t) = (\cos(t), \sin(t))$ and $\beta : [-1, 1] \longrightarrow \mathbb{R}^2$, $\beta(t) = (t, t^2)$ are pictured below. We imagine $\gamma(t)$ as the position of a hiker at time t; the arrows then indicate the direction of traversal.



Note that a 'path' is not just some subset of $\mathbb{R}^n$, it is a *function* that gives position at a given time. In other words, the path you traversed last night to the grocery store is not 'two blocks on Main Street and three blocks on Wall Street', it is 'at 8:30 you began walking along Main Street, then at 8:32 you made a right turn onto Wall Street and continued until you arrived at the grocery store at 8:35.'

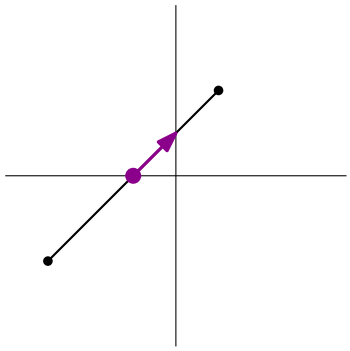
In order to show that shortest paths are line segments, we'll need to understand how to write a line ℓ in coordinates. One way is to pick a point p on ℓ and a vector v that lies along ℓ, in which case we can write

$$\ell = \{p + t \cdot v \mid t \in \mathbb{R}\}.$$

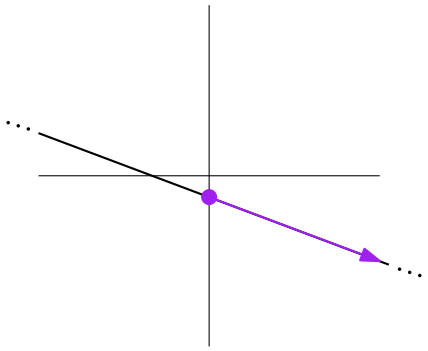
Similarly, a line segment can be described by limiting t to an interval:

$$\{p + t \cdot v \mid t \in [a, b]\}.$$

Here are two examples – the first is a line segment and the latter a full line:



$\{(-1, 0) + t \cdot (1, 1) \mid t \in [-2, 2]\}$



$\{(0, -1/2) + t \cdot (4, -3/2) \mid t \in \mathbb{R}\}$

If $x, y \in \mathbb{R}^n$, the line segment between x and y can be written as

$$xy = \{(1 - t)x + ty \mid t \in [0, 1]\}$$

As $(1 - t)x + ty = x + t(x - y)$, we can describe xy as before by setting $p = x$ and $v = y - x$. However, this new presentation is sometimes preferred because it illustrates that the points of xy are 'weighted averages' of x and y.

Exercise 1.4.2. If $0 \neq v = (v_1, v_2) \in \mathbb{R}^2$, let $v^\perp = (-v_2, v_1)$. Show that the line

$$\{tv \mid t \in \mathbb{R}\} = \{x \in \mathbb{R}^2 \mid x \cdot v^\perp = 0\}.$$

Geometrically, this means that the line through 0 in the direction of v consists of all vectors that are perpendicular to $v^\perp$. Then show that if $p \in \mathbb{R}^2$, we also have

$$\{p + tv \mid t \in \mathbb{R}\} = \{x \in \mathbb{R}^2 \mid x \cdot v^\perp = p \cdot v^\perp\}.$$

Hint: use the first part. As an aside, note that writing $x = (x_1, x_2)$, $v^\perp = (a, b)$ and $p \cdot v^\perp = c$, the equation $x \cdot v^\perp = p \cdot v^\perp$ is really just

$$ax_1 + bx_2 = c,$$

a usual Cartesian equation for a line. So, when you are writing down such a Cartesian equation, remember that the condition you are really imposing is that your point has a prescribed dot product with some vector!

So far, our lines are subsets of $\mathbb{R}^n$ rather than paths. However, a line

$$L = \{p + t \cdot v \mid t \in \mathbb{R}\}$$

is clearly the image of the path

$$\gamma : \mathbb{R} \longrightarrow \mathbb{R}^n, \quad \gamma(t) = p + tv,$$

which we say *parameterizes* L. Restricting the domain to some interval $[a, b]$ gives a path that parameterizes a line segment.

Lines can be described in many different ways. For instance,

$$\{(-4, 1) + t \cdot (8, -3) \mid t \in \mathbb{R}\} = \{(0, -1/2) + t \cdot (4, -3/2) \mid t \in \mathbb{R}\}.$$

To see this, observe that any point $(-4, 1)+t\cdot(8, -3)$ can also be written as $(0, -1/2)+(2t - 1) \cdot (4, -3/2)$, while any point $(0, -1/2) + t \cdot (4, -3/2)$ can also be written as $(-4, 1) + \frac{t+1}{2} \cdot (8, -3)$. So, the two distinct paths

$$\gamma : \mathbb{R} \longrightarrow \mathbb{R}^2, \ \ \gamma(t) = (-4, 1) + t \cdot (8, -3),$$
$$\delta : \mathbb{R} \longrightarrow \mathbb{R}^2, \ \ \delta(t) = (0, -1/2) + t \cdot (4, -3/2)$$

both parametrize the same line! The map $t \longmapsto 2t - 1$ that translates between the t-coordinates in both expressions is called a 'reparameterization'.

Here is a formal definition.

Definition 1.4.3. Suppose that $\gamma : [c, d] \longrightarrow \mathbb{R}^n$ is a path and $f : [a, b] \longrightarrow [c, d]$ is a *homeomorphism*, a continuous bijection with continuous inverse. Then $\gamma \circ f : [a, b] \longrightarrow \mathbb{R}^n$ is a new path that is a *reparameterization* of γ.

Note that $\gamma \circ f$ and γ have the same image in $\mathbb{R}^n$. The only difference is that this image is being traced at different rates, and possibly different times or in different directions, by the two paths. As usual, we also allow the intervals above to be half-infinite or infinite. In the example above with γ, δ above,

$$\gamma(t) = (-4, 1) + t \cdot (8, -3) \quad \text{and} \quad f(t) = \frac{t + 1}{2},$$

and we have

$$\gamma \circ f(t) = (-4, 1) + \frac{t + 1}{2} \cdot (8, -3) = (0, -1/2) + t \cdot (4, -3/2) = \delta(t).$$

For another example, consider the two paths

$$\alpha, \beta : [0, 1] \longrightarrow \mathbb{R}^2, \quad \alpha(t) = (\cos(t), \sin(t)), \quad \beta(t) = (\cos(t^2), \sin(t^2)).$$

Clearly, β is a reparameterization of α by the map $f : [0, 1] \longrightarrow [0, 1]$, $f(t) = t^2$. The path β starts out slower than α, increases its speed until it is going as fast as α at $t = 1/2$, and continues to increase speed until its position catches up with α at $t = 1$.

Exercise 1.4.4. Show that if $f : [c, d] \longrightarrow [a, b]$ is a homeomorphism, then either

(a) f is increasing, $f(c) = a$, $f(d) = b$, or

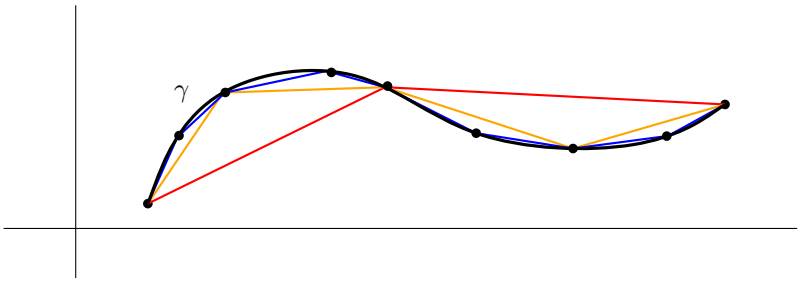
(b) f is decreasing, $f(c) = b$, $f(d) = a$.

Hint: since f is a bijection, either $f(c) < f(d)$ or $f(c) > f(d)$. Let's assume that we are in the first case. Given $c < x < y < d$, first show that $f(x) < f(d)$ using the 'intermediate value theorem', which says that a continuous function on an interval takes on every value between the values of its endpoints. Then use the same argument to show that $f(x) < f(y)$, which means f is increasing.

EXERCISE 1.4.5. Prove the 'Sylvester-Gallai Theorem': for any finite set S of non-collinear points in $\mathbb{R}^2$, there is a line passing through exactly two of the points in S. *Hint: look at pairs (p, ℓ), where $p \in S$ and ℓ passes through at least two points of S, but not through p. Take such a pair where the distance from p to ℓ is minimal, and show that l only passes through two points of S.*

1.5. Path length

How does one define the length of the path? Presumably the length of a line segment should be just the distance between its endpoints. Similarly, the length of a *piecewise linear* path, i.e. a concatenation of line segments, should be the sum of the lengths of the line segments. To measure the length of an arbitrary path, we take a limit of the lengths of piecewise linear approximations.



Above, we have a path $\gamma : [a, b] \longrightarrow \mathbb{R}^n$ drawn in black. The red, orange and blue paths are increasingly better piecewise linear approximations of γ. Each is constructed by choosing a *partition* $\{a = t_0 < \cdots < t_n = b\}$ of $[a, b]$, and then connecting each $\gamma(t_i)$ to $\gamma(t_{i+1})$ with a line segment. The length of such an approximation is then

$$\sum_{i=0}^{n-1} d\left(\gamma(t_i), \gamma(t_{i+1})\right).$$

Consider the effect of inserting a new element t into a partition between t_k and t_{k+1}. By the triangle inequality,

$$d\left(\gamma(t_k), \gamma(t_{k+1})\right) \leqslant d\left(\gamma(t_k), \gamma(t)\right) + d\left(\gamma(t), \gamma(t_{k+1})\right),$$

so the new partition will give a path with length at least that of the old partition. That is, as more points are added to a partition, length increases. We define:

DEFINITION 1.5.1. The *length* of a path $\gamma : [a, b] \longrightarrow \mathbb{R}^n$ is defined as

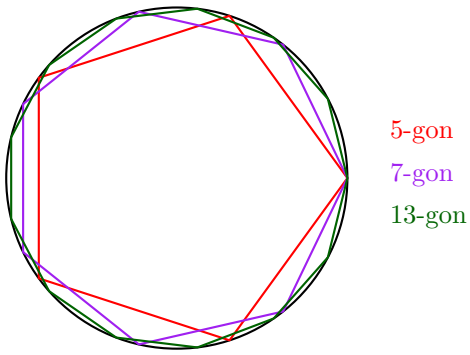
$$\text{length}(\gamma) = \sup_{\{a=t_0,\ldots,t_n=b\}} \sum_{i=0}^{k} d\left(\gamma(t_i), \gamma(t_{i+1})\right),$$

where the supremum is taken over all partitions $\{a = t_0 < \cdots < t_n = b\}$ of $[a, b]$.

REMARK 1.5.2. For the unfamiliar, a *supremum* is a 'least upper bound' of a set. For example, the suprema of the intervals $[0, 1]$ and $(-5, 1)$ are both 1, while the suprema of $\mathbb{R}$ and $[2, \infty)$ are ∞. One might be tempted to say 'maximum' instead of supremum, but unless γ is itself piecewise linear, there will not be any piecewise linear approximation that has exactly the same length as γ. For example, if $\gamma : [0, \pi] \longrightarrow \mathbb{R}^2$, $\gamma(t) = (\cos t, \sin t)$ traces out a semicircle, it turns out that the lengths of γ's piecewise linear approximations fill out exactly the interval $[2, \pi)$. So, the length of γ is π.

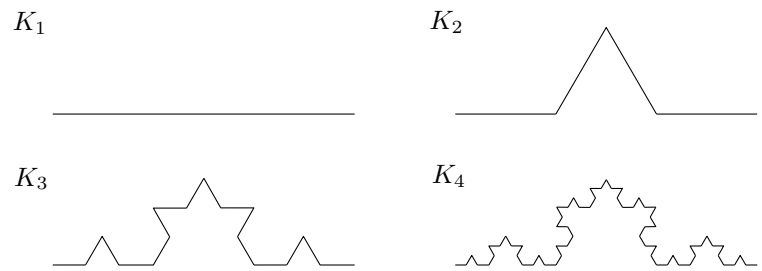
EXERCISE 1.5.3. Verify that the length of a line segment is the distance between its endpoints. *Hint: use the triangle inequality!*

EXERCISE 1.5.4. Regular n-gons can be use as piecewise linear approximations of the unit circle, which we expect has length 2π. Show that the perimeters of regular n-gons whose vertices lie on the unit circle do indeed converge to 2π as $n \to \infty$. *Hint: use the law of cosines. You might also find L'Hopital's rule useful.*



In fact, there are paths $\gamma : [a, b] \longrightarrow \mathbb{R}^2$ that have piecewise linear approximations with arbitrarily large length, so the supremum is ∞! Let's quickly sketch an iterative construction of such a path, the 'Koch snowflake curve'.

The game starts with K_1, which is a line segment of unit length. K_2 is created by replacing the middle third of K_1 by two other sides with which it makes an equilateral triangle, as below. We then continue inductively, replacing the middle third of each line segment in K_n with two line segments as before.



It turns out that as $n \longrightarrow \infty$, the paths K_n converge to a limiting path K, of which the K_n are piecewise linear approximations. This K is the Koch snowflake curve.

EXERCISE 1.5.5. Find a formula for the length of K_n, and show that it goes to infinity as n increases. *It is true that we should really regard the K_n as images of paths $\gamma_n : [0,1] \longrightarrow \mathbb{R}^2$, in which case K is the image of some $\gamma : [0,1] \longrightarrow \mathbb{R}^2$. However, as this is just a sketch of a construction, we won't bother with that. In any case, the particular parameterization shouldn't enter in your computation of the length of K_n.*

So, as there are piecewise linear approximations of K with arbitrarily large length, the length of K is infinite. Such curves are often called *non-rectifiable*. The Koch curve is also an example of a 'fractal'. One of the features of fractality is self similarity, which is evident in the the fact that each K_n subdivides into three pieces, each of which is a scaled copy of the entire K_n. This self similarity deepens in the limit, so that smaller copies of the Koch curve can be seen at many scales.

Of course, the length of a path should not depend on the speed at which it was traversed. In other words, reparametrization does not change path length.

PROPOSITION 1.5.6. *Suppose that $\gamma : [a,b] \longrightarrow \mathbb{R}^n$ is a path and $f : [c,d] \longrightarrow [a,b]$ is a homeomorphism. Then* $\text{length}(\gamma \circ f) = \text{length}(\gamma)$.

PROOF. Let's suppose that $f(c) = a$ and $f(d) = b$, rather than the other way around. If $\{a = t_0 < \cdots < t_n = b\}$ is a partition of $[a,b]$, then $\{c = f^{-1}(t_0) < \cdots < f^{-1}(t_n) = b\}$ is a partition of $[c,d]$, and

$$\sum_{i=0}^{k} d\left(\gamma(t_i), \gamma(t_{i+1})\right) = \sum_{i=0}^{k} d\left(\gamma \circ f(f^{-1}(t_i)), \gamma \circ f(f^{-1}(t_{i+1}))\right).$$

Similarly, to every partition of $[c,d]$ there is a corresponding partition of $[a,b]$ that gives the same length sum. In other words, for every piecewise linear approximation of γ, there is a piecewise linear approximation of $\gamma \circ f$ with the same length, and vice versa. So, the suprema of these lengths are the same. □

We can now prove that a shortest path between two points is a line segment.

THEOREM 1.5.7. *If $\gamma : [a,b] \longrightarrow \mathbb{R}^n$ is a path, then* $\text{length}(\gamma) \geqslant d(\gamma(a), \gamma(b))$. *If this is an equality, then γ lies on the line segment between $\gamma(a)$ and $\gamma(b)$.*

PROOF. Taking the partition $\{a, b\}$ gives a length estimate of $d(\gamma(a), \gamma(b))$, so the supremum must be at least that large. If there is some $\gamma(t)$ that doesn't lie on the line segment between $\gamma(a)$ and $\gamma(b)$, then the partition $\{a, t, b\}$ gives a length estimate that is strictly bigger than $d(\gamma(a), \gamma(b))$, by the triangle inequality (Theorem 1.1.6). □

The major drawback of our definition of length is that it is pretty useless for hand calculations. We now describe a different approach using calculus.

Suppose that $\gamma : [a, b] \longrightarrow \mathbb{R}^n$ is a differentiable path, i.e. we can write $\gamma(t) = (\gamma_1(t), \ldots, \gamma_n(t))$ where each $\gamma_1, \ldots, \gamma_n$ is a differentiable real valued function. The *velocity vector* of γ at time t is the vector

$$\gamma'(t) = (\gamma_1'(t), \ldots, \gamma_n'(t)).$$

We often draw $\gamma'(t)$ as a vector based at the point $\gamma(t)$ in $\mathbb{R}^2$. Note that

$$\lim_{h \to 0} \frac{\gamma(t+h) - \gamma(t)}{h} = \gamma'(t), \tag{3}$$

since one can distribute the limit inside each coordinate. So, the velocity vector indicates the infinitesimal rate of change of γ through $\mathbb{R}^n$.

The *speed* of γ at time t is the length $|\gamma'(t)|$ of its velocity vector. Here, the *length* of a vector $(a_1, \ldots, a_n) \in \mathbb{R}^n$ is defined by

$$|(a_1, \ldots, a_n)| = \sqrt{a_1^2 + \ldots + a_n^2}.$$

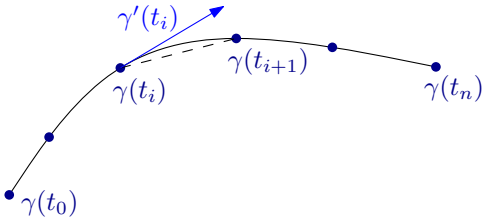
In other words, it is the distance from the tail of the vector to its head. The central theorem of this section is then the following:

THEOREM 1.5.8. *If $\gamma : [a, b] \longrightarrow \mathbb{R}^n$ is a differentiable path, we have*

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

A true proof of the theorem is too technical for us, but we can give the intuition.

PROOF IDEA. Suppose we have a path $\gamma : [a, b] \longrightarrow \mathbb{R}^2$ and partition $[a, b]$ into subintervals using the points $a = t_0 < t_1 < \ldots < t_n = b$. Connecting the points $\gamma(t_0), \ldots, \gamma(t_n)$ gives an approximation of γ.



When the partition is very fine, Equation (3) implies that

$$\frac{\gamma(t_{i+1}) - \gamma(t_i)}{t_{i+1} - t_i} \approx \gamma'(t_i).$$

Therefore, the length of the piecewise linear approximation is

$$\begin{aligned}\sum_{i=0}^{n-1} d(\gamma(t_{i+1}),\gamma(t_i)) &= \sum_{i=0}^{n-1} |\gamma(t_{i+1})-\gamma(t_i)| \\ &\approx \sum_{i=0}^{n-1} |\gamma'(t_i)|\cdot(t_{i+1}-t_i),\end{aligned}$$

which is exactly a Riemann sum approximating the integral of $|\gamma'(t)|$. □

EXAMPLE 1.5.9. Let $\gamma : [0, 4\pi] \longrightarrow \mathbb{R}^2$, $\gamma(t) = (\cos t, \sin t)$ be a path that winds twice around the unit circle in $\mathbb{R}^2$. Then $\gamma'(t) = (-\sin t, \cos t)$, so

$$|\gamma'(t)| = \sqrt{\sin^2 t + \cos^2 t} = 1$$

for all $t \in [0, 4\pi]$. Thus,

$$\text{length}(\gamma) = \int_0^{4\pi} 1\, dt = 4\pi.$$

EXAMPLE 1.5.10. Let $\alpha : [0, 1] \longrightarrow \mathbb{R}^3,\ \alpha(t) = (t, 0, t^2)$ be a path in $\mathbb{R}^3$ that traces out a portion of a parabola in the xz-plane. Then $\alpha'(t) = (1, 0, 2t)$, so

$$|\alpha'(t)| = \sqrt{1 + 4t^2}$$

for all $t \in [0, 1]$. Thus,

$$\text{length}(\alpha) = \int_0^1 \sqrt{1 + 4t^2}\, dt \approx 1.47894.$$

EXERCISE 1.5.11. Verify that the length of a line segment is the distance between its endpoints using the integral formula, rather than the definition of path length.

EXERCISE 1.5.12. Let $S = \{v \in \mathbb{R}^3 \mid |v| = 1\}$ be the unit sphere in $\mathbb{R}^3$, let $v, w \in S$ be perpendicular vectors and let P be the plane that spanned by v, w. The path

$$\gamma : [0, 2\pi] \longrightarrow \mathbb{R}^3,\ \gamma(t) = \cos(t)v + \sin(t)w$$

parameterizes the intersection $P \cap S$, which is called a *great circle* on S. Show that $|\gamma'(t)| = 1$ for all t and verify that $\text{length}(\gamma) = 2\pi$.

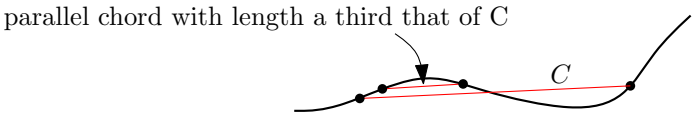
EXERCISE 1.5.13. Suppose that $\gamma : [a, b] \longrightarrow \mathbb{R}^n$ is a differentiable path and $f : [c, d] \longrightarrow [a, b]$ is a *diffeomorphism*, i.e. a differentiable bijection with differentiable inverse.

Prove that $\text{length}(\gamma \circ f) = \text{length}(\gamma)$ using just the integral formula for length, rather than the definition. *Hint: show that either $f'(t) > 0$ for all t or $f'(t) < 0$ for all t, depending on whether f is increasing or decreasing, see Exercise 1.4.4. Split into two cases and use the chain rule, which says that $(\gamma \circ f)'(t) = \gamma'(f(t))f'(t)$.*

1.6. The Chord Theorem

Suppose that $\gamma : [a, b] \longrightarrow \mathbb{R}^2$ is a path. A *chord* for γ is a line segment both of whose endpoints lie on γ.

THEOREM 1.6.1 (The Chord Theorem). *If C is a chord for γ with length a, then for every $n = 1, 2, \ldots$, there is another chord for γ with length a/n that is parallel to C.*



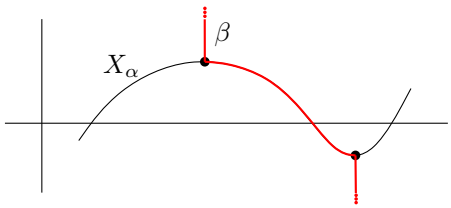
The proof will use the following lemma.

LEMMA 1.6.2. *Suppose C is a chord for γ with length c. Then for every $\alpha \in (0, 1)$, there is a parallel chord either with length αc or length $(1 - \alpha)c$.*

PROOF. After rotating, scaling and translating the picture, let's assume for simplicity that C is the line segment joining the origin to $(1, 0)$. Let

$$X_s = \{\gamma(t) + (s, 0) \mid t \in [a, b]\}.$$

We'll be particularly interested in X_0, which is just the image of γ, and X_α, X_1, which are obtained by shifting the image of γ to the right by α and 1, respectively. It suffices to show that either $X \cap X_\alpha \neq \varnothing$ or $X_\alpha \cap X_1 \neq \varnothing$, for γ has a horizontal chord of length α exactly when X_0 intersects its translate X_α, while a length $1 - \alpha$ chord amounts to the second intersection being nonempty. So, hoping for a contradiction, assume that $X \cap X_\alpha = \varnothing = X_\alpha \cap X_1$.



Construct a bi-infinite path β by taking the part of X_α between its highest point and its lowest point, and concatenating with vertical rays emanating up from the highest point and down from the lowest point, as in the picture above. We claim that β is disjoint from X_0. First, X_0 cannot intersect the part of β that lies along X_α, since we assumed that $X_0 \cap X_\alpha = \varnothing$. But X_α is a horizontal shift of X_0, so no point of X_0 can be higher than the highest point of X_α, or lower than the lowest point of X_α, so X_0 cannot intersect the other two parts of β. Similarly, X_1 is disjoint from β.

The path β splits the plane into two pieces. We saw above that β is disjoint from both X_0 and X_1. In fact, X_0, X_1 lie on different sides of β, where X_0 is on the 'left' and X_1 is on the 'right'. So, X_0 and X_1 are disjoint.

Now, we started out by assuming that the line segment joining the origin to $(1,0)$ was a chord, so both the origin and $(1,0)$ lie in X_0. But if the origin lies in X, then $(1,0)$ lies in X_1 as well! So, X_0, X_1 intersect. This is a contradiction. □

EXERCISE 1.6.3. Prove the chord theorem, using the lemma. *Hint: try to prove the following statement using induction on n: for every n, whenever C is a chord for γ of length a, there is a parallel chord for γ with length a/n.*

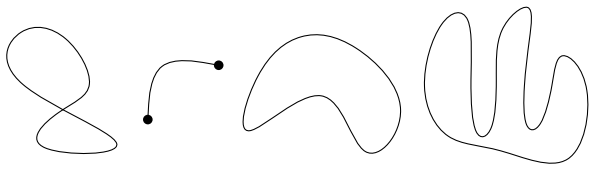
In fact, for every $\alpha \in (0,1)$ that is *not* of the form $\frac{1}{n}$, for some natural number n, the conclusion of the chord theorem fails! That is, for each such α, there is a path γ with a chord of length a that has no parallel chord with length αa! Here is an explicit example.

EXERCISE 1.6.4. Suppose that $\alpha \in (0,1)$ and that $\alpha \neq 1/n$ for any $n \in \mathbb{N}$. Show that the graph of the function $f(x) = \sin^2(\pi x/\alpha) - x\sin^2(\pi/\alpha)$ has a horizontal chord of length 1, but no horizontal chord of length α.

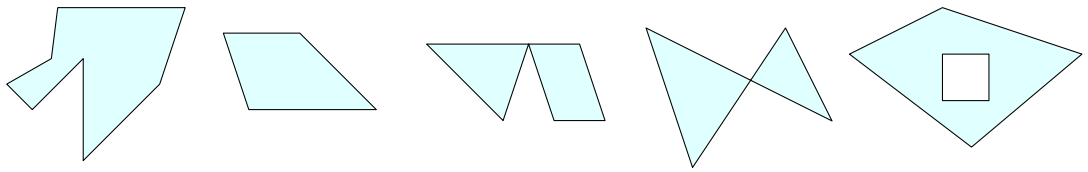
EXERCISE 1.6.5. Suppose that $f : \mathbb{R} \longrightarrow \mathbb{R}$ is continuous and *periodic*, meaning that for some $a \in \mathbb{R}$, we have $f(x+a) = f(x)$ for all x. Show that the graph of f has horizontal chords of every length.

1.7. Polygons and Triangulations

A *path* in $\mathbb{R}^2$ is a continuous map $\gamma : [a,b] \longrightarrow \mathbb{R}^2$. A *loop* is a path that doesn't intersect itself and comes back to where it started, i.e. a path γ such that if $\gamma(a) = \gamma(b)$, and where if $x \neq y \in [a,b]$ then $\gamma(x) = \gamma(y)$ only when one of x, y is a and the other is b. The first two paths below are not loops, while the latter two are loops.



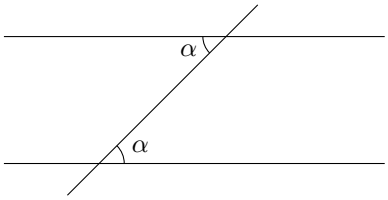
A *polygon* is a region of the plane bounded by a finite number of line segments that form a loop. Polygons with n sides are also called n-gons, and for small n we also use the conventional terms triangle, quadrilateral, pentagon, hexagon, etc.



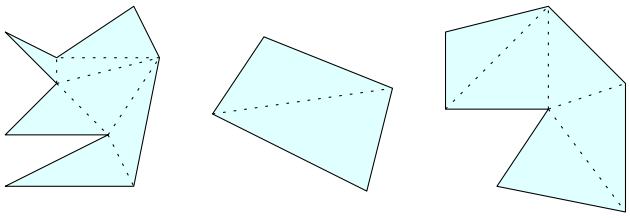
EXERCISE 1.7.1. Show that the following are equivalent, for a quadrilateral Q. We call a quadrilateral satisfying any/all of these four conditions a *parallelogram*.

(a) opposite sides of Q have the same length,
(b) angles at opposite vertices of Q are equal,
(c) opposite sides of Q are parallel,
(d) the diagonals of Q bisect each other.

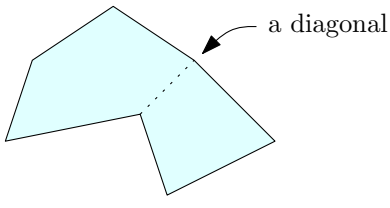
Hint: prove $(a) \implies (b) \implies (c) \implies (d) \implies (a)$. *You might find the SSS, SAA and SSA conditions for congruence of triangles useful, as described in Exercises 1.1.11 and 1.1.12, the fact that the angle sum of a quadrilateral is* 2π*, and the fact that two lines are parallel if and only if whenever another line intersects them both, the 'alternate interior angles' are equal, as pictured below.*



A *triangulation* of a polygon P is a collection of triangles that union to P, whose vertices are all vertices of P, and where any two of the triangles intersect exactly in an entire edge of each, or in a vertex of each.



So, can every polygon be triangulated? You can probably guess that the answer is yes, but how do you prove it in general? Looking at the examples above, the first step in triangulating a polygon is to show it has a *diagonal*, a line segment connecting two nonadjacent vertices of P that is entirely contained in P.



LEMMA 1.7.2. *If* $n \geqslant 4$*, any* n*-gon has a diagonal.*

PROOF. Pick a vertex p of P such that the interior angle at P is less than π, and let q, r be the adjacent vertices. (For instance, you can take p to be a 'leftmost' vertex of P, i.e. a vertex where the first coordinate is as small as possible. Then q and r both lie to the right of p, so the interior angle is less than π.) The point of requiring that the interior angle is less than π is that then, if you start moving into the triangle pqr from p, you move *into* P rather than out of it.

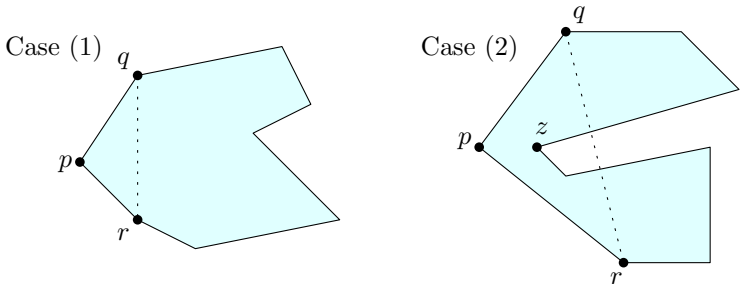


FIGURE 2. The two cases in the inductive step.

If the segment qr is a diagonal, we are done. So, assume qr is not contained in P. In this case, there must be vertices of P inside the interior of the triangle pqr. Let z be the vertex in the interior of pqr that lies farthest from the segment qr.

We claim that pz is a diagonal. Since the segment pz starts out at p by going into P, the only way it can fail to be a diagonal is if it hits the boundary of P before it hits z. It cannot hit a vertex of P before z, since that vertex would be farther from qr than z. And it cannot hit an edge of P before z, since if it did, one of the two vertices of that edge has to lie in pqr and be farther from qr than z, contrary to assumption. So, pz is a diagonal. □

THEOREM 1.7.3. *Every n-gon in $\mathbb{R}^2$ admits a triangulation with $n-2$ triangles.*

PROOF. The proof is by induction, and the base case $n = 3$ is obvious. So, suppose that the theorem is true for $(n-1)$-gons, and let P be an n-gon.

By the lemma, P has a diagonal, which cuts P into two polygons with fewer vertices, say an i-gon and a j-gon. Note that $i + j = n + 2$, since the vertices of the diagonal appear in both polygons. By induction, the i-gon and j-gon admit triangulations with $i-2$ vertices and $j-2$ triangles, respectively. The union of the two is a triangulation of P with $i - 2 + j - 2 = i + j - 4 = n - 2$ triangles, as desired. □

If $f, g : \mathbb{N} \longrightarrow \mathbb{N}$ are functions, we say that $f = O(g)$ if there is some real number C such that $f(n) \leqslant C \cdot g(n)$ for all n. This is called *big O* notation, and is especially common in computer science. For example, you can check that

$$(1 + n^5)(\sin(n) + n^2) = O(n^7).$$

EXERCISE 1.7.4. Analyzing the proof above, explain why the number of steps required to triangulate an n-gon is $O(n^4)$. *Here, we're not being so careful with*

defining 'step'. Like, if you're doing an arithmetic computation, is computing $13 + 9$ *a single step, or do you have multiple steps corresponding to how you'd write out the computation on paper? The advantage of the big O notation is that you don't have to sweat these kind of details, since if the total number of steps is* $100n$ *vs* $2n$*, it's still* $O(n)$*.*

There are faster ways to triangulate polygons, though. Here's one approach.

DEFINITION 1.7.5. An *ear* of a polygon P is a triangle consisting of three consecutive vertices r, p, q on the boundary of P such that p, q is a diagonal. Two ears of P *overlap* if their interiors intersect, which happens when they're the same ear, or when we have 4 consecutive vertices r, p, q, z on the boundary of P, and the two ears we're considering are the triangles rpq and pqz.

THEOREM 1.7.6. *If* $n \geqslant 4$*, every* n*-gon* P *has two non-overlapping ears.*

PROOF. We proceed by (strong) induction. If $n = 4$, we're done, since either of the two possible diagonals splits P into two non-overlapping ears. For the inductive case, suppose we have n-gon P, and that the theorem holds for polygons with fewer sides. Pick a diagonal xy, and split P along it into two polygons with fewer vertices. By induction, each of these has two nonoverlapping ears, so there's one ear of each that doesn't use the edge xy. These are two nonoverlapping ears of P. □

EXERCISE 1.7.7. By looking for ears instead of arbitrary diagonals, give an algorithm to triangulate an n-gon in $O(n^3)$ steps.

This strategy for triangulating a polygon is called *ear clipping*. It turns out that if done extra intelligently, see this article, ear clipping actually gives an $O(n^2)$-time algorithm for triangulating an n-gon. There are also good $O(n \log n)$ algorithms, which are what are used in practice. Chazelle (1991) even gave an $O(n)$ algorithm, but it's so ridiculously complicated that noone uses it in practice.

EXERCISE 1.7.8. Show that for each n, there is an n-gon with a *unique* triangulation!

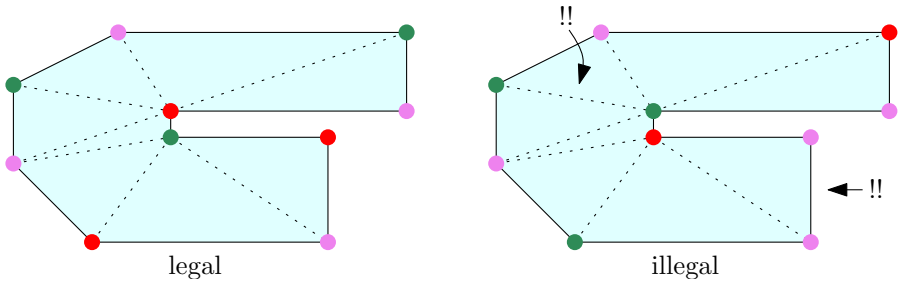
Here's a nice application of ears and triangulations. Suppose that we have a polygon that represents an art gallery. The *art gallery problem* asks 'if the polygon has n sides, how many cameras do we need to install in the gallery so that every point is always on camera'? Assume that the cameras have a full 360° field of vision.

THEOREM 1.7.9. $\lfloor n/3 \rfloor$ *cameras always suffice, and for each* n*, there is an example in which* $\lfloor n/3 \rfloor$ *cameras are necessary.*

Here, $\lfloor x \rfloor$ is the greatest integer less than or equal to x, so $\lfloor 2.32 \rfloor = 2$. To prove the theorem, we'll need the following lemma, which we prove via ear clipping.

LEMMA 1.7.10. *Suppose that* P *is a triangulated polygon. Show that the vertices of* P *can be colored with three colors so that vertices that share an edge of the triangulation have different colors.*

We will call such a coloring of the vertices a 3-*coloring* of P.

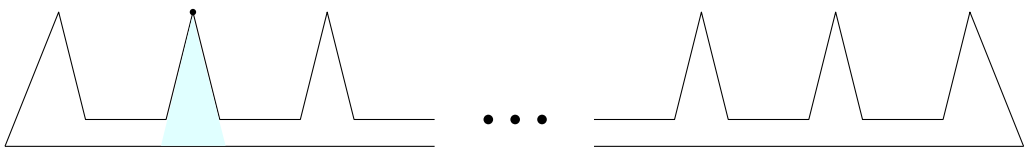


PROOF. We use strong induction on the number of sides of the polygon P. Certainly, the vertices of a triangle can be thus colored, just by using different colors for the three vertices. So, assume that the vertices of any triangulated polygon with less than n vertices can be three colored, and let P be a triangulated n-gon.

Let xy be a diagonal of P. Then xy splits P into two triangulated polygons Q, R with less than n vertices, which can both be 3-colored. The vertices x, y must have different colors in both Q and R, so by permuting the colors in R, we can assume that x is the same color in Q, R and y is the same color in Q, R. The two colorings then combine to give a coloring of the vertices of P as desired. Since every diagonal of P is a diagonal of either Q or R, this coloring is a 3-coloring as desired. □

PROOF OF THEOREM 1.7.9. Triangulate our n-gon P, and color the vertices of P so that vertices was that share an edge of the triangulation have different colors. One of the colors, say 'blue', appears at most $\lfloor n/3 \rfloor$ times, and we station our cameras at the blue vertices. Every triangle in the triangulation must have a blue vertex, since otherwise one of its edges would connect two vertices of the same color. As a camera stationed at such a blue vertex can see the entire adjacent triangle, our blue-positioned cameras can see the entire art gallery.

Here is an example indicating that $\lfloor n/3 \rfloor$ cameras are necessary.

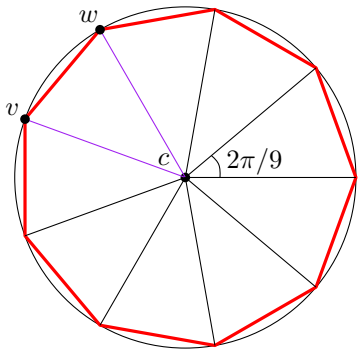


If there are k peaks in the triangle above, there are $n = 3k$ total vertices. Each peak vertex casts a blue shadow in the polygon, and a camera that sees the peak must be positioned somewhere in this shadow. As all the shadows are disjoint, we need at least $k = n/3$ cameras. This gives an example as long as n is a multiple of three. To make examples with $n = 3k + 1$ or $n = 3k + 2$, just insert either one or two additional vertices into the bottom edge of the polygon drawn above. □

1.7.1. Exercises.

EXERCISE 1.7.11. As an extension of the art gallery problem, construct a polygon P and a placement of cameras in P such that every point of the loop bounding P is on camera, but some point of the interior of P is not.

A polygon is *regular* if all its side lengths are the same and all its angles are the same. One can construct a regular n-gon by choosing a center c, then laying the vertices of the polygon at angle increments of $2\pi/n$ along a circle centered at c.



a regular 9-gon

EXERCISE 1.7.12. Show that all regular n-gons are of this form. *Hint: to prove this, you must take a polygon P all of whose side lengths and angles are equal, construct the center c and show that the vertices lie as described along a circle centered at c. If v, w are adjacent vertices of P, construct a triangle using the line segment vw and segments of the interior angle bisectors at the vertices v, w, pictured above in purple. Explain why the third vertex of this triangle is always the same, no matter which adjacent pair v, w is chosen. Then use this as your c.*

EXERCISE 1.7.13. Draw an example of a pentagon with all the same side lengths, but not all the same angles. Is it possible to do this with a quadrilateral?

EXERCISE 1.7.14. The interior angles of a triangle sum to π. Show that the interior angles of an n-gon sum to $\pi(n-2)$, and then use this to find a formula for the individual interior angles of a regular n-gon.

EXERCISE 1.7.15. Show that any triangulation of an n-gon has exactly $n-2$ triangles. *Hint: you can do this with an induction proof.*

A polygon P is *convex* if whenever $x, y \in P$, the line segment $xy \subset P$. Try to draw some examples of convex, and non-convex polygons.

EXERCISE 1.7.16. Let's say a vertex of a polygon P is *convex* if its interior angle is at most π, and concave otherwise. Show that P is convex if and only if all its vertices are convex.

EXERCISE 1.7.17. How many triangulations can a polygon have? We saw in Exercise 1.7.8 that the answer may be 1. On the flip side, any *convex* polygon has many triangulations, since the line segment connecting any two vertices is a diagonal. Indeed, it turns out that convex n-gons have the most triangulations out of all n-gons, and that all convex n-gons have the same number of triangulations. (Try to convince yourself of this if you like, but you don't have to write a proof.) In this problem, we'll try to calculate this number.

Let t_n be the number of triangulations of (any) convex n-gon. For convenience, let's also define $t_2 = 1$, even though there's no actual polygon with only two sides.

(a) Calculate t_3, t_4, t_5 explicitly, by drawing all possible triangulations.
(b) Show that $t_{n+1} = \sum_{i=2}^{n} t_i t_{n-i+2}$ *Hint: Fix some edge e of a $(n+1)$-gon. When constructing a triangulation, there are $n-1$ options for the third vertex in the triangle that is adjacent to e, and there are $n-1$ terms in the sum...*
(c) Use the formula in (b) to calculate both t_6 and t_7. Do not try to draw any triangulations, but show the work in your calculations.

EXERCISE 1.7.18 (A continuation of Exercise 1.7.17). The n^{th} *Catalan number* is

$$C_n = \frac{1}{n+1}\binom{2n}{n}.$$

These numbers come up in a bunch of different counting problems—check out the Wikipedia entry for more informations if you like.

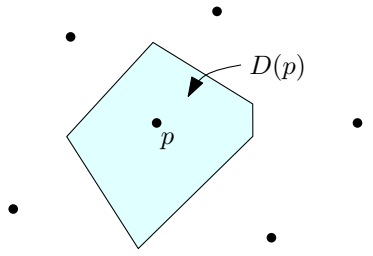
(a) Show that C_n satisfies

$$C_0 = 1, \quad C_{n+1} = \sum_{i=0}^{n} C_i C_{n-i}.$$

(b) Show that $t_n = C_{n-2}$, where t_n is as in Exercise 1.7.17.

1.8. School districts and Convexity

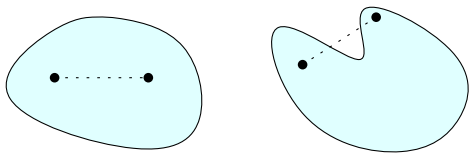
There are 26 high schools in Idealtown. How should school districts be determined so that each student attends the closest school to his/her house?



Above, schools are represented as points in the plane. The school district $D(p)$ associated to p is drawn in blue. It has an interesting feature: if both Alice and Bob

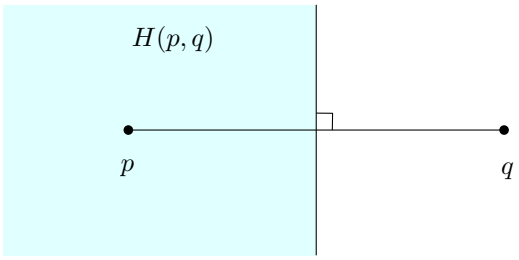
attend p, and Alice walks straight from her house to Bob's house, she will never leave their school district!

More precisely, a subset $C \subset \mathbb{R}^n$ is called *convex* if for every $x, y \in C$, the line segment xy is contained in C. We'll see why school districts are convex in a minute, but first here are two illustrative examples in $\mathbb{R}^2$:



EXERCISE 1.8.1. What are the convex subsets of $\mathbb{R}$? You don't need to prove your answer, but you can if you want.

For simplicity, let's say from now on that the district of a school p contains all points that are *at least* as close to p as to any other school. So, points on the boundaries of two districts are in both districts. Let's see what a school district looks like when there are only two schools, at p and q in $\mathbb{R}^2$. In this case, the set of points in $\mathbb{R}^2$ that are at least as close to p as to q is the *half plane* $H(p, q)$ bounded by the perpendicular bisector of the line segment ℓ from p to q; that is, it consists of everything that lies on the p side of ℓ, together with ℓ.



So, the school district $D(p)$ should just be this half-plane $H(p, q)$.

EXERCISE 1.8.2. Using the law of cosines, show that every point in $H(p, q)$ is actually closer to p than to q. *Hint: make triangles where one of the vertices is at the midpoint of the line segment pq.*

Note that the half plane $H(p, q)$ is convex, since as lines in $\mathbb{R}^2$ can intersect at most once, once a line segment leaves $H(p, q)$ it can't come back.

EXERCISE 1.8.3. Explain why $H(p, q) = \{x \in \mathbb{R}^2 \mid \frac{(x-p)\cdot(q-p)}{|q-p|} < \frac{1}{2}|q - p|\}$, and use this to give an algebraic proof that $H(p, q)$ is convex, by showing explicitly that if $x, y \in H(p, q)$ and $t \in [0, 1]$, then $tx + (1 - t)y \in H(p, q)$ as well. *Hint: For the first part, you might find the perspective of Exercise 1.1.4 useful.*

In general, there may be many schools, say at $p_1, \ldots, p_k$. The school district of p_i is then (modulo boundary issues, as before) the set of points in $\mathbb{R}^2$ that are closer to p_i than to any other p_j. In other words,

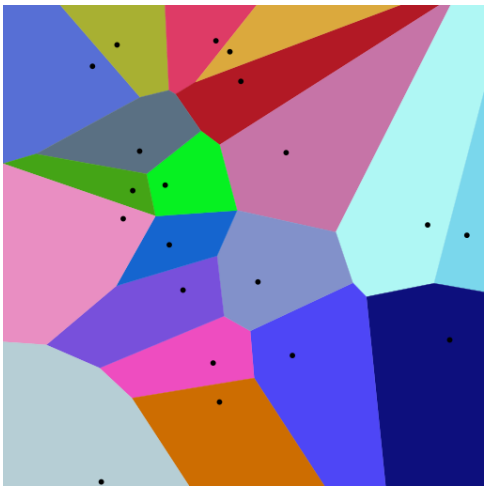
$$D(p_i) = \bigcap_{i \neq j} H(p_i, p_j).$$

LEMMA 1.8.4. *If A_i are convex subsets of $\mathbb{R}^n$, so is the intersection $\cap_i A_i$.*

PROOF. Let $x, y \in \cap_i A_i$. Applying convexity of each A_i, the line segment xy is contained in each A_i, hence in $\cap_i A_i$. So, the intersection is convex. □

So, this proves that $D(p_i)$ is convex. In fact, it is a convex (generalized) polygon, possibly with infinitely long sides, where each side is a piece of the boundary line of some $H(p_i, p_j)$.

Abstracting away from the school district problem, if $p_1, \ldots, p_k \in \mathbb{R}^2$ the set $D(p_i)$ of points that are at least as close to p_i as to any other p_j is called the *Voronoi cell* of p_i, and the cells $D(p_1), \ldots, D(p_k)$ give a *Voronoi decomposition* of $\mathbb{R}^2$.



EXERCISE 1.8.5. Suppose that $p_1, \ldots, p_5$ are the vertices of a regular pentagon. Draw the associated Voronoi decomposition of $\mathbb{R}^2$.

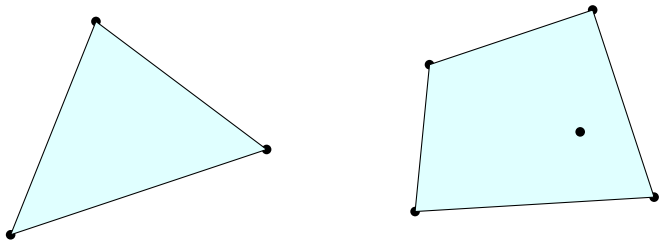
EXERCISE 1.8.6. Show more generally that if $p_1, \ldots, p_k$ are the vertices of a convex k-gon, then all Voronoi cells are unbounded. *Here, a subset $A \subset \mathbb{R}^2$ is* unbounded *if for each $r > 0$, there is an element $p \in A$ with $d(0, p) > r$. In other words, A has points that are arbitrarily far away from the origin. To prove a Voronoi cell is unbounded, you can just find an infinite length ray that is contained in it.*

1.9. Convex Hulls

In this section, we'll investigate how to take a subset of $\mathbb{R}^n$ and enclose it in a convex set in a minimal way.

DEFINITION 1.9.1. The *convex hull* of a subset $X \subset \mathbb{R}^n$, written $\text{hull}(X) \subset \mathbb{R}^n$, is the intersection of all convex sets containing X.

By Lemma 1.8.4, the convex hull is convex; it is then the smallest convex set containing X, in the sense that any other convex set containing X contains $\text{hull}(X)$. The convex hull of three non-collinear points is a triangle; the three line segments connecting the points must be in the convex hull, and then all line segments joining points on *those three segments* must also be in the convex hull, filling out the interior of the triangle. In general, the convex hull of n points in $\mathbb{R}^n$ will be a polygon, possibly with less than n vertices. For instance, on the right we have a set of 5 points in the plane whose convex hull is a quadrilateral.



A *convex combination* of points $x_1, \ldots, x_k \in \mathbb{R}^n$ is a sum of the form

$$\lambda_1 x_1 + \cdots + \lambda_k x_k,$$

where all $\lambda_i \geqslant 0$ and $\lambda_1 + \cdots + \lambda_k = 1$. When $k = 2$, these are exactly the sums $(1-t)x_1 + tx_2$, where $t \in [0,1]$, so the convex combinations of two points just make up the line segment between them.

THEOREM 1.9.2. *If $X \subset \mathbb{R}^n$, the convex hull of X is the set of all convex combinations of points in X.*

PROOF. First, we show that $\text{hull}(X)$ is contained in the set C of all convex combinations of points in X. To do this, it suffices to show that C is a convex set containing X, since we defined $\text{hull}(X)$ to be the intersection of all such sets. First, $C \supset X$, since every element $x \in X$ is the trivial convex combination $1 \cdot x$. Next, suppose

$$\sum_i t_i x_i \quad \text{and} \quad \sum_j s_j y_j,$$

are two convex combinations of elements of X, i.e. $t_i, s_j \in [0,1]$, $\sum_i t_i = \sum_j s_j = 1$ and $x_i, y_j \in X$. Then if $s, t \in [0,1]$ with $s + t = 1$, we have

$$t \sum_i t_i x_i + s \sum_j s_j y_j = \sum_i t t_i x_i + \sum_j s s_j y_j,$$

which is a convex combination of points in X since the coefficients sum to 1:

$$\sum_i tt_i + \sum_j ss_j = t\sum_i t_i + s\sum_j s_j = t + s = 1.$$

In other words, any point on the line between any two convex combinations of points in X is also a convex combination of points in X, so the set of all convex combinations is convex.

We now show that C is contained in hull(X), by induction on the number of points in the convex combination. The $k = 1$ case is trivial, since a convex combination of one point in X is just that point, which lies in hull(X). So, suppose that we know that all convex combinations of k points in X lie in hull(X), and consider

$$p = \sum_{i=1}^{k+1} t_i x_i, \quad \text{where } t_i \in (0,1), x_i \in X, \ \sum_i t_i = 1.$$

Well,

$$p = (t_1 + \cdots + t_k)\sum_{i=1}^{k} \frac{t_i}{t_1 + \cdots + t_k} x_i + t_{k+1}x_{k+1},$$

which is a point on the line segment between $q = \sum_{i=1}^{k} \frac{t_i}{t_1+\cdots+t_k} x_i$ and x_{k+1}. This q is a convex combination of k points of X, so by induction is contained in hull(X). As x_{k+1} is also contained in hull(X) and hull(X) is convex, we have $p \in$ hull(X). □

Exercise 1.9.3. Show that if X, Y are convex subsets of $\mathbb{R}^n$, then so is the subset

$$X + Y = \{x + y \mid x \in X, y \in Y\} \subset \mathbb{R}^n.$$

Exercise 1.9.4 (Cartheodory's Theorem in $\mathbb{R}^2$). Show that if $X \subset \mathbb{R}^2$, then every point in hull(X) is a convex combination of *three* points in X. *Hint: argue geometrically that any point in the convex hull of four points in $\mathbb{R}^2$ lies in the convex hull of three of the points. Then use induction.*

That is, hull(X) is the union of all triangles with vertices in X. As a challenge, you could also try proving the general version of Cartheodory's Theorem: every point in the convex hull of a set $X \subset \mathbb{R}^n$ is a convex combination of $n + 1$ points in X.

Definition 1.9.5. A function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is *convex* if when $x, y \in \mathbb{R}$ and $t \in [0, 1]$,

$$(1 - t)f(x) + tf(y) \geqslant f((1 - t)x + ty).$$

We say f is *strictly convex* if this is a strict inequality whenever $x \neq y$, $t \in (0, 1)$.

Exercise 1.9.6. If $f : \mathbb{R} \longrightarrow \mathbb{R}$, let $C_f = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_2 \geqslant f(x_1)\}$. Show that if f is convex, then C_f is a convex subset of $\mathbb{R}^2$.

Exercise 1.9.7. Show that if $f : \mathbb{R} \longrightarrow \mathbb{R}$ and $f''(x) > 0$ for all x, then f is strictly convex. *Hint: Prove that in the definition of convexity, subtracting the left side from the right gives you a function that is zero when $t = 0, 1$ and (unless $x = y$)*

has positive second derivative. Then show that such a function is always less than zero for $t \in (0,1)$. You might find the mean value theorem useful.

So for instance, the functions $f(x) = x^2$, $f(x) = e^x$, and $f(x) = -\ln(x)$ are all strictly convex, and the regions of $\mathbb{R}^2$ above their graphs are convex.

EXERCISE 1.9.8. Suppose that $f : \mathbb{R} \longrightarrow \mathbb{R}$ is convex. If we have $x_1, \ldots, x_n \in \mathbb{R}$ and $t_1, \ldots, t_n \in (0,1)$ with $t_1 + \cdots + t_n = 1$, show that

$$\sum_{i=1}^{n} t_i f(x_i) \geqslant f\left(\sum_{i=1}^{n} t_i x_i\right).$$

If f is strictly convex, show that this can be an equality only if $x_1 = \cdots = x_n$.

EXERCISE 1.9.9 (Inequality of means). If $x_1, \ldots, x_n \in \mathbb{R}$, show that

$$\frac{x_1 + \cdots + x_n}{n} \geqslant \sqrt[n]{x_1 \cdots x_n},$$

with equality if and only if $x_1 = \cdots = x_n$. *Hint: use the previous exercise and the fact that* $-\ln$ *is a decreasing, strictly convex function.*

In Exercise 1.9.9, the left side is call the *arithmetic mean*, while the right side is the *geometric mean*. One useful feature of the geometric mean is that

$$\sqrt[n]{\frac{x_1}{y_1} \cdots \frac{x_n}{y_n}} = \frac{\sqrt[n]{x_1 \cdots x_n}}{\sqrt[n]{y_1 \cdots y_n}},$$

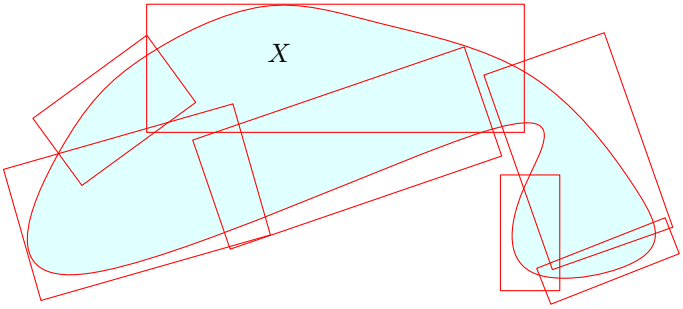
which makes the geometric mean the only suitable mean for averaging 'normalized' values. In other words, suppose that two star students each take 10 exams. Two professors then come in and each grades both the students' exams as follows. For each exam, each professor decides what point value a '100%' score will be on that exam, and then divides each student's score by that to obtain a percentage. For instance, if a student scores 64 and the professor decides that 32 points are sufficient, then the student gets 200%, while if the professor decides that the exam is out of 200, the student gets 32%. Finally, each professor *averages* all the percentages and then ranks the students according to the final average.

It turns out that if this average is an arithmetic mean, then the final rankings may be different depending on how the professors calculate the percentages for the exams! On the other hand, if a geometric mean is used, the ranking will always be the same, even though the actual percentages calculated might differ.

EXERCISE 1.9.10. Justify the previous paragraph. In particular, come up with an example in which the ranking determined by an arithmetic mean depends on how the percentages are calculated, and show how the equality of quotients for the geometric mean implies that the ranking will always be the same if the geometric mean is used.

1.10. Area

The area of a rectangle should be its base length times its height. Just as we defined path length by approximating with line segments, area of arbitrary subsets is defined by approximating with rectangles. Namely, to estimate the area of a subset $X \subset \mathbb{R}^2$, we cover X with rectangles and then sum up the areas of the rectangles.



Of course, such estimates will not be perfect, since the rectangles will overlap and will contain points not in R. Attempting to take the rectangular approximation with the least possible excess of area leads to the following definition.

DEFINITION 1.10.1. If $X \subset \mathbb{R}^2$, the *area* of X is defined by

$$\text{Area}(X) = \inf \left\{ \sum_i \text{Area}(R_i) \;\middle|\; \text{rectangles } R_1, R_2, \dots \text{ with } \cup_i R_i \supset X \right\}.$$

In the definition of path length, our piecewise linear approximations gave underestimates to the length of the path, which we then defined as a supremum, or least upper bound, of the approximating lengths. Here, our rectangular approximations give overestimates for area, so $\text{Area}(X)$ is their *infimum*, or greatest lower bound.

As with length, there may not be a rectangular approximation of X that realizes its area. For instance, if X is a point then $\text{Area}(X) = 0$, since there are rectangles containing X with arbitrarily small area. Similarly, the area of any finite set is zero.

In general, the definition of area given above is quite difficult to use. Even verifying that Definition 1.10.1 gives the right answer when X is a rectangle is not obvious! So for sanity, we will just summarize area's fundamental properties here and refer the reader to a course in measure theory for proofs.

THEOREM 1.10.2 (The fundamental theorem of area).

(a) If $f : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is an isometry and $X \subset \mathbb{R}^2$, then $\text{Area}(f(X)) = \text{Area}(X)$.

(b) If $X_1, X_2, \dots$ are disjoint measurable *subsets of $\mathbb{R}^2$, then*

$$\text{Area}(\cup_i X_i) = \sum_i \text{Area}(X_i).$$

(c) If $X = \{(x, y) \in \mathbb{R}^2 \mid f(x) \leqslant y \leqslant g(x),\ a \leqslant x \leqslant b\}$, *where* $a, b \in \mathbb{R}$ *and* $f, g : [a, b] \longrightarrow \mathbb{R}$ *are continuous functions, then*

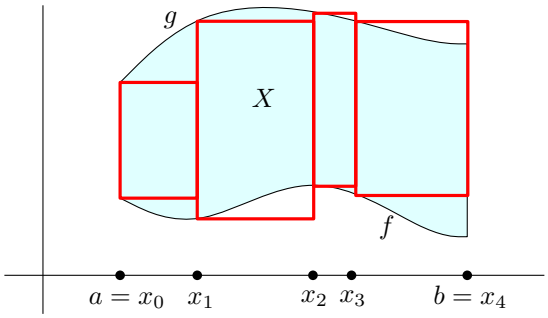
$$\text{Area}(X) = \int_a^b g(x) - f(x)\, dx.$$

The same formula also holds if any of the inequalities defining X *are strict.*

Property (a) says that congruent figures, i.e. sets that differ by an isometry, have the same area. The conclusion of Property (b) should also seem completely natural, so the presence of the mysterious 'measurability' condition is perhaps worrisome. We'll give an example of a non-measurable set at the end of this section, but for the moment we aassure the reader that every set that we will discuss in this course will be measurable. Property (c) says that area is obtained by integrating the lengths of vertical cross-sections, at least when X has a form that allows such an integral to be easily expressed. The intuition is that the integral in (b) can be approximated as

$$\int_a^b g(x) - f(x)\, dx \approx \sum_i (g(x_i) - f(x_i))(x_{i+1} - x_i),$$

where $a = x_0 < \cdots < x_n = b$ is a partition of $[a, b]$, and that this precisely corresponds to the sum of the areas of a collection of rectangles approximating the region!



As the partitions become finer, the associated Riemann sums approach the integral above, and also converge to the area of X.

Let's compute area in another example. Suppose that X is any *countable* set, i.e. a set that can be written as $X = \{x_1, x_2, \ldots\}$. Then we have

$$X = \cup_i \{x_i\},\ \Longrightarrow\ \text{Area}(X) = \sum_i \text{Area}\{x_i\} = \sum_i 0 = 0.$$

EXERCISE 1.10.3. Prove that the area of a countable subset of $\mathbb{R}^2$ is zero just using the definition of area, not the theorem.

The set of rational numbers $\mathbb{Q} \subset \mathbb{R}$ is countable. For instance, one can first list rational numbers p/q where, in lowest terms, both $|p|, |q| \leqslant 1$. We then move on

to the remaining points where $|p|, |q| \leqslant 2$, etc... Each time there only finitely many points, so assembling all these lists one after another allows us to write

$$\mathbb{Q} = \{0, 1, -1, 2, -2, \frac{1}{2}, -\frac{1}{2}, 3, -3, \frac{1}{3}, -\frac{1}{3}, \frac{2}{3}, -\frac{2}{3}, \frac{3}{2}, -\frac{3}{2}, \ldots\}.$$

On the other hand, Cantor's diagonalization argument (take an analysis class!) shows that any interval in $\mathbb{R}$ is uncountable – there are too many real numbers in any role to list them as above! Turning back to the plane, the set of points $\mathbb{Q}^2$ in $\mathbb{R}^2$ with rational coordinates is countable, while any subset of $\mathbb{R}^2$ that contains a line segment is uncountable as well.

In fact, Property (c) still holds if the roles of x and y are switched.

EXERCISE 1.10.4. Suppose that $X = \{(x, y) \in \mathbb{R}^2 \mid f(y) \leqslant x \leqslant g(y),\ a \leqslant y \leqslant b\}$, where $a, b \in \mathbb{R}$ and $f, g : [a, b] \longrightarrow \mathbb{R}$ are continuous functions. Show that

$$\text{Area}(X) = \int_a^b g(y) - f(y)\, dy.$$

In the following exercises, remember that you can only use the definition of area and the theorem above! Don't just rely on your intuition. However, you can and should *assume that all sets of interest in your proofs are measurable.*

EXERCISE 1.10.5. Prove that the area of any line in $\mathbb{R}^2$ is zero.

EXERCISE 1.10.6. If $A \subset B$, show that $\text{Area}(A) \leqslant \text{Area}(B)$.

EXERCISE 1.10.7. Show that the area of any rectangle in $\mathbb{R}^2$ is the length of its base times its height. *The most efficient proof uses both parts (a) and (c) of the theorem!*

EXERCISE 1.10.8. (For expert integrators only) Show that the area of the unit disc $D = \{v \in \mathbb{R}^2 \mid |v| \leqslant 1\}$ is 2π.

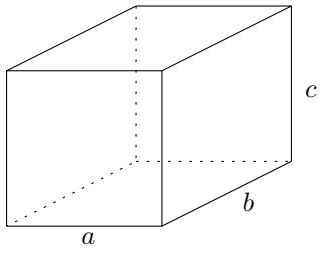
EXERCISE 1.10.9. Show that the area of a triangle is given by the formula $\frac{1}{2}\, base \times height$.

A triangle can be divided into two right triangles with the same height and whose bases add to the base of the original triangle.

EXERCISE 1.10.10. Calculate the area of a regular n-gon with vertices on the unit circle and show that these areas limit to π as $n \to \infty$.

1.11. Volume and non-measurable sets

Although our discussion in the previous section was two-dimensional, almost all of it generalizes to n-dimensions. Rectangles in $\mathbb{R}^2$ are replaced by n-dimensional rectangular solids, or *n-rectangles*, and we begin by stipulating that volume of an n-rectangle should be the product of its side lengths.



$vol = abc$

Then for the *n-dimensional volume* $\mathrm{vol}(X)$ of a set $X \subset \mathbb{R}^n$, we define

$$\mathrm{vol}(X) = \inf \left\{ \sum_i \mathrm{vol}(R_i) \;\middle|\; n\text{-rectangles } R_1, R_2, \dots \text{ with } \cup_i R_i \supset X \right\}.$$

There is a direct n-dimensional analogue of Theorem 1.10.2. Properties (a) and (b) are unchanged: isometries of $\mathbb{R}^n$ still preserve volume and volume sums under measurable disjoint unions. Versions of Property (c) also hold. For instance, if

$$X = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n \;\middle|\; \begin{array}{c} a_i \leqslant x_i \leqslant b_i \ \forall i = 1, \dots, n-1 \\ f(x_1, \dots, x_{n-1}) \leqslant x_n \leqslant g(x_1, \dots, x_{n-1}) \end{array} \right\},$$

where $a_i, b_i \in \mathbb{R}$ and f, g are continuous functions, then

$$\mathrm{vol}(X) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} g(x_1, \dots, x_{n-1}) - f(x_1, \dots, x_{n-1}) \; dx_1 \; \dots \; dx_n.$$

This discussion is even interesting in one dimension, where 'volume' should be interpreted as length. While you may not think that Property (c) makes much sense for $\mathbb{R}$, it can be interpreted to say that the volume of an interval is its length. So, for $\mathbb{R}$ the fundamental properties of volume are the following:

(a) If $f : \mathbb{R} \longrightarrow \mathbb{R}$ is an isometry and $X \subset \mathbb{R}$, then $\mathrm{vol}(f(X)) = \mathrm{vol}(X)$.
(b) If $X_1, X_2, \dots$ are disjoint and measurable, $\mathrm{vol}(\cup_i X_i) = \sum_i \mathrm{vol}(X_i)$.
(c) $\mathrm{vol}\,[a, b] = \mathrm{vol}\,[a, b) = \mathrm{vol}\,(a, b) = \mathrm{vol}\,(a, b] = b - a$.

Let's now return to the discussion of the measurability condition in Property (c). Things are a little easier in $\mathbb{R}$ than in $\mathbb{R}^2$, and the issues are all the same, so we'll stick to one dimension. Here's a first example of a non-measurable subset of $\mathbb{R}$:

Exercise 1.11.1 (*). Given a set $A \subset [0, 1)$ and an element $\alpha \in \mathbb{R}$, define

$$A + \alpha \pmod 1 = \{x + \alpha \pmod 1 \mid x \in A\} \subset [0, 1).$$

We call this set the *mod* 1 *translate of* A *by* α.

(a) Show that if $A \subset [0, 1)$ and $\alpha \in \mathbb{R}$, then $\mathrm{vol}(A) = \mathrm{vol}(A + \alpha \,(\mathrm{mod}\ 1))$.
(b) If $\alpha \in \mathbb{R}$, let $Q_\alpha = \mathbb{Q} \cap [0, 1) + \alpha \pmod 1$. Show that Q_α, Q_β are either the same or disjoint, depending on whether $\alpha - \beta \in \mathbb{Q}$ or not, and $\cup_\alpha Q_\alpha = [0, 1)$.

Part (b) means that the translates of $\mathbb{Q}$ *partition* $[0, 1)$, i.e. the interval is a disjoint union of these translates. Now form a subset $N \subset [0, 1]$ by choosing exactly one

element from each translate of Q_α. In other words, N is a subset of $[0,1)$ such that $N \cap Q_\alpha$ always contains exactly one point.

(c) Show that if $x, y \in \mathbb{Q}$, then $N + x \pmod 1 \cap N + y \pmod 1 = \varnothing$ unless $x \pmod 1 = y \pmod 1$. Then show that $\cup_{x\in\mathbb{Q}} N + x \pmod 1 = [0,1)$.

(d) Show that the sets $N + x \pmod 1$, where $x \in \mathbb{Q}$, cannot all be measurable, using parts (a) and (c) above. *(In fact, none of them are measurable.)*

When constructing N in the exercise above, we made implicit use of the 'axiom of choice', which tells us that if we have a collection of sets, we can create a new set by choosing one element from each. In fact, Solovay proved that any construction of a non-measurable set *must* use the axiom of choice, which in particular means that it is impossible to write down an actual formula for a non-measurable set!

For an even more dramatic example, the *Banach-Tarski paradox* states that a solid ball in $\mathbb{R}^3$ can be cut into five pieces and reassembled into two solid balls, each with the same volume as the first! The point is that the pieces are not measurable, and their volumes sum to more than that of the original ball.

1.12. Isoperimetric inequalities

How does one enclose the largest amount of area using a rope of a certain length? Physical experience suggests that one should lay the rope along a circle. In this section, we will formulate this problem mathematically and present a solution!

A *loop* in $\mathbb{R}^2$ is just an injective path $\gamma : [a,b] \longrightarrow \mathbb{R}^2$ such that $\gamma(a) = \gamma(b)$. Injectivity means that the path cannot cross itself, while the latter condition ensures that it closes up. The celebrated *Jordan curve theorem*, which is intuitively obvious but actually quite involved to prove, says that any loop separates $\mathbb{R}^2$ into two pieces, one bounded and the other unbounded, i.e. the inside and outside.

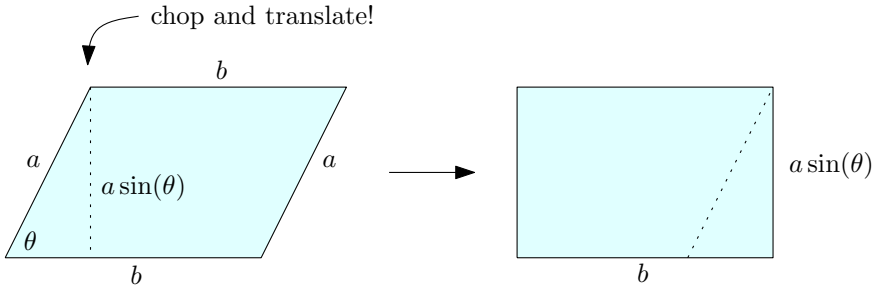
THEOREM 1.12.1. *Among all loops γ in $\mathbb{R}^2$ with length at most l, any loop that maximizes the enclosed area must be a circle.*

As a circle with circumference l has radius $\frac{l}{2\pi}$, the area enclosed is $\pi\left(\frac{l}{2\pi}\right)^2 = \frac{\pi}{4}l^2$. One would like to then use the theorem to say that a general loop with length l never encloses a region with area greater than $\frac{\pi}{4}l^2$. However, note the careful wording of the theorem! Nowhere is it actually stated that there *is* a loop that encloses a maximal amount of area, and if there is not then the theorem has no content. Of course, as you might expect, there is actually such a loop, but a proof of its existence is more difficult than what we'll do below.

Before beginning the proof, let's do a similar problem involving parallelograms.

LEMMA 1.12.2. *Fix $a, b > 0$. Among all the parallelograms with side lengths a and b, the rectangle encloses the most area.*

PROOF. The area of such a parallelogram is $ab\sin(\theta)$, where θ is pictured below.

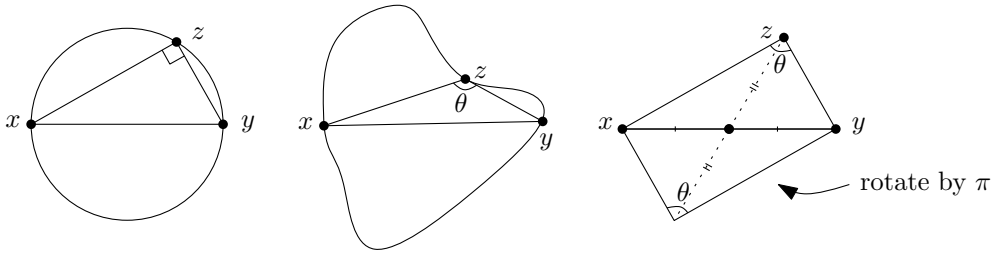


This is maximal when $\sin(\theta)$ is maximal, i.e. when $\theta = \pi/2$, which happens if and only if the parallelogram is a rectangle. □

To prove Theorem 1.12.1, we need a criterion that characterizes circles.

LEMMA 1.12.3. *A loop γ describes a circle if and only if for all x, y, z on γ such that x, y divide γ in half by length, the segments xz and yz meet at right angles.*

PROOF. The figure below illustrates a right angle when γ is a circle, and a non-right angle when γ is not a circle. To prove that this is always the case, rotate a copy of the triangle xyz by π and adjoin it to xyz to make a parallelogram $xyzz'$.



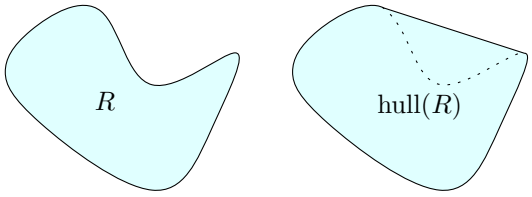
The diagonals of the parallelogram bisect each other, and have the same length if and only if the parallelogram is a rectangle, i.e. if the angle $\theta = \pi/2$.

Now, if γ describes a circle, the segment xy is a diameter, and hence its midpoint is the center of the circle. As x, y, z are all equidistant from the center, the diagonals of the parallelogram $xyzz'$ have the same length, so the angle at z is $\pi/2$.

Conversely, if the angle at z is always $\pi/2$, the diagonals of $xyzz'$ always have the same length, so the distance from z to the midpoint of xy is always half the length of xy. Therefore, γ is a circle centered at the midpoint of xy with radius $\frac{1}{2}d(x, y)$. □

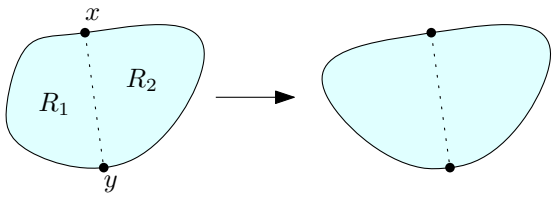
We are now ready for the proof of Theorem 1.12.1. As the argument we will give is a little more hand wavy than others so far, we will call it a 'proof sketch'.

PROOF SKETCH. Assume that γ has length at most l and encloses a region R with maximal area. As in Section 1.8, the convex hull $\text{hull}(R)$ is the smallest convex set containing R; intuitively, it is the region enclosed by a rubber band snapped around R.

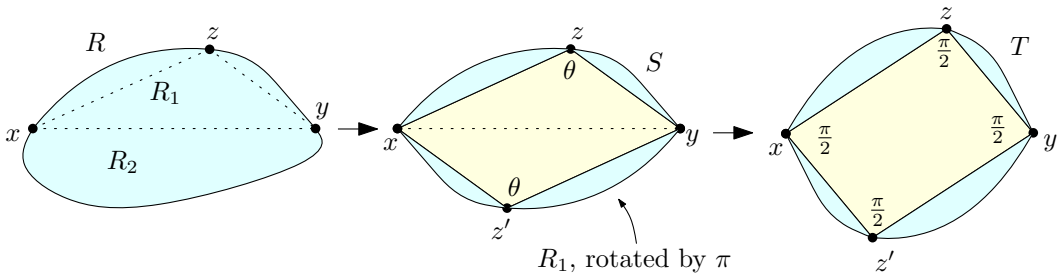


The perimeter of hull(R) consists of pieces of γ interspersed with a number of line segments, each of which replaces a (longer, by Theorem 1.5.7) portion of γ with the same endpoints. Unless R is already convex, hull(R) is a region of larger area enclosed by a loop with length at most l, contradicting maximality of R. Therefore, R is convex.

Let x, y be points on γ that separate it into two paths of equal length. By convexity, the line segment xy llies in R, and separates it into two pieces, R_1 and R_2. These two pieces must have the same area, since otherwise we could reflect the larger piece across xy to create a new region with perimeter at most l and larger area.



If z is another point on γ, it suffices by Lemma 1.12.3 to show that the segments xz and yz meet at a right angle. Assume without loss of generality that z lies on the half of γ adjacent to R_1. Replace R_2 with a copy of R_1 rotated by π, and call this new region S, as pictured below. The triangle xyz and its rotated twin join to form a parallelogram $xyzz' \subset S$, as below in yellow.



If the angle at z is not $\pi/2$, we know by Lemma 1.12.2 that the rectangle with the same side lengths has greater area than $xyzz'$. So, create a new region T by pasting the four blue pieces of $S \backslash xyzz'$ onto the sides of such a rectangle. This T has the same perimeter as R, but has larger area, contradicting maximality of R. □

Exercise 1.12.4. At the end of the proof of Theorem 1.12.1, you may have wondered why the four pieces of $S \backslash xyzz'$ do not intersect each other when they are stuck onto the rectangle to form T. This is your chance to explain!

EXERCISE 1.12.5. Fix $l > 0$. Show that among parallelograms with perimeter l, the square with side lengths $l/4$ maximizes area.

EXERCISE 1.12.6 (Heron's formula). If T is a triangle with side lengths a, b, c, show

$$\text{Area}(T) = \sqrt{s(s-a)(s-b)(s-c)}, \text{ where } s = \frac{a+b+c}{2}.$$

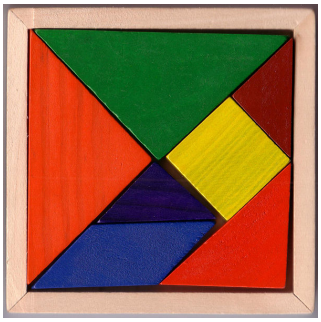
Hint: use the law of cosines and the identity $\sin^2 + \cos^2 = 1$ *to express the* sin *of one of the angles of the triangle just in terms of the side lengths. Then, use this* sin *to find the height of the triangle, plug that into the area formula for triangles and do a bunch of algebraic manipulations. There may come a point where you just want to explain loosely why a product of four terms eventually simplifies to something – feel free to just do that instead of writing out the entire calculation.*

EXERCISE 1.12.7. Show that among triangles with perimeter l, the area is largest for equilateral triangles. *Hint: apply Heron's formula, and use the inequality of means described in Exercise 1.9.9.*

More generally, regular n-gons maximize area among n-gons with specified perimeter. As a challenge, try to think about a proof of this statement!

1.13. Tangrams and Scissors Congruence

The *tangram* is a puzzle in which one is given a set of seven pieces (five triangles of varying sizes, a parallelogram and a square) and is asked to arrange them into prescribed configurations. Only the outline of the desired shape is given, and the appropriate configuration can be difficult to find.



a tangram set

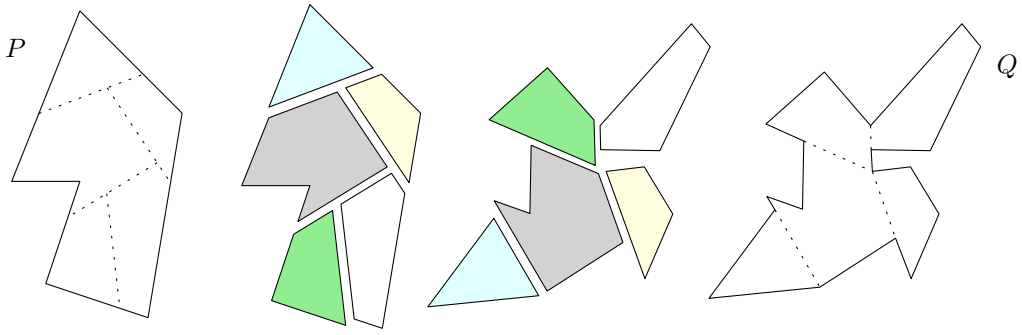
the puzzles

The puzzle originated in China. Although the creator of the game is no longer known, you can find many fictionalized origin stories on the Internet. Many of them begin with a sentence like "Once upon a time, a man had a treasured clay tile..."; often, you can imagine the rest. There are even creation myths based on tangrams! Try googling tangram history if you want to take a trip down the rabbit hole.

In 1815, the tangram puzzle was brought from China to the US on the ship *Trader*, by Capt. M. Donaldson. It was then exported to Britain, Germany and Denmark. Also known as "the anchor puzzle" and "the Sphinx", it became one of the most popular games of the 19th century in America and Europe. One reason for the popularity of such puzzles at the time was that the Catholic Church tolerated playing them on the Sabbath.

There is some interesting mathematics related to the tangram puzzle.

DEFINITION 1.13.1. Two subsets P, Q of $\mathbb{R}^2$ are *scissors congruent* if they are the unions of polygons $P_1, \ldots, P_n$ and $Q_1, \ldots, Q_n$, respectively, intersecting only on their edges, such that P_i and Q_i are congruent for each i.

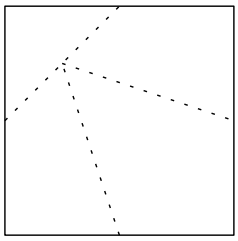


In other words, P and Q are scissors congruent if one can be cut along line segments and reassembled into the other. As an example, any tangram puzzle that has a solution must be scissors congruent to a square!

EXERCISE 1.13.2. If P and Q are scissors congruent and Q and R are scissors congruent, show that P and R are scissors congruent. In other words, scissors congruence is an 'equivalence relation'.

Note that the subsets P, Q in the definition of scissors congruence may be polygons, but they may also be unions of disjoint polygons! For instance, the two unions of polygons in the middle of the figure above are both scissors congruent to P and Q. Even if one is only interested in polygons, this extended point of view is useful, since it is often useful to use the Exercise repeatedly to prove that polygons are scissors congruent by passing through intermediate subsets of $\mathbb{R}^2$ that are not polygons.

EXERCISE 1.13.3. Cut a square along the following line segments. Show that the resulting pieces can be rearranged into an isosceles triangle.



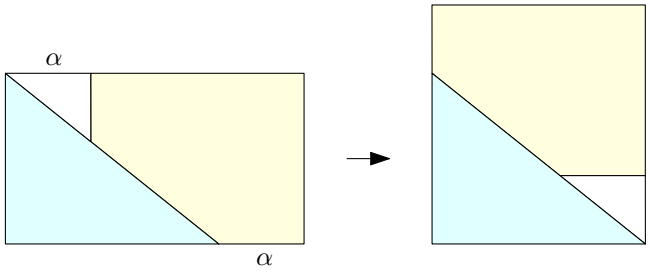
So, when are two polygons scissors congruent? Well, certainly the two polygons must have the same area. In fact, the converse is true:

THEOREM 1.13.4 (Wallace-Bolyai-Gerwein). *Two polygons A and B are scissors congruent if and only if they have the same area.*

A word is in order about the triple attribution. Some sources say that the problem was posed by Bolyai, then solved by Gerwein in 1833 and by Wallace in 1807. Others say that it was proved by Bolyai in 1835. So to be safe, we give credit to everyone!

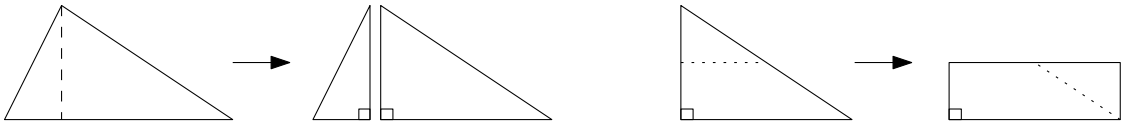
LEMMA 1.13.5. *Any two rectangles with the same area are scissors congruent.*

PROOF. The following move on rectangles is called a 'P-slide' - the only constraint here is that α is less than or equal to half the width of the rectangle.



Using a P-slide or its inverse, a rectangle with width a is scissors congruent to any rectangle with the same area and width in $[a/2, 2a]$. So, repeated P-slides can be used to show that any two rectangles with the same area are scissors congruent. □

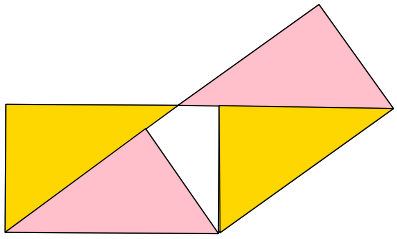
PROOF OF THEOREM 1.13.4. Let A be a polygon. We will show that A is scissors congruent to a square with the same area. Cut A into triangles using Theorem 1.7.3. Each triangle can be cut into two right triangles, which can be reassembled into a rectangle.



Using a P-slide, alter each rectangle so that its width is $\sqrt{\text{Area}(A)}$. Stacking the rectangles must give a square, since it is a rectangle with the same area as A. □

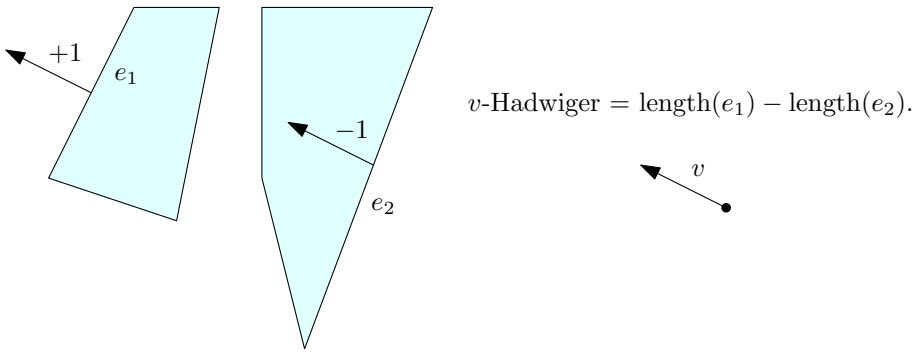
Two subsets P, Q of $\mathbb{R}^2$ are *scissors congruent via translations* if they are the unions of polygons $P_1, \ldots, P_n$ and $Q_1, \ldots, Q_n$, respectively, intersecting only on their edges, such that P_i and Q_i differ by a translation for each i. Similarly, one could consider 'scissors congruence via translations and rotations by π'. The point is that now we are limiting the movement of the polygons to certain isometries.

EXERCISE 1.13.6. (Hard) Show that any two rectangles with the same area are scissors congruent via translations. *Hint: P-slides can be used to alter the dimensions of a rectangle. The real trick is to say why you can rotate a rectangle. Here's a hint:*



EXERCISE 1.13.7. Show that any two polygons with the same area are scissors congruent via translations and rotations by π. *Hint: repeat the proof of Theorem 1.13.4 using Exercise 1.13.6.*

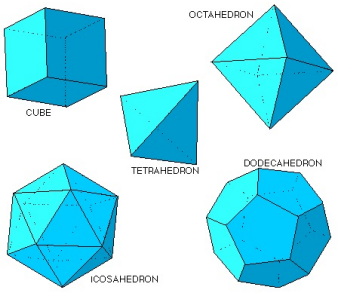
Given a vector $v \in \mathbb{R}^2$, the *v-Hadwiger invariant* of a collection of polygons is the real number obtained by summing up the signed lengths of all edges perpendicular to v, where the sign of an edge is $+1$ if v points outward and -1 if v points inward.

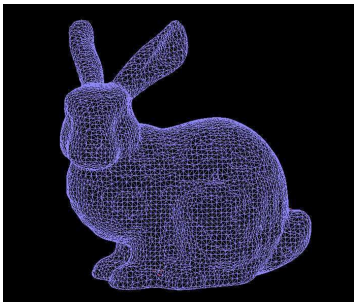


EXERCISE 1.13.8. Show that if $v \in \mathbb{R}^2$, the v-Hadwiger invariants of two collections of polygons that are scissors congruent via translations must be equal. Use this to give an example of two polygons that are not scissors congruent via translations.

1.14. Polyhedra and the Dehn invariant

Loosely[1], a *polyhedron* is a solid in $\mathbb{R}^3$ bounded by a collection of polygons (*faces*) that meet along their edges. Here are some examples of polyhedra.

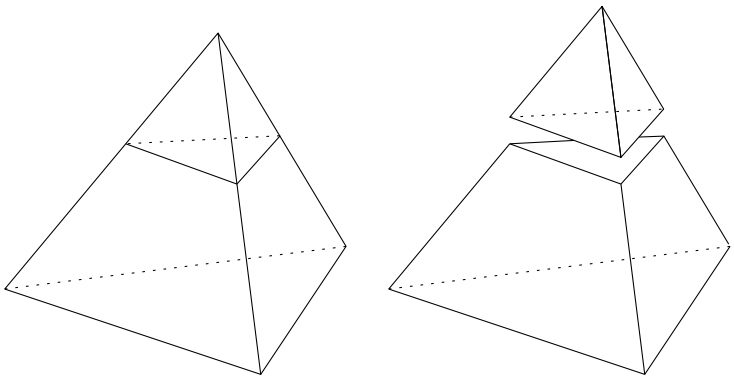




The polyhedra on the left are the *Platonic solids*, which may be familiar from high school geometry. On the right is the famous 'Rabbitic solid', which has thousands of faces, but not the one that counts.

Just as for polygons, we say that two subsets P, Q of $\mathbb{R}^3$ are *scissors congruent* if they are the unions of polyhedra $P_1, \ldots, P_n$ and $Q_1, \ldots, Q_n$, respectively, intersecting only on their faces, such that P_i and Q_i are congruent for each i.

Again, scissors congruence is an equivalence relation, and the cut-and-reassemble picture is the same as before, except that we cut along planes instead of lines.



The volume of a *polygon* in $\mathbb{R}^3$ is zero, so volume sums when polyhedra are glued along their polygonal faces. So, scissors congruent subsets of $\mathbb{R}^3$ have the same volume.

QUESTION. Is it true that any two polyhedra in $\mathbb{R}^3$ with the same volume are scissors congruent?

[1]It's surprisingly difficult to give a good definition of a polyhedron that conforms exactly to one's intuition, and many early treatments of polyhedra suffered from the lack of a precise definition. The definition we give here is a little bit vague, but we'll be content with it and use our intuition. Note that for instance, if you put a delete a small cube from the interior of a bigger cube, the result is a polyhedron under most reasonable interpretations of our definition.

In 1900, the mathematician David Hilbert devised a list of 23 problems to focus research in the 20th century. The innocuous question above was the third problem. Three months later, it was solved by Hilbert's student Max Dehn.

THEOREM 1.14.1 (Dehn, 1900). *A cube and regular tetrahedron of the same volume are not scissors congruent.*

A tetrahedron is featured in the picture above. Regular tetrahedrons are those that have all their side lengths and dihedral angles equal. Here, a 'dihedral angle' is the angle at which two faces meet along an edge.

EXERCISE 1.14.2. The four points $(\pm 1, 0, -1/\sqrt{2})$, $(0, \pm 1, 1/\sqrt{2})$ in $\mathbb{R}^3$ form the vertices of a regular tetrahedron.

(a) Show that indeed, all the distances between pairs of these points are the same. (In your write-up, just do 3 of the 6 pairs.)
(b) Pick some edge of the tetrahedron, and show that the associated dihedral angle is $\cos^{-1}(1/3)$. (All edges will give you the same dihedral angle, but I'm only asking you to do the computation for one edge, of your choice.)

Hint: for (b) you'll need to do a little bit of multivariable calculus to compute dihedral angles. The point is that the angle at which two planes intersect is the same as the angle between two vectors that are respectively perpendicular ('normal') to the two planes. To find these normal vectors, note that if vectors v, w are not scales of each other, and both lie along the plane P (i.e. their heads and tails both lie in P) then a normal vector for P is the cross product

$$v \times w = \det \begin{pmatrix} i & j & k \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix} = (v_2w_3 - w_2v_3)i - (v_1w_3 - w_1v_3)j + (v_1w_2 - w_1v_2)k.$$

Here, we are using the multivariable calculus notation $(a, b, c) = ai + bj + ck$, so in our usual notation the cross product is $v \times w = (v_2w_3 - w_2v_3, w_1v_3 - v_1w_3, v_1w_2 - w_1v_2)$.

We will give a proof of Dehn's theorem in the remainder of the section. The point is to come up with an appropriate *invariant* – a number that one can associate to a union of polyhedra that does not change under scissors congruence. This requires a sort of lengthy digression into some (fascinating) algebra.

We say that a function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is *additive* if $f(x) + f(y) = f(x + y)$ for all $x, y \in \mathbb{R}$. As an example, if $c \in \mathbb{R}$, then the function

$$f : \mathbb{R} \longrightarrow \mathbb{R}, \ f(x) = cx$$

is additive, since $c(x + y) = cx + cy$. We call additive functions of this type *linear*. So, are there any nonlinear additive functions? We will prove:

PROPOSITION 1.14.3. *There is an additive function $f : \mathbb{R} \longrightarrow \mathbb{R}$ such that*

$$f(\cos^{-1}(1/3)) = 1, \quad f(\pi) = 0.$$

Note that this f cannot be linear, since it cannot be that $c\pi = 0$ while $c\cos^{-1}(1/3) = 1$. And what is truly bizarre is that the proof of Proposition 1.14.3 just guarantees the *existence* of such an f, it does not actually construct one explicitly. In fact, it is provably impossible to write down a formula for a nonlinear additive function!

EXERCISE 1.14.4. Show that if f is additive, $f(qx) = qf(x)$ for all $x \in \mathbb{R}$ and $q \in \mathbb{Q}$.

Assuming Proposition 1.14.3, let's see how to prove that a cube and a regular tetrahedron are never scissors congruent. Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be an additive function such that $f(\pi) = 0$, so for instance f could be the function from Proposition 1.14.3. The *Dehn invariant* D_f associated to f is defined as follows. If P is a union of polyhedra, let

$$\text{length}(e) \quad \text{and} \quad \angle(e)$$

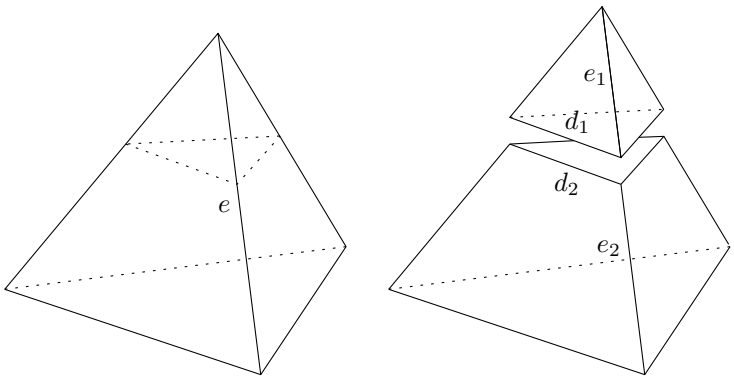
be the length and dihedral angle of an edge e of P, and define

$$D_f(P) = \sum_{\text{edges } e \text{ of } P} \text{length}(e) f(\angle(e)).$$

THEOREM 1.14.5. *If P and Q are unions of polyhedra that are scissors congruent, then $D_f(P) = D_f(Q)$ for any additive f such that $f(\pi) = 0$.*

PROOF. We just need to show that D_f does not change when a polyhedron of P is cut by a plane, for moving the pieces around by isometries does not change D_f.

To do this, we examine the effect that a cut has on a given term $\text{length}(e)f(\angle(e))$ of the summation defining $D_f(P)$. Clearly, this term is only affected if the cutting plane intersects e.



Suppose first that the cut divides e into two edges e_1 and e_2, as in the picture above. The joint contribution of e_1 and e_2 to D_f is then the same as that of e:

$$\begin{aligned} \text{length}(e)f(\angle(e)) &= (\text{length}(e_1) + \text{length}(e_2))f(\angle(e)) \\ &= \text{length}(e_1)f(\angle(e)) + \text{length}(e_2)f(\angle(e)) \\ &= \text{length}(e_1)f(\angle(e_1)) + \text{length}(e_2)f(\angle(e_2)). \end{aligned}$$

So, any change in D_f cannot come from the e edges. Similarly, if the edge e actually lies on the cutting plane, then after the cut, e becomes two edges e_1 and e_2, each of which has the same length as e, and where $\angle(e_1) + \angle(e_2) = \angle(e)$. So,

$$\begin{aligned}\text{length}(e_1)f(\angle(e_1)) + \text{length}(e_2)f(\angle(e_2)) &= \text{length}(e)\Big(f(\angle(e_1)) + f(\angle(e_2))\Big)\\ &= \text{length}(e)f(\angle(e)).\end{aligned}$$

But wait, you say, there are additional edges introduced by these cuts that we have not accounted for! These new edges come in pairs, e.g. d_1 and d_2 above, where the edges in a pair have the same length and have dihedral angles summing to π. So,

$$\text{length}(d_1)f(\angle(d_1)) + \text{length}(d_2)f(\angle(d_2)) = 0,$$

by additivity of f and the fact that $f(\pi) = 0$. This means that the new edges introduced by the cut do not contribute to D_f. Thus, D_f is unchanged when a polyhedron of P is cut by a plane. □

We can now prove the main result of this section, that a cube and a regular tetrahedron are never scissors congruent.

Proof of Theorem 1.14.1. Now suppose that both $f(\pi) = 0$ and $f(\cos^{-1}\left(\frac{1}{3}\right)) = 1$, as in Proposition 1.14.3, and let C and T be a cube and a regular tetrahedron. Since the dihedral angles of C and T are $\pi/2$ and $\cos^{-1}\left(\frac{1}{3}\right)$, respectively, we see that

$$\begin{aligned}D_f(C) &= 12 \cdot \ell_C \cdot f\left(\frac{\pi}{2}\right) = 12 \cdot \frac{1}{2} \cdot f(\pi) = 0,\\ D_f(T) &= 4 \cdot \ell_T \cdot f\left(\cos^{-1}\left(\frac{1}{3}\right)\right) \neq 0,\end{aligned}$$

where here, ℓ_C and ℓ_T are the lengths of the edges in C and T, respectively. Since $D_f(C) \neq D_f(T)$, Theorem 1.14.5 says that C and T are not scissors congruent. □

1.14.1. Linear algebra over $\mathbb{Q}$: a proof of Proposition 1.14.3. We now prove that there is an additive function f with $f(\cos^{-1}(1/3)) = 1$ and $f(\pi) = 0$, as required by Proposition 1.14.3. The proof involves some elementary number theory, and some arguments that you may have seen in linear algebra, adapted to a new setting.

The existence of such an f doesn't have much to do with the particular numbers given; rather, what matters is that $\cos^{-1}(1/3)$ and π are not *rational multiples of each other*. Let's prove this first. To do so, we'll need the following tool, which one can use to prove that numbers are irrational.

Theorem 1.14.6 (The Rational Root Theorem). *If $x = \frac{p}{q}$ is a fraction in lowest terms that is a solution of $a_n x^n + \cdots + a_0 = 0$, where each $a_i \in \mathbb{Z}$, then $p|a_0$ and $q|a_n$.*

PROOF. As $a_n(\frac{p}{q})^n + \cdots + a_0 = 0$, multiplying by q^n and shifting the constant term,

$$p\left(a_n p^{n-1} + \cdots + a_1 q^{n-1}\right) = -a_0 q^n.$$

As p, q are co-prime, so are p, q^n. So, as the expression in parentheses is an integer, p divides a_0. The proof that $q|a_n$ is similar. □

A polynomial $f(x) = a_n x^n + \cdots + a_0$ with integer coefficients is called an *integer polynomial*. It is called *monic* if $a_n = 1$. We call a number $x \in \mathbb{R}$ such that $f(x) = 0$ a *root* of f.

COROLLARY 1.14.7. *If $x \in \mathbb{Q}$ is a root of a monic, integer polynomial, then $x \in \mathbb{Z}$.*

PROOF. Suppose $(\frac{p}{q})^n + a_{n-1}(\frac{p}{q})^{n-1} + \cdots + a_0 = 0$, where p/q is in lowest terms. By the Rational Root Theorem, $q|1$, so $q = \pm 1$, implying $\frac{p}{q}$ is an integer. □

As an application, note that if $m \in \mathbb{N}$, then $x = \sqrt{m}$ is a root of $x^2 - m = 0$, so $\sqrt{m}$ is only rational when m is a perfect square. In particular, $\sqrt{2}$ is irrational.

Here is the application that we are most interested in, though.

PROPOSITION 1.14.8. *Within the interval $[-\pi, \pi]$, the only rational multiples of π that have rational cosine are $0, \pm\frac{\pi}{3}, \pm\frac{\pi}{2}, \pm\pi$.*

Note that the cosines of the angles $0, \pm\frac{\pi}{3}, \pm\frac{\pi}{2}, \pm\pi$ are $1, \pm\frac{1}{2}, 0, -1$, respectively. So, if q is any rational number except $1, \pm\frac{1}{2}, 0, -1$, Proposition 1.14.8 implies that $\cos^{-1}(q)$ is not a rational multiple of π. In particular,

COROLLARY 1.14.9. $\cos^{-1}(\frac{1}{3})$ *is not a rational multiple of π.*

The proof of the proposition is an application of Corollary 1.14.7.

PROOF OF PROPOSITION 1.14.8. By the angle sum formula, for $\alpha \in \mathbb{R}$ and $n \in \mathbb{N}$, we have

$$\begin{aligned}
&\cos\big((n+1)\alpha\big) + \cos\big((n-1)\alpha\big) \\
&= \cos(n\alpha)\cos(\alpha) - \sin(n\alpha)\sin(\alpha) + \cos(n\alpha)\cos(-\alpha) - \sin(n\alpha)\sin(-\alpha) \\
&= 2\cos(n\alpha)\cos(\alpha).
\end{aligned}$$

So, setting $x = 2\cos(\alpha)$ and '$P_n(x)$'$= 2\cos(n\alpha)$, this implies

$$P_{n+1}(x) = xP_n(x) - P_{n-1}(x), \quad P_1(x) = x$$

By induction, $P_n(x)$ must be a monic degree n polynomial in x, that is a polynomial whose leading term is x^n, with coefficient 1.

Now if $\alpha = \frac{m}{n}\pi$, we have $P_n(x) = 2\cos(n\frac{m}{n}\pi) = 2(-1)^m$. So, $x = 2\cos(\alpha)$ is a root of the monic polynomial

$$P_n(x) - 2(-1)^m = 0.$$

If $\cos\alpha$ is a rational, so is $x = 2\cos(\alpha)$, so Corollary 1.14.7 implies that x is an integer, implying $\cos(\alpha)$ is one of $0, \pm\frac{1}{2}, \pm 1$. This implies α is one of $0, \pm\frac{\pi}{3}, \pm\frac{\pi}{2}, \pm\pi$. □

So, now we know that $\cos^{-1}(1/3)$ and π are not rational multiples of each other, and we want to conclude that there is an additive f with $f(\cos^{-1}(1/3)) = 1$ and $f(\pi) = 0$. To do this, it will be convenient to adopt a more general viewpoint. *Note: if you have taken an abstract linear algebra class that works over arbitrary fields, you may have seen much of the following material, although you may not have thought to consider $\mathbb{R}$ as a vector space over $\mathbb{Q}$.*

DEFINITION 1.14.10. A subset $S \subset \mathbb{R}$ is $\mathbb{Q}$-*independent* if whenever $q_1 s_1 + \cdots + q_n s_n = 0$, with $q_i \in \mathbb{Q}, s_i \in S$, then $q_1 = \cdots = q_n = 0$.

For instance, if $x \neq 0$, then $\{x\}$ is $\mathbb{Q}$-independent, since $qx = 0$ implies $q = 0$. Here, we call an expression of the form $q_1 s_1 + \cdots + q_n s_n$, where $q_i \in \mathbb{Q}$, a $\mathbb{Q}$-*linear combination* of the elements s_i.

LEMMA 1.14.11. *If $a \neq 0$, then $\{a, b\}$ is $\mathbb{Q}$-independent if and only if b is not a rational multiple of a.*

PROOF. Suppose first that b is a rational multiple of a, so $b = qa$, where $a \in \mathbb{Q}$. Then $b - qa = 0$, contradicting that $\{a, b\}$ is $\mathbb{Q}$-independent.

Conversely, suppose that $\{a, b\}$ is not $\mathbb{Q}$-independent, i.e. there exist $q, r \in \mathbb{Q}$, not both zero, such that $qa + rb = 0$. If $r = 0$, then $qa = 0$, implying that $q = 0$ since $a \neq 0$. This cannot happen since q, r aren't both zero. So, $r \neq 0$. Hence, $b = (-q/r)a$ is a rational multiple of a. □

So for instance, $\{1, \sqrt{2}\}$ and $\{\cos^{-1}(1/3), \pi\}$ are both $\mathbb{Q}$-independent sets.

EXERCISE 1.14.12. Prove that there is no triple of integers (m, n, p) except $(0, 0, 0)$ such that $m + n\sqrt{2} + p\sqrt{3} = 0$. *Hint: move m to the other side and square both sides. Clearing the denominators, this implies the same result where m, n, p are rational numbers. In other words, $\{1, \sqrt{2}, \sqrt{3}\}$ is a $\mathbb{Q}$-linearly independent subset of $\mathbb{R}$.*

The connection with additive functions is the following.

THEOREM 1.14.13. *If $S \subset \mathbb{R}$ is a $\mathbb{Q}$-independent set and for each $s \in S$ we have some real number a_s, then there is an additive function $f : \mathbb{R} \to \mathbb{R}$ such that $f(s) = a_s$ for all $s \in S$.*

In other words, the values of an additive function on a $\mathbb{Q}$-independent set can be prescribed arbitrarily. In contrast, note that if $qa + rb = 0$ for some $q, r \in \mathbb{Q}$, then using Exercise 1.14.4, if f is additive

$$0 = f(qa + rb) = qf(a) + rf(b),$$

so there is a definite relationship between $f(a)$ and $f(b)$; one cannot prescribe their values independently of each other. Note that as $\{\cos^{-1}(1/3), \pi\}$ is $\mathbb{Q}$-independent, it follows that there is an additive function $f : \mathbb{R} \longrightarrow \mathbb{R}$ such that

$$f(\cos^{-1}(1/3)) = 1, \quad f(\pi) = 0,$$

so *Proposition 1.14.3 is a direct corollary.*

We will now proceed toward a proof of Theorem 1.14.13. We define:

DEFINITION 1.14.14. If $S \subset \mathbb{R}$, the $\mathbb{Q}$-*span* of S is

$$span_{\mathbb{Q}}(S) = \{q_1 s_1 + \cdots q_n s_n \mid q_i \in \mathbb{Q}, s_i \in S\}.$$

For example, note that $span_{\mathbb{Q}}(\{1\}) = \mathbb{Q}$, and $span_{\mathbb{Q}}(\{1, \sqrt{2}\})$ is, well, the set of all numbers that can be written as $q + r\sqrt{2}$, where $q, r \in \mathbb{Q}$.

LEMMA 1.14.15. *Suppose that $S \subset \mathbb{R}$ is $\mathbb{Q}$-independent, and $x \in \mathbb{R} \backslash span_{\mathbb{Q}}(\mathbb{R})$. Then $S \cup \{x\}$ is $\mathbb{Q}$-independent.*

PROOF. Suppose that we have $q_1 s_1 + \cdots + q_n s_n + rx = 0$, where $q_i, r \in \mathbb{Q}$, $s_i \in S$. We want to show that all the coefficients q_i and r are zero. If $r = 0$, then this is a $\mathbb{Q}$-linear combination of the s_i, so as S is $\mathbb{Q}$-independent, all the coefficients q_i must be zero, and we are done. Otherwise, if $r \neq 0$, then we have

$$x = (-q_1/r)s_1 + \cdots + (-q_n/r)s_n \in span_{\mathbb{Q}}(\mathbb{R}),$$

a contradiction. □

DEFINITION 1.14.16. A $\mathbb{Q}$-*basis* for $\mathbb{R}$, or alternatively a *Hamel basis*, is a $\mathbb{Q}$-independent set $S \subset \mathbb{R}$ such that $span_{\mathbb{Q}}(S) = \mathbb{R}$.

LEMMA 1.14.17. *$S \subset \mathbb{R}$ is a $\mathbb{Q}$-basis if and only if every $x \in \mathbb{R}$ can be written* uniquely *as a $\mathbb{Q}$-linear combination $x = q_1 s_1 + \cdots + q_n s_n$, where $q_i \in \mathbb{Q}$ and $s_i \in S$.*

PROOF. Suppose $S \subset \mathbb{R}$ is a $\mathbb{Q}$-basis and let $x \in \mathbb{R}$. Since $x \in span_{\mathbb{Q}}(S)$, we can write x as a $\mathbb{Q}$-linear combination of elements of S. So, assume that we can write x as such a combination in two different ways. Now, given any $\mathbb{Q}$-linear combination, we can certainly add on other elements of S multiplied by the coefficient zero, without changing the result. So, we can assume that our two linear combinations include the same elements of S:

$$x = q_1 s_1 + \cdots + q_n s_n = r_1 s_1 + \cdots + r_n s_n.$$

But then $(q_1 - r_1)s_1 + \cdots + (q_n - r_n)s_n = 0$, so $\mathbb{Q}$-independence implies that $q_i = r_i$ for all i.

Now suppose that every $x \in \mathbb{R}$ can be written *uniquely* as a $\mathbb{Q}$-linear combination $x = q_1 s_1 + \cdots + q_n s_n$, where $q_i \in \mathbb{Q}$ and $s_i \in S$. Clearly, $span_{\mathbb{Q}}(S) = \mathbb{R}$, and since 0 can be written as a $\mathbb{Q}$-linear combination with only zero coefficients, that is the only way to write 0 as such, so S is $\mathbb{Q}$-independent. □

So, do $\mathbb{Q}$-bases exist? Yes, in abundance!

THEOREM 1.14.18. *If $S \subset \mathbb{R}$ is $\mathbb{Q}$-independent, then there is a $\mathbb{Q}$-basis $T \subset \mathbb{R}$ that contains S.*

PROOF SKETCH. The idea is simple. Start with S. If $span_{\mathbb{Q}}(S) = \mathbb{R}$, we're done. If not, take some $x_1 \in \mathbb{R} \backslash span_{\mathbb{Q}}(S)$. By Lemma 1.14.15, $S \cup \{x_1\}$ is $\mathbb{Q}$-independent. So, now we just continue this process, each time adding some $x_i + 1$ outside the span of our current set $S_i := S \cup \{x_1, \ldots, x_i\}$, and preserving independence in every step. One would like to say that this terminates with some S where $span_{\mathbb{Q}}(S) = \mathbb{R}$. However, this may not be the case: one could potentially construct an infinite sequence $x_1, x_2, \ldots$ such that the $\mathbb{Q}$-span of $S_\infty := S \cup \{x_1, x_2, \ldots\}$ is still not $\mathbb{R}$! However, in this case one just continues as before: pick some y_1 outside the span of S_∞ and add it to S_∞, etc... This can get a little confusing, but intuitively, if one can keep doing this process, even after you get to infinity, or even an infinity's worth of infinities, eventually you should end up with some T that's a $\mathbb{Q}$-basis. If you're interested, there is an important tool from logic called *Zorn's Lemma* that tells you rigorously that you can do this. Look it up! □

EXERCISE 1.14.19 (For those who know about countability). Show that any $\mathbb{Q}$-basis for $\mathbb{R}$ is uncountably infinite.

Now let's show how to use a $\mathbb{Q}$-basis S to create additive functions. Given $x \in \mathbb{R}$, we know we can write x uniquely as a $\mathbb{Q}$-linear combination of elements of S. For convenience, let's write this as

$$x = \sum_{s_i \in S} q_i s_i.$$

Now, it may look like there are infinitely many terms in this summatiy all but finitely many of the coefficients q_i are zero. So, this is really just a $\mathbb{Q}$-linear combination as before, but written slightly differently. Define $\langle x, s_i \rangle := q_i$, so that

$$x = \sum_{s_i \in S} \langle x, s_i \rangle s_i.$$

That is, $\langle x, s_i \rangle$ is the coefficient of s_i in the unique $\mathbb{Q}$-linear combination of elements of S that gives you x. We often call $\langle x, s \rangle$ the *s-coefficient of x* or the *s-coordinate of x*. For instance, if S is a $\mathbb{Q}$-basis that contains $1, \sqrt{2}, \sqrt{3}$, then

$$\langle 5 + 8\sqrt{2}, 1 \rangle = 5, \quad \langle 5 + 8\sqrt{2}, \sqrt{2} \rangle = 8, \quad \langle 5 + 8\sqrt{3}, \sqrt{3} \rangle = 0.$$

EXERCISE 1.14.20. Explain why it is that if S is a $\mathbb{Q}$-basis and $s, t \in S$, then $\langle s, t \rangle = 0$ unless $s = t$, in which case $\langle s, t \rangle = 1$.

FACT 1.14.21. *If $S \subset \mathbb{R}$ is a $\mathbb{Q}$-basis and $s \in S$, the following function is additive:*

$$f : \mathbb{R} \longrightarrow \mathbb{R}, \quad f(x) = \langle x, s \rangle$$

PROOF. Given $x, y \in \mathbb{R}$, we have

$$x + y = \sum_{s_i \in S} \langle x, s_i \rangle s_i + \sum_{s_i \in S} \langle y, s_i \rangle s_i = \sum_{s_i \in S} (\langle x, s_i \rangle + \langle y, s_i \rangle) s_i,$$

so by definition, $\langle x + y, s_i \rangle = \langle x, s_i \rangle + \langle y, s_i \rangle$. □

We can now prove Theorem 1.14.13, that for any $\mathbb{Q}$-independent set S, we can prescribe the values of an additive function arbitrarily on elements of S.

PROOF OF THEOREM 1.14.13. Say $S \subset \mathbb{R}$ is $\mathbb{Q}$-independent, and that we are given real numbers a_s for every element $s \in S$. Let T be a $\mathbb{Q}$-basis containing S, as given by Theorem 1.14.18. Given $s \in S \subset T$, let $\langle x, s\rangle$ denote the s-coefficient of x with respect to the $\mathbb{Q}$-basis T and define

$$f : \mathbb{R} \longrightarrow \mathbb{R}, \quad f(x) = \sum_{s \in S} a_s \langle x, s\rangle.$$

By Exercise 1.14.20, $f(s) = a_s$ for all s. Additivity of f follows from Fact 1.14.21: we leave the details as an exercise. □

EXERCISE 1.14.22 (For students with an analysis background). Show that if $f : \mathbb{R} \longrightarrow \mathbb{R}$ is additive and continuous, then $f(x) = ax$ for some $a \in \mathbb{R}$.

A function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is called *multiplicative* if

$$f(xy) = f(x)f(y), \quad \forall x, y \in \mathbb{R}.$$

Examples include functions $f(x) = x^a$, where $a \in \mathbb{R}$.

EXERCISE 1.14.23. Suppose that $f : \mathbb{R} \longrightarrow \mathbb{R}$ is additive, and define

$$g : \mathbb{R} \longrightarrow \mathbb{R}, \; g(x) = \begin{cases} e^{f(\log |x|)} & x \neq 0 \\ 0 & x = 0 \end{cases}.$$

Show that g is multiplicative.

EXERCISE 1.14.24. Suppose that $f : \mathbb{R} \longrightarrow \mathbb{R}$ is both additive and multiplicative, and that f *is not the zero function.* In this problem, we will show that the only other option is the identity function, i.e. that $f(x) = x$ for all $x \in \mathbb{R}$.

(a) Show that $f(q) = q$ for all $q \in \mathbb{Q}$.
(b) Show that if $x \geqslant 0$, then $f(x) \geqslant 0$ too.
(c) Using (b), show that f is *order preserving*, i.e. that $x \leqslant y$ implies $f(x) \leqslant f(y)$.
(d) Show that $f(x) = x$ for all $x \in \mathbb{R}$. *Hint: you may find it useful that between every two distinct real numbers, there's a rational number.*

EXERCISE 1.14.25 (For students with an analysis background). Suppose that x is irrational. Show that the set of all numbers of the form $m + nx$, where $m, n \in \mathbb{Z}$, is *dense* in $\mathbb{R}$, meaning that between any two real numbers there is a number of this form. *This set should be considered as the* $\mathbb{Z}$-span *of* $\{1, x\}$. *It is obvious that the* $\mathbb{Q}$*-span is dense, since it contains* $\mathbb{Q}$*, which is dense.*

EXERCISE 1.14.26. Suppose $S \subset \mathbb{R}$ is a Hamel basis and that $a \in \mathbb{R}$, where $a \neq 1$. Show that there is some $x \in S$ with $ax \notin S$. *Hint: if* $ax \in S$ *for all* $x \in S$*, the sum* $\sum_i q_i$ *of the coefficients in* (4) *will be the same for* x *as it is for* ax*. (Why?)*

In Exercise 1.14.26, if a is rational, it can *never* be the case that $ax \in S$. For if so, we'd have $a \cdot x - 1 \cdot (ax) = 0$, implying that $0 \in \mathbb{R}$ can be expressed as a $\mathbb{Q}$-linear combination of elements of S in more than one way: as above, and with the trivial linear combination, in which all coefficients are zero. However, in the problem a doesn't have to be rational, and the question is really different.

EXERCISE 1.14.27. Show there is no subset $S \subset \mathbb{R}$ such that every $x \in \mathbb{R}$ can be represented uniquely as a '$\mathbb{Z}$-linear combination' of elements of S:

$$x = q_1 s_1 + \cdots + q_k s_k, \qquad q_1, \ldots, q_k \in \mathbb{Z},\ s_1, \ldots, s_k \in S. \tag{4}$$

Hint: assume there is, and take some $x \in S$. Write $x/2$ as a $\mathbb{Z}$-linear combination of elements of S.

EXERCISE 1.14.28. Find two disjoint subsets A, B such that $A \cup B = \{x \in \mathbb{R} \mid x > 0\}$, and where both A, B are *closed under addition*, meaning

$$a + a' \in A, \text{ for all } a, a' \in A, \text{ and } b + b' \in B \text{ for all } b, b' \in A.$$

EXERCISE 1.14.29. A function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is *periodic* with period ℓ if $f(x + \ell) = f(x)$ for all $x \in \mathbb{R}$. For example, sin and cos are periodic with period 2π.

(a) Suppose that f, g are periodic with periods a, b, respectively. If b is a rational multiple of a, show that $f + g$ is periodic.

(b) Using a Hamel basis, show that there are periodic functions $f, g : \mathbb{R} \longrightarrow \mathbb{R}$ (with different periods) such that $f + g = id$, that is

$$f(x) + g(x) = x, \ \forall x \in \mathbb{R}.$$

Hint: given $x \in \mathbb{R}$ and $s \in S$, let $\langle x, s \rangle$ be the 'coefficient' of s in $\mathbb{Q}$-linear combination expressing x. That is, $x = \sum_{s \in S} \langle x, s \rangle s$. Pick some s, and let

$$f(x) = \langle x, s \rangle s.$$

(c) (Harder) Can you write $x \mapsto x^2$ as a sum of three periodic functions?

(d) Show that $x \mapsto e^x$ is not a sum of (any finite number of) period functions.

EXERCISE 1.14.30. The dihedral angles in a regular octahedron are $\pi - \cos^{-1}(1/3)$. Show that a regular octahedron O is neither scissors congruent to a cube C, nor to a regular tetrahedron T.

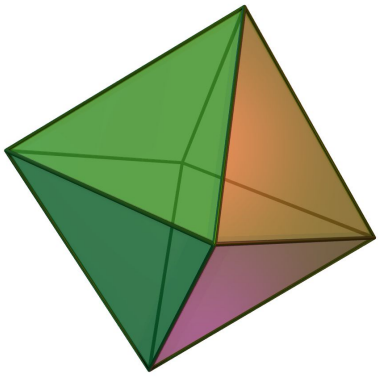


FIGURE 3. A regular octahedron, courtesy of Wikipedia.

CHAPTER 2

Spherical Geometry

2.1. Distance on the sphere

We have talked a lot about distance so far, and about realizing distances between points in $\mathbb{R}^2$ by shortest paths, which are line segments. This work is physically meaningful because $\mathbb{R}^2$ is a good model for the geometry of the earth, at least when distances are small.

As a curious example, which US city do you think is closest to Dakhla, one of the closest cities to the US in Africa? Looking at a map, it seems like a bit of a tossup between all the cities on the eastern seaboard.

Computing the actual distances, though, one finds that the distance from Miami to Dakhla is 3,991 miles, while the distance from Boston to Dakhla is 3,374 miles, which is considerably smaller! Even more strikingly, the distance from Iceland to Dakhla is around 2800 miles, even though on the map it looks somewhat comparable.

This discussion motivates the investigation of a new kind of geometry, *spherical geometry*, which more closely models the geometry of the earth when large distances are concerned. Specifically, our model will be the radius r sphere

$$S_r = \{x \in \mathbb{R}^3 \mid |x| = r\} \subset \mathbb{R}^3,$$

and our goal is to understand the geometry of S_r in a manner similar to our investigation of the geometry of $\mathbb{R}^2$.

How does one define distance on S_r? Just using the usual definition of distance in $\mathbb{R}^3$ doesn't seem like a good idea, since it can only be realized using paths that go

through the interior of the earth. However, we know how to measure the lengths of paths on S_r, and we can just define the distance between two points on S_r as the shortest length of a path between them: if $p, q \in S_r$,

$$d_{S_r}(p,q) = \inf\big\{ \text{length}(\gamma) \;\big|\; \gamma : [a,b] \longrightarrow S_r, \gamma(a) = p, \gamma(b) = q \big\}.$$

We write 'inf' here so as to avoid issues about whether length minimizing paths actually exist. However, we'll see in a minute that they do exist, so one could write 'min' instead if desired.

EXERCISE 2.1.1. Suppose that $A \subset \mathbb{R}^n$ and define, for $p, q \in A$,

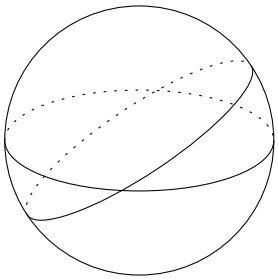
$$d_A(p,q) = \inf\big\{ \text{length}(\gamma) \;\big|\; \gamma : [a,b] \longrightarrow A, \gamma(a) = p, \gamma(b) = q \big\}.$$

Show that the function d_A satisfies all the usual properties of a metric, except that $d_A(p,q)$ may equal ∞ if there is no path from p to q in A that has finite length.

Examples of $A \subset \mathbb{R}^n$ where d_A can be infinite are 'disconnected sets' like the union of two points, or the images of non-rectifiable paths like the Koch curve.

Of course, this is not a problem for S_r. We'll find paths of finite length that join two given points while answering another important question: what are the shortest paths between points on S_r, so the spherical analogue of lines in $\mathbb{R}^2$?

DEFINITION 2.1.2. A *great circle* on the sphere S_r is an intersection $P \cap S_r$, where P is a plane in $\mathbb{R}^3$ going through the origin.



If p and q are points on S_r, then any plane P going through $0, p, q$ intersects S_r in a great circle that passes through p and q. There is a unique such plane except when $0, p, q$ are colinear, in which case we say that p and q are *antipodes*. In other words,

PROPOSITION 2.1.3. *Every pair of points p, q on S_r lies on a great circle. The great circle is unique except if p and q are antipodes, in which case there are infinitely many great circles passing through p and q.*

Regarding two points $p, q \in S_r$ as vectors based at the origin, suppose p, q meet at an angle $\theta \in [0, \pi]$. Pick a plane P through the origin containing p, q and let v be the unit vector in P that is perpendicular to p and closest to q. Then the path

$$\gamma : [0, 2\pi] \longrightarrow \mathbb{R}^3, \;\; \gamma(t) = r\cos(t)p + r\sin(t)v$$

parameterizes $P \cap S_r$, and $\gamma(0) = p$ while $\gamma(\theta) = q$. Since

$$\operatorname{length} \gamma|_{[0,\theta]} = \int_0^\theta |\gamma'(t)|\, dt = \int_0^\theta r\, dt = r\theta,$$

we see that *when the angle between $p, q \in S_r$ is θ, there is a segment of a great circle cconnecting p, q that has length $r\theta$.*

It turns out that this is actually the shortest path between p and q!

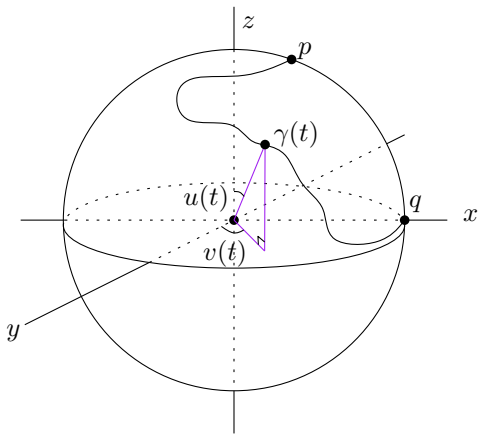
PROPOSITION 2.1.4. *If p, q are points on S_r with angle θ, any path γ from p to q has length at least $r\theta$, with equality only if γ is an arc of the great circle containing p, q. In particular, the spherical distance between $p, q \in S_r$ is $d_{S_r}(p, q) = r\theta$.*

In the proof, we will assume γ is piecewise differentiable, so that we can calculate its length via an integral formula. In general, a continuous path γ on S_r can be approximated by a piecewise differentiable path with almost the same length; for instance, one can project a piecewise linear approximation to γ radially onto the sphere. So, any continuous path with length less than θ would yield a piecewise differentiable path with length less than θ. With a bit more work, one could also prove the statement about equality for continuous paths, but we'll not do so here.

PROOF. It suffices to prove the proposition when γ is contained in an open hemisphere around p, and in particular $\theta < \pi/2$. If not, cut γ into small pieces, and observe that if the length of γ is less than the angle between its endpoints, the same must be true for one of these subpaths. A similar argument reduces the 'with equality' part of the proposition to this case.

Rotations around lines through the origin are isometries of $\mathbb{R}^3$ and preserve S_r. They preserve the lengths of paths and the angles between vectors, and take great circles to great circles. So using a composition of two rotations, we may assume

$$p = \left(r \sin\left(\frac{\pi}{2} - \theta\right), 0, r\cos\left(\frac{\pi}{2} - \theta\right)\right), \quad q = (r, 0, 0).$$



Using spherical coordinates, γ can be written as

$$\gamma(t) = \big(r\sin u(t)\cos v(t), r\sin u(t)\sin v(t), r\cos u(t)\big),$$

where $u : [a,b] \longrightarrow [0,\pi]$ and $v : [a,b] \longrightarrow [0,2\pi)$. Note that as

$$\gamma(a) = \left(r\sin\left(\frac{\pi}{2}-\theta\right), 0, r\cos\left(\frac{\pi}{2}-\theta\right)\right),$$

and $\gamma(b) = (r,0,0)$, we have

$$u(a) = \frac{\pi}{2}-\theta,\ v(a) = \frac{\pi}{2}-\theta, \quad u(b) = \frac{\pi}{2}.$$

The square of the speed of γ at time t is then:

$$\begin{aligned}
|\gamma'(t)|^2 &= \big|(ru'\cos u\cos v - rv'\sin u\sin v, ru'\cos u\sin v + rv'\sin u\cos v, -ru'\sin u)\big|^2 \\
&= (ru'\cos u\cos v - rv'\sin u\sin v)^2 + (ru'\cos u\sin v + rv'\sin u\cos v)^2 + \\
&\quad (-ru'\sin u)^2 \\
&= (ru'\cos u\cos v)^2 + (rv'\sin u\sin v)^2 + (ru'\cos u\sin v)^2 + \\
&\quad (rv'\sin u\cos v)^2 + (-ru'\sin u)^2, \qquad \textit{as the middle terms above cancel,} \\
&= (ru'\cos u)^2 + (rv'\sin u)^2 + (-ru'\sin u)^2, \qquad \textit{using } \sin^2+\cos^2 = 1, \\
&= (ru')^2 + (rv'\sin u)^2,
\end{aligned}$$

so putting this into the integral that we use to calculate length,

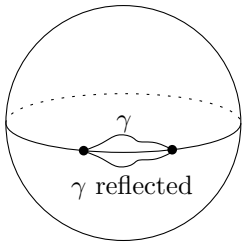
$$\begin{aligned}
\text{length}(\gamma) &= \int_a^b \sqrt{(ru'(t))^2 + (rv'(t)\sin u(t))^2}\,dt \\
&\geqslant \int_a^b |ru'(t)|\,dt \\
&\geqslant \int_a^b ru'(t)\,dt \\
&= r\frac{\pi}{2} - r\left(\frac{\pi}{2}-\theta\right) \\
&= r\theta.
\end{aligned}$$

Moreover, we have equality above if and only if $v'(t) = 0$ and $u'(t) > 0$ for all t. If this happens, $v(t) = \pi/2$ for all t, so γ lies along the great circle determined by the xz-plane, and the condition $u'(t) > 0$ means that it always goes clockwise, so it traces out exactly the segment of the great circle between p and q. □

While the proof of Proposition 2.1.4 above used some slightly nasty calculus, here's a result with a simple proo that should convince you intuitively that great circles are the correct paths to consider here.

FACT 2.1.5. *Suppose that $p, q \in S_r$ and that there is a* unique *shortest path γ from p to q. Then γ is a segment of a great circle.*

PROOF. Let P be a plane in $\mathbb{R}^3$ passing through the origin and p, q'. The reflection through P sends γ to a path on S_r with the same endpoints that has the same length as γ, which must then be γ, by our assumption. This means that the reflection fixes γ, so γ' lies along the great circle $P \cap S$. □



EXERCISE 2.1.6. (Quite Hard) Using just that $d_{S_r}(p, q) = r\theta$, where θ is the angle between p, q, taken within the interval $[0, \pi]$, prove that d_{S_r} is a metric on S_r. *Hint: the hard part here is proving a triangle inequality for the angles between vectors u, v, w in $\mathbb{R}^3$, i.e. that $\theta(u, v) + \theta(v, w) \geqslant \theta(u, w)$. Try doing it first if u, v, w all lie in a plane, then in general, try projecting w onto the plane spanned by u, v.*

EXERCISE 2.1.7. Show that if five points are placed on S_r, there is a closed hemisphere containing four of them. Show by example that this is not true for open hemispheres. *Here,* closed *means that the hemisphere includes its great circle boundary, while* open *hemispheres don't contain the boundary.*

If a point $p \in S_r$ is given, the *circle* of radius s around p is the set

$$C(p, s) = \{q \in S_r \mid d_{S_r}(p, q) = s\}.$$

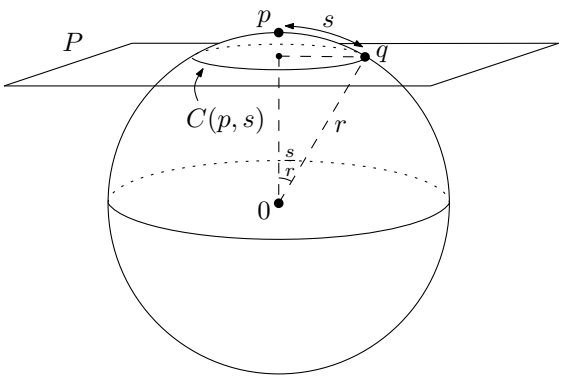
Recall that the dot product $p \cdot q$ of two points in $\mathbb{R}^3$ can be interpreted geometrically as $p \cdot q = |p| \cdot |q| \cos\theta$, where θ is the angle between p, q. As $d_{S_r}(p, q) = r\theta$,

$$d_{S_r}(p, q) = s \iff \theta = \frac{s}{r} \iff p \cdot q = |p||q| \cos\left(\frac{s}{r}\right),$$

so since $|p| = |q| = r$ for $p, q \in S_r$, the circle $C(p, s)$ is just the intersection with S_r of the plane $P \subset \mathbb{R}^3$ defined by the equation

$$P = \{q \in \mathbb{R}^3 \mid p \cdot q = r^2 \cos(s/r)\}.$$

Using elementary Euclidean geometry, we see that in fact $C(p, s)$ is a Euclidean circle of *Euclidean* radius $r \sin\frac{s}{r}$ around the point $(r \cos\frac{s}{r})p \in P$.



Hence, the circumference of $C(p, s)$ is $2\pi r \sin \frac{s}{r}$. So, we have proved:

COROLLARY 2.1.8. *The circumference of a circle of radius s on S_r is $2\pi r \sin \frac{s}{r}$.*

Here are some exercises.

EXERCISE 2.1.9. Suppose that P is a plane in $\mathbb{R}^3$ defined by $ax + by + cz = d$. Show algebraically that the intersection $P \cap S_r$ is either empty, a point, or a circle.

EXERCISE 2.1.10. Show that a circle $C(p, \alpha)$ is a great circle if and only if $\alpha = \frac{r\pi}{2}$.

EXERCISE 2.1.11. The circumference of a Euclidean circle of radius s is $2\pi s$. Show

$$\lim_{r\to\infty} 2\pi r \sin \frac{s}{r} = 2\pi s.$$

This should make sense, since at moderate scales a very large sphere (like the earth) is almost indistinguishable from a plane.

EXERCISE 2.1.12 (Compare with 8.11). Show that for fixed r, we have

$$\lim_{s\to 0} \frac{2\pi r \sin \frac{s}{r}}{2\pi s} = 1.$$

Explain what this means about small spherical circles, and their relationship to Euclidean circles.

EXERCISE 2.1.13. The following is a quotation from the Bible, 1 Kings 7.23:

Then he made the molten sea; it was round, ten cubits from brim to brim, and five cubits high. A line of thirty cubits would encircle it completely.

Let's interpret this as describing an above-ground pool ten cubits in diameter and 30 cubits in circumference. Some doubters like to say that this is impossible, since the ratio of circumference to diameter should be π, not 3. In a scathing rebuttal, explain how the pool could be built exactly as specified on the surface of S_r, for some r. *Hint: you don't need to explicitly find the appropriate r. Just argue that it exists using the intermediate value theorem.*

2.2. Spherical isometries

The distance function d_{S_r} is a metric on S_r, giving S_r the structure of a metric space. What are the isometries of S_r? It turns out there is a dictionary between isometries of S_r and certain isometries of $\mathbb{R}^3$.

PROPOSITION 2.2.1. *Let $r > 0$. Any isometry $f : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ with $f(0) = 0$ restricts to an isometry of S_r, considered with the metric d_{S_r}. Conversely, every isometry $f : S_r \longrightarrow S_r$ is the restriction of a unique isometry of $\mathbb{R}^3$ fixing the origin.*

PROOF. If $p \in \mathbb{R}^3$ lies in S_r, then

$$d(0, f(p)) = d(f(0), f(p)) = d(0, p) = r,$$

so $f(p) \in S_r$ as well. Therefore, f restricts to a map of S_r. Since the same is true for f^{-1}, the restriction map $f : S_r \longrightarrow S_r$ is a bijection.

We now show that the restriction of f to S_r preserves d_{S_r}. If $p, q \in S_r$, then $d_{S_r}(p, q) = r\theta$, where θ is the angle between the segments $0p$ and $0q$. By Lemma 1.2.12, as f is an isometry the angle between $0f(p)$ and $0f(q)$ is θ as well. So,

$$d_{S_r}(p, q) = r\theta = d_{S_r}(f(p), f(q)),$$

implying that f restricts to a d_{S_r}-isometry.

Conversely, suppose we start with an isometry $f : S_r \longrightarrow S_r$. Define

$$F : \mathbb{R}^3 \longrightarrow \mathbb{R}^3, \quad F(x) = \frac{|x|}{r} f\left(\frac{rx}{|x|}\right).$$

In other words, to define $F(x)$ we first scale the vector x so that its head lies on S_r, then apply f, then scale the result back to its original length.

EXERCISE 2.2.2. Prove that $d(F(x), F(y)) = d(x, y)$ for all $x, y \in \mathbb{R}^3$.

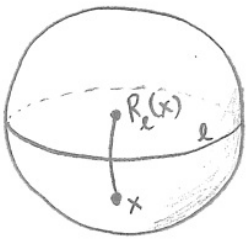
It's easy to see that F is a bijection - its inverse is obtained from the isometry f^{-1} of S_r in the same way that we obtained F from f. So, F is an isometry. □

By Theorem 1.3.1, the isometries of $\mathbb{R}^3$ that fix the origin are the identity, reflections through planes containing the origin, rotations around lines through the origin, and 'twist reflections' that are compositions of rotations around lines through the origin and perpendicular planes containing the origin. So, this gives a complete classification of isometries of S_r. However, the reliance on $\mathbb{R}^3$ here is a bit unsatisfying, and we'll see that in fact the isometries of S_r can be described *intrinsically,* in a way that parallels the definitions of isometries of $\mathbb{R}^2$.

Let's take as an example the restriction to S_r of a reflection through a plane $P \subset \mathbb{R}^3$. The plane P intersects S_r in a great circle, which we call ℓ, remembering that great circles play the role of lines on S_r. Denoting by

$$R_\ell : S_r \longrightarrow S_r$$

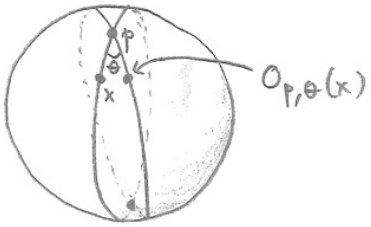
the restricted reflection, we see that $R_\ell(p) = p$ whenever $p \in \ell$, while if $q \notin \ell$, then any segment of a great circle connecting q and $R_\ell(q)$ is perpendicularly bisected by ℓ. These properties determine R_ℓ, and are completely analogous to those in the geometric definition of a reflection in $\mathbb{R}^2$. So, by analogy, we call R_ℓ the *reflection of S_r through the great circle ℓ.*



Similarly, any line l through the origin in $\mathbb{R}^3$ intersects S_r in a point p and its antipode. The rotation around l by angle θ restricts to a map

$$O_{p,\theta} : S_r \longrightarrow S_r$$

that fixes p and its antipode, and otherwise takes x to the unique point $O_{p,\theta}(x)$ such that any great circle segments px and $pO_{p,\theta}(x)$ have the same length and meet with an angle of θ, measured counterclockwise from px to $pO_{p,\theta}(x)$.



We call this map $O_{p,\theta}$ the *rotation of S_r around p.* In some sense, we should call $O_{p,\theta}$ a 'rotation around both p and its antipode', but this is too much of a mouthful.

We now have spherical versions of rotations and reflections of $\mathbb{R}^2$. What about translations? In fact, a spherical rotation plays a dual role, analogous to both a Euclidean rotation and a translation! To see why, note that a translation of $\mathbb{R}^2$ preserves all the lines in the direction of translation, and acts on each one as a shift. Well, a spherical rotation $O_{p,\theta}$ preserves all the circles $C(p, s)$, where $s \in [0, \pi]$, but the circle $C(p, \pi/2)$ is a great circle, and therefore plays the role of a line on S_r! We can also consider $O_{p,\theta}$ as 'shifting' along $C(p, \pi/2)$, except that since $C(p, \pi/2)$ closes up we eventually get back to where we started. So, a spherical rotation $O_{p,\theta}$ could also be considered as a 'spherical translation' along the great circle $C(p, \pi/2)$.

Moreover, if rotations can be considered as 'spherical translations', then twist reflections are 'spherical glide reflections'. For if $p \in S_r$, the line through $0, p$ is perpendicular to the plane cutting out $\ell = C(p, \pi/2)$, so t(W)ist reflections are compositions

$$W_{p,\theta} : S_r \longrightarrow S_r, \quad W_{p,\theta} = O_{p,\theta} \circ R_\ell$$

of 'spherical translations' along ℓ and reflections through ℓ, in perfect analogy with glide reflections in $\mathbb{R}^2$. We'll refer to $W_{p,\theta}$ above as the *twist reflection of S_r along ℓ by angle θ*.

With this new terminology, the classification of spherical isometries becomes:

THEOREM 2.2.3. *The only isometries of S_r are the identity, rotations around points, and reflections and twist reflections along great circles.*

Here are some exercises on spherical isometries.

EXERCISE 2.2.4. Show that every isometry of S_r is the product of at most three reflections.

EXERCISE 2.2.5 (The antipodal map). The map $A : S_r \longrightarrow S_r$, $A(p) = -p$ is called the *antipodal map*, since it takes every point to its antipode.

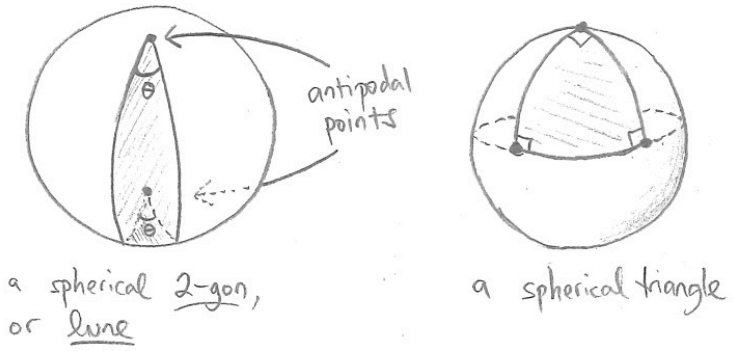
(a) Show that A is an isometry.
(b) Where does the antipodal map appear in Theorem 2.2.3?
(c) Without using Theorem 2.2.3, show that if $f : S_r \longrightarrow S_r$ is an isometry and $p, q \in S_r$ are antipodes, then $f(p), f(q)$ are antipodes as well.
(d) Show that $f \circ A = A \circ f$ for every isometry $f : S_r \longrightarrow S_r$. We summarize this property by saying that A is *central*.
(e) Using Theorem 2.2.3, show that A and the identity are the only central isometries of S_r.

EXERCISE 2.2.6. A metric space X is *homogenous* if for all $x, y \in X$, there is an isometry $f : X \longrightarrow X$ with $f(x) = y$. Show that $\mathbb{R}^n$ and S_r are homogenous. Show that $\{1, 2, 3\}$, with the metric $d(x, y) = |x - y|$, is a non-homogenous metric space.

2.3. Spherical area and polygons

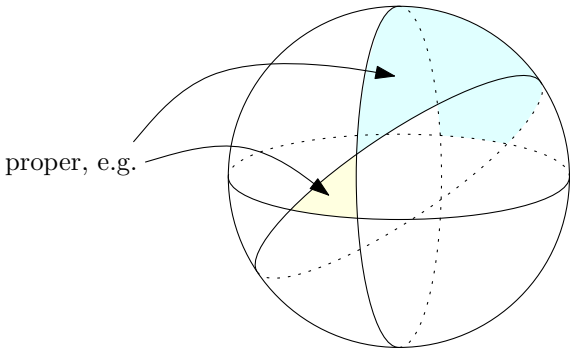
Now that we have a little familiarity with distance on the sphere S_r, what about area? Surface area is sometimes covered in a good multivariable calculus class, and if you're comfortable with it you might at least remember that the surface area of the sphere is supposed to be $4\pi r^2$. This, combined with believable facts like *congruent subsets of S_r have equal area* and *the area of a union of two regions of S_r with disjoint interiors is the sum of the two component areas* are all you will need here.

In search of some examples where we can calculate area, we are led to consider spherical polygons. As great circles on the sphere play the role of lines in the plane, it is natural to define a *spherical polygon* as a region on the sphere bounded by a finite number of segments of great circles that form a loop.



As pictured above, there are two sided spherical polygons, called *lunes* or *bigons*! Even worse/better, a hemisphere of S_r can be considered as a polygon with one side, or *monogon*; we usually plant a vertex somewhere on the great circle boundary so that we still have the picture of edges connecting vertices, though.

A spherical polygon is *proper* if it is contained in an open hemisphere. For a triangle T, properness has another interpretation. The three great circles determined by T's sides divide the sphere into eight triangles, as in the picture below, and T is either one of these six or is a union of some of them.



If T is one of the eight, then T is proper: a great circle bounding an open hemisphere containing T can be created by slightly rotating any of the three great circles pictured. On the other hand, if T is a union of more than one of these triangles, it will contain a pair of antipodal points, so cannot be contained in an open hemisphere.

EXERCISE 2.3.1. Show that a spherical triangle T is proper if and only if all its interior angles are less than π.

EXERCISE 2.3.2. Show that a proper triangle on S_r has side lengths less than πr. *Note: the converse is not true, as the complement of a proper triangle is a nonproper triangle with the same side lengths.*

We now give a first interesting example where we can calculate area. Intuitively, a lune with angle θ takes up a proportion of $\frac{\theta}{2\pi}$ of the sphere, so its area should be

$$\frac{\theta}{2\pi} 4\pi r^2 = 2\theta r^2.$$

So, we can then work in stages. All lunes of a certain angle are congruent, i.e. there's an isometry of S_r taking one to the other, so they all have the same area. As the sphere S_r is the union of $2n$ lunes with angle π/n and disjoint interiors, a lune with angle $\theta = \frac{\pi}{n}$ must have area $4\pi r^2/2n = 2(\pi/n)r^2$. Taking the union of m such lunes, a lune with angle $\theta = \frac{\pi m}{n}$ has area $2(\pi m/n)r^2$. This proves the formula when θ is rational. In general, given any angle θ, we can write θ/π using decimal notation

$$\theta/\pi = a_0 + \frac{a_1}{10} + \frac{a_2}{100} + \cdots, \quad \text{where } a_i \in \mathbb{Z},$$

and then $\theta = \pi a_0 + \pi\frac{a_1}{10} + \cdots$ is a sum of rational multiples of π, so a lune with angle θ is the union of lunes with angles that are rational multiples of π and disjoint interiors. Summing their areas gives $2\theta r^2$.

Returning to our discussion of polygons, note that in the picture above on the right we have a triangle T with three right angles that occupies a quarter of the upper hemisphere. However, the clever reader will notice that is actually another triangle pictured: the complement of T's interior, which has three $3\pi/2$ angles!

Exercise 2.3.3. Given an angle $\alpha \in (0, 2\pi)$, show that there is a spherical triangle that has α as one of its interior angles.

There is some debate whether to even include triangles like the complement of T's interior in a definition of 'polygon', since although they are bounded by a number of great circle segments, they are too large and curved to really look like a polygon.

We now come to the central result of this section.

Girard's Theorem. *If a proper triangle T on S_r has interior angles α, β, γ, then*

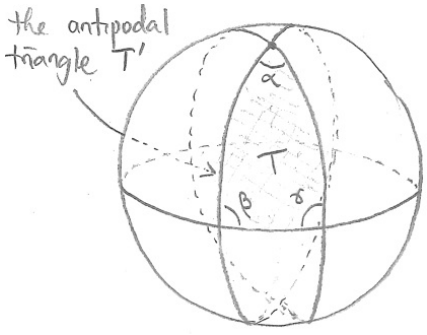
$$\text{Area}(T) = r^2(\alpha + \beta + \gamma - \pi).$$

In particular, this implies that the angle sum of a spherical triangle is always greater than π, since area must be positive. The quantity $\alpha + \beta + \gamma - \pi$ is often called the *angle excess* of the triangle, since it is the amount by which the interior angle sum exceeds the corresponding sum for Euclidean triangles.

Proof. Suppose T is a triangle on S_r with interior angles α, β, γ. Extend the sides of T to the full great circles on which they lie. There are then two cases, depending on whether T is proper or not.

First, assume T is proper, as pictured below.

We see six lunes that cover S_r, two each with angles α, β, and γ. Every point in the sphere is contained in exactly one of the lunes, with the exception that points in



T and in the antipodal triangle T' are contained in 3 lunes. So,

$$\begin{aligned} 4\pi r^2 &= \text{Area}(S) \\ &= 2 \cdot (2\alpha r^2) + 2 \cdot (2\beta r^2) + 2 \cdot (2\gamma r^2) - 2\,\text{Area}(T) - 2\,\text{Area}(T') \\ &= 4r^2\,(\alpha + \beta + \gamma) - 4\,\text{Area}(T), \end{aligned}$$

where the second equality decomposes the area of S into the regions covered by the lunes, while subtracting off twice the areas of T and T' to compensate for the overcounting. Solving for $\text{Area}(T)$ proves the proposition. □

EXERCISE 2.3.4. Show that in fact, the area formula in Girard's Theorem is also true for non-proper triangles.

EXERCISE 2.3.5. Using Girard's theorem, show that if r is large, then the interior angle sum of a small triangle (say, with unit area) on S_r is close to π. This is another example of the philosophy that at small scales, a large sphere looks Euclidean.

EXERCISE 2.3.6 (SAS?). Two spherical polygons are *congruent* if there is an isometry of S_r taking one to the other. Explain why it is that if two triangles on S_r share two side lengths that meet at the same interior angle, then the triangles are congruent.

If you're interested, think about spherical analogues of the other congruence conditions for Euclidean triangles. In fact, there's also an AAA condition in spherical geometry! This should be plausible, since Girard's Theorem implies that you can't scale a triangle without altering its angles like you can in Euclidean space.

EXERCISE 2.3.7. (Hard) Prove the 'spherical law of cosines': if a proper triangle on S_r has side lengths a, b, c, and θ is the angle opposite c, then

$$\cos\frac{c}{r} = \cos\frac{a}{r}\cos\frac{b}{r} + \sin\frac{a}{r}\sin\frac{b}{r}\cos\theta.$$

2.4. Projections

Now that we're somewhat comfortable with spherical geometry, how do we relate it back to Euclidean geometry? A cartographer can tell you the answer! There are

many ways to create maps of a sphere. We'll describe three of the most important, the *cylinder projections*, *stereographic projections* and *gnomonic projections*.

Throughout this section, we'll focus on the sphere S with unit radius. Everything works in the same way for spheres with arbitrary radius, but the formulas are nastier.

2.4.1. Cylindrical projection. Enclose S in the vertical cylinder

$$C = \{(x, y, z) \mid x^2 + y^2 = 1\}$$

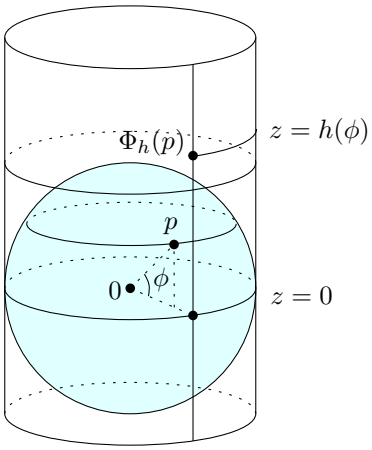
tangent to S along the equator and fix a differentiable, increasing function

$$h : \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \longrightarrow \mathbb{R}.$$

The *cylindrical projection associated to* h is the function

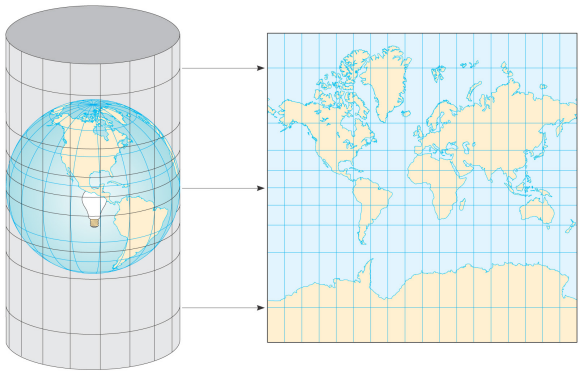
$$\Phi_h : S \backslash \{(0, 0, 1),\ (0, 0, -1)\} \longrightarrow C$$

defined as follows. If $p \in S$, let ϕ be the angle that the line segment $0p$ makes with the horizontall, and $\Phi_h(p)$ be the point with height $h(\phi)$ that lies on the intersection of C and the vertical plane through 0 and p, as shown in the picture below.



Note that by cutting the cylinder along a vertical line and unrolling, we can create a flat map of the sphere from a cylindrical projection.

If $h(\phi) = \tan(\phi)$, then $\Phi_h(p)$ is the point at which the ray from 0 in the direction of p intersects C. So, imagining a light bulb at 0 and a translucent sphere, Φ_h is the 'radial projection' in which the light paints the sphere onto the cylinder.



In the picture above, it is clear that distances on the sphere become quite distorted upon cylindrical projection. In the map, Greenland looks three times as big as Australia, but in actuality it is the other way around. Moreover, distances may be scaled differently in different directions. To understand this, let's compare the distance distance distortion along *meridians* (or *longitudes*) on the sphere, great circles through the North Pole, with that along *parallels* (or *lattitudes*), spherical circles centered at the North Pole.

First, consider a parallel ℓ, say consisting of all point p such that $0p$ makes an angle of ϕ with the horizontal. Then ℓ is a Euclidean circle in $\mathbb{R}^3$ with radius $\cos(\theta)$, and hence its length is $2\pi\cos(\theta)$. The image $\Phi_h(\ell)$ is a Euclidean circle of radius 1 that lies at height $h(\phi)$ on the cylinder C. So, Φ_h stretches ℓ by a factor of $1/\cos(\theta) = \sec(\theta)$. Furthermore, since cylindrical projections have a rotational symmetry, all parts of ℓ are uniformly stretched by this constant factor.

Now consider a meridian m, say the one that is the intersection of S with the xz-plane. Here, different parts of m are stretched by different amounts, so to figure out how much m is stretched at a given point, we need to parametrize m and take derivatives. Write

$$m : (-\pi, \pi) \longrightarrow S, \quad m(\phi) = (\cos\phi, 0, \sin\phi).$$

Then $|m'(\phi)| = 1$ for all ϕ, i.e. the path is unit speed on S. Moreover,

$$\Phi_h \circ m(\phi) = (1, 0, h(\phi)),$$

so we have $|(\Phi_h \circ m)'(\phi)| = |(0, 0, h'(\phi))| = h'(\phi)$, and hence

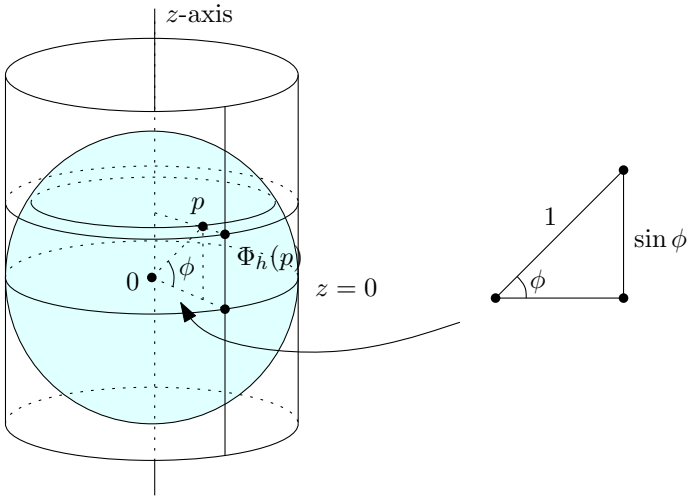
$$\frac{|(\Phi_h \circ m)'(\phi)|}{|m'(\phi)|} = h'(\phi).$$

In summary: at a point p such that $0p$ makes an angle of ϕ with the horizontal, Φ_h *distorts distances by a factor of* $\sec\phi$ *in the parallel direction and by* $h'(\phi)$ *in the meridian direction.*

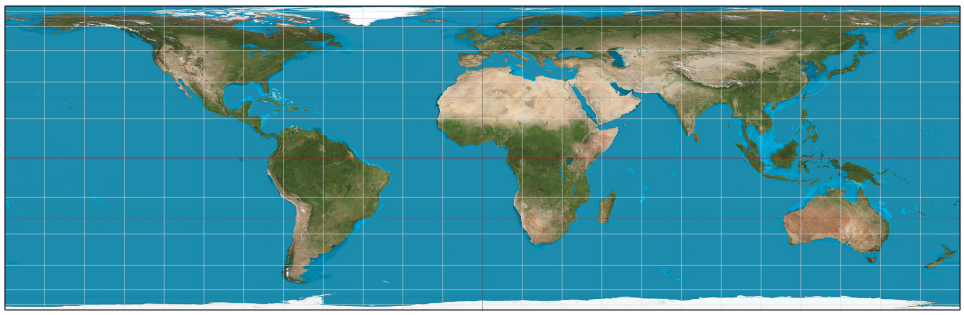
When $h(\phi) = \sin\phi$, we have $h'(\phi) = \cos\phi$, so the parallel and meridian distortion factors of Φ_h are inverses. This Φ_h, called the *Lambert projection*, preserves area.

Informally, a small 'rectangle' on S with meridian and parallel sides will be taken to a small Euclidean rectangle by Φ_h. As the parallel and meridian distortion factors of Φ_h are inverses, the two pairs of opposite sides will have been scaled by inverse factors. So, the two rectangles will have the same area.

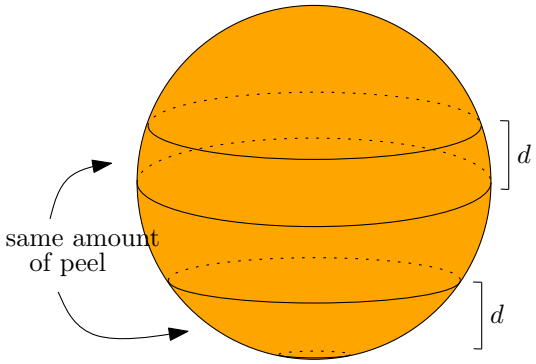
Geometrically, the Lambert projection takes a point $p \in S$ to the point on C with the same x, y-coordinates as the radial projection, but with z-coordinate the same as p. You can think of it as projecting straight out horizontally from the z-axis.



You can see the area preservation in the following picture: e.g. Greenland's area is to scale with the rest of the map, even though the country appears very stretched.



EXERCISE 2.4.1. Hold two knives in parallel at a distance of d from each other, and use them to cut a ring out of a (radius 1, spherical) orange. Using the Lambert projection, show that the area of the peel produced depends only on d, not on where the cuts are made.



On the other hand, if the two distortions agree at a point p then the projection of a very small right triangle near p that has base and height along a parallel and a meridian will look like a small right triangle near $\Phi_h(p)$, and the base and height will both have been scaled by the common distortion factor. So, the two triangles will be similar, and in particular the angles are preserved by Φ_h. As an example, if we set

$$h(\phi) = \ln\left(\sec\phi + \tan\phi\right),$$

then you can check $h'(\phi) = \sec\phi$, so the two distortion factors agree, and the projection Φ_h, called the *Mercator projection*, is conformal. In the picture of the Mercator projection below, areas are massively distorted, but the distortion at a point happens uniformly in all directions, so angles are undistorted.

There is a beautiful physical way to picture the Mercator projection. Imagine that the sphere is a balloon, and the enclosing cylinder is rigid. Inflating the balloon smashes it against the cylinder, and if the continents were recently painted, the wet paint will transfer to the cylinder to give the Mercator projection. It's not hard to convince oneself that this is the right picture; the inflation will give a cylindrical

projection, by symmetry, and it'll be conformal since the balloon stretches equally in all directions near a point on its surface.

Besides its aesthetic value, conformality of the projection is useful for navigation. Namely, imagine you are a ship's captain and you have your trusty compass at hand. The easiest way for you to navigate is to try to keep your ship heading always in the same compass direction. Since 'north' always points along *meridians*, i.e. great circles going through the north and south poles, your course will always make the same angle with all meridians. Since meridians are mapped to vertical lines under cylindrical projections, if h is conformal your course will track a path on the map that always makes the same angle with vertical lines. In fact, such a path is a line:

EXERCISE 2.4.2 (Read the previous paragraph first). Suppose that a path $\gamma : [0,1] \longrightarrow \mathbb{R}^2$ has the property that $\gamma'(t)$ always makes an angle of θ with the vertical. Show that γ is a line segment. *Hint: integrate $\gamma'(t)$ to find $\gamma(t)$.*

So, courses traveling in a constant compass direction, called *rhumb lines* in navigation, correspond to lines in the Mercator projection. *The reader that is too smart for his/her own good might object that bearing 'due north' takes one to the magnetic north pole, rather than the North Pole. In this discussion, and in the following exercise, we ignore such technicalities and assume they are one and the same.*

EXERCISE 2.4.3. Describe, and draw, the rhumb lines on the sphere that you get from traveling indefinitely in the directions north, east, and northeast, respectively. You should see some surprising behavior in the last case! *(Your pictures should be on a sphere, not on a map.)*

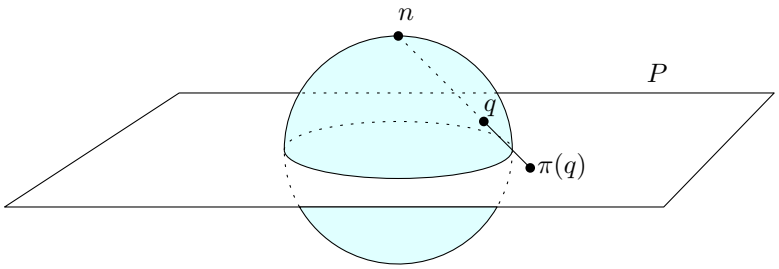
Incidentally, Mercator definitely had navigation by compass in mind when he developed his projection, which he called *Nova et Aucta Orbis Terrae Descriptio ad Usum Navigantium Emendata*. This translates to 'A new and augmented description of Earth corrected for the use of sailors.'

Our next projection, *stereographic projection*, is less useful for navigation but very important in mathematics. It is also used in 'stereographic fisheye lenses' in photography. Below are two examples of pictures taken with such a lens. The first is of Paris (see the Eiffel tower?) and the second of Ueno station in Tokyo.


Writing $n = (0, 0, 1)$, the *stereographic projection of S onto* $\mathbb{R}^2$ is the map

$$\pi : S \backslash \{n\} \longrightarrow \mathbb{R}^2$$

defined by letting $\pi(q)$ be the point at which the ray from n in the direction of q intersects the plane xy-plane, which we identify with $\mathbb{R}^2$.



To write down π in coordinates, note that if $q = (x, y, z)$ then the line through n and q can be parameterized as $\gamma(t) = t(x, y, z) + (1-t)(0, 0, 1) = (tx, ty, t(z-1)+1)$. The intersection of γ with the xy-plane happens when $t = \frac{1}{1-z}$, so

$$\pi(x, y, z) = \left(\frac{x}{1-z}, \frac{y}{1-z} \right).$$

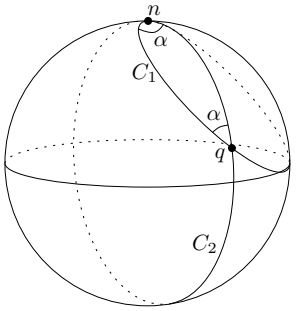
Exercise 2.4.4. Using a similar argument, show that if $(x, y) \in \mathbb{R}^2$ then we have

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

Stereographic projection has a number of interesting properties. It takes meridians of S to lines passing through the origin of $\mathbb{R}^2$, and takes parallels to circle centered at the origin. Moreover, like the Mercator projection,

Proposition 2.4.5. *Stereographic projection is conformal, i.e. angle preserving.*

Proof. Two directions at a point $q \in S$ can be realized as the tangents of two circles C_1 and C_2 on S passing through q and n. Let P_i, where $i = 1, 2$, be the planes in $\mathbb{R}^3$ with $P_i \cap S = C_i$. Note that $\pi(C_i)$ is the line in which P_i intersects the xy-plane.



Let P be the perpendicular plane through the midpoint of the segment nq in $\mathbb{R}^3$. As the triangle $0nq$ is isosceles, P passes through the origin, and P is perpendicular to both P_1, P_2, which contain nq. So, reflection through P preserves P_1, P_2, and restricts to an isometry of S that preserves the two circles but exchanges their intersection points. So, the angles α at the two intersection points of C_1 and C_2 are the same.

Stereographic projection takes C_i to the intersection of P_i with the xy-plane. The angle made by two horizontal vectors in P_1 and P_2 based at n is α, and the same angle is made by two horizontal vectors in P_1 and P_2 based at $\pi(q)$, which is the angle at which the lines $\pi(C_1)$ and $\pi(C_2)$ meet. Thus, stereographic projection is conformal. □

Exercise 2.4.6. Give a less rigorous argument for conformality by showing that the lattitudinal/longitudinal stretch factors of the map π agree at every point, as we did when discussing cylindrical projections. *Hint: parametrize a longitude via*

$$\beta(\phi) = (\cos\phi, 0, \sin\phi).$$

The longitudinal stretch factor at $\beta(\phi)$ will be $|(\pi\circ\beta)'(\phi)|$, while the lattitudinal stretch factor you can calculate by comparing the length of the lattitude through $\beta(\phi)$ to the length of its image. Since π is rotationally symmetric, we'd get the same result looking at any other longitude.

Here's another even more amazing property of stereographic projection.

Proposition 2.4.7. *If C is a circle in S, then $\pi(C)$ is either a circle or a line. Conversely, if C is any circle or line in $\mathbb{R}^2$, then $\pi^{-1}(C)$ is a circle on S.*

There are a number of beautiful geometric proofs of this fact, but an algebraic argument is simpler and less confusing, in my opinion. The key is to have a uniform algebraic description of lines and circles in $\mathbb{R}^2$: they are exactly the infinite, proper solution sets of equations of the form

$$ax^2 + ay^2 + bx + cy + d = 0. \tag{5}$$

When $a = 0$, such an equation describes a line in $\mathbb{R}^2$. If $a \neq 0$, the equation may have no solutions, e.g. $x^2 + 1 = 0$, or a single solution, e.g. $x^2 + y^2 = 0$. However,

EXERCISE 2.4.8. Show that if Equation (8) has more than one solution and $a \neq 0$, then it describes a circle in $\mathbb{R}^2$. *Hint: complete the square.*

In fact, we'll discuss in Section 3.1 how it is worth having a blanket term *'lircle'* for lines and circles, which we'll often use without distinguishing between the two cases.

PROOF OF PROPOSITION 2.4.7. A circle C in S is the intersection of a plane with S – let's assume the plane is given by the equation $ax + by + cz = d$. Using the formula for the inverse of stereographic projection, $\pi(C)$ consists of all points $(x, y) \in \mathbb{R}^2$ satisfying

$$a\frac{2x}{1+x^2+y^2} + b\frac{2y}{1+x^2+y^2} + c\frac{x^2+y^2-1}{1+x^2+y^2} = d.$$

After some algebraic manipulation, this becomes

$$(c-d)(x^2+y^2) + 2xa + 2yb - c - d = 0. \tag{6}$$

Since a circle C on the sphere has infinitely many points, so does $\pi(C)$, so this equation must have infinitely many solutions. As it exactly has the form of Equation (8), it describes a line in $\mathbb{R}^2$ when $c = d$ and a circle otherwise. □

EXERCISE 2.4.9. Analyzing the proof of Proposition 2.4.7, show that $\pi(C)$ is a line exactly when C passes through $(0, 0, 1)$, and is a circle otherwise.

An advantage of stereographic projection in cartography is that if one is making maps of the moon, one sometimes wants craters to appear round. Proposition 2.4.7 ensures that this will be the case.

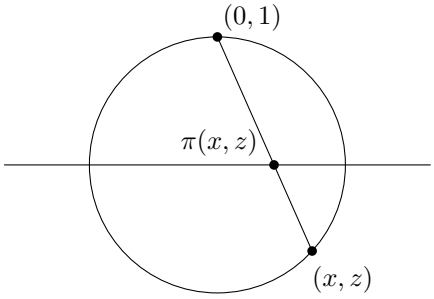
Stereographic projection has an interesting number theoretic property. Let's restrict our attention to the xz-plane in $\mathbb{R}^3$. Dropping the y-coordinate from our notation, the unit sphere in $\mathbb{R}^3$ intersects the xz-plane in the circle

$$C = \{(x, z) \mid x^2 + z^2 = 1\}$$

and stereographic projection restricts to the map

$$\pi : C\backslash\{(0,1)\} \longrightarrow \mathbb{R}, \quad \pi(x,z) = \frac{x}{1-z},$$

where the geometric interpretation is the same as before:



EXERCISE 2.4.10. If $x^2 + z^2 = 1$ and $z \neq 1$, show that $\frac{x}{1-z} \in \mathbb{Q}$ if and only if both $x, z \in \mathbb{Q}$.

A *Pythagorean triple* is a set of three positive integers x, y, z such that

$$x^2 + y^2 = z^2.$$

One example is $3^2 + 4^2 = 5^2$. There are certainly infinitely many such triples, since any Pythagorean triple can be scaled by an integer to produce another. For instance,

$$(3n)^2 + (4n)^2 = (5n)^2, \quad \text{for all } n \geqslant 1.$$

A Pythagorean triple (a, b, c) is called *primitive* if the greatest common divisor of a, b, c is one. For instance, $(3, 4, 5)$ is primitive but $(6, 8, 10)$ is not.

EXERCISE 2.4.11. *Using the previous problem,* show that there are infinitely many primitive Pythagorean triples.

In fact, Euclid's formula states that every primitive Pythagorean triple (a, b, c) can be expressed as $a = m^2 - n^2$, $b = 2mn$, $c = m^2 + n^2$ for some integers $m > n > 0$.

2.5. Triangulating spherical polygons

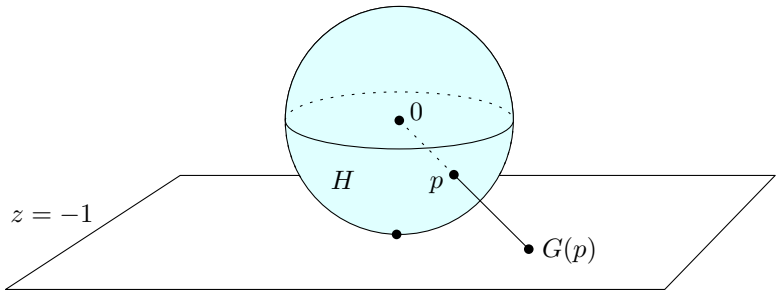
If P is a polygon on S_r, a *triangulation* of P is a collection of triangles on S_r (as always, with sides that are great circle segments) that union to P, and where two triangles intersect either along an edge of both, or a vertex of both.

We saw in §1.7 that Euclidean polygons can always be triangulated. Is the same true for proper spherical polygons? One way to try to answer this question would be to repeat the proof that we did in the Euclidean setting, hoping that all the details still work. Instead of attempting this, we'll show that proper spherical polygons can be transformed into Euclidean polygons, which can then be triangulated, and such a Euclidean triangulation gives a triangulation of the original polygon.

The key is to study the 'gnomonic projection' of a hemisphere of S_r. Set

$$H = \{(x, y, z) \in S \mid z < 0\}, \quad Z = \{(x, y, z) \in \mathbb{R}^3 \mid z = -1\}.$$

The *gnomonic projection* is the map $G : H \longrightarrow Z$, where if $p \in S$, we let $G(p)$ be the point where the ray from 0 in the direction of p intersects the plane $z = -1$.

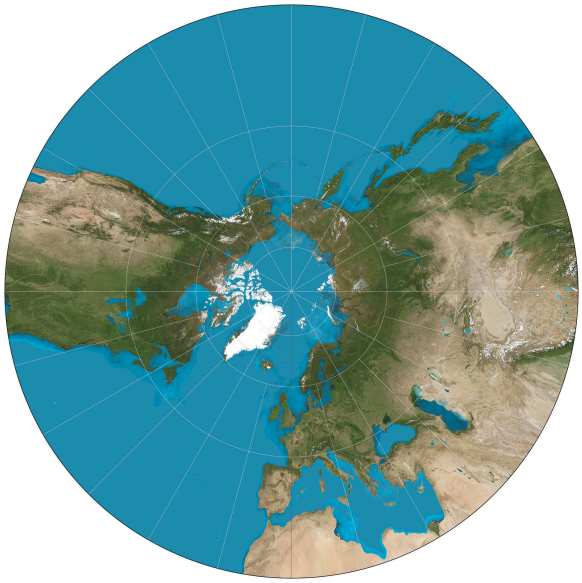


In coordinates, the ray from the origin through a point $p = (x, y, z)$ can be parameterized as $\gamma(t) = (tx, ty, tz)$. This ray intersects the plane at height -1 when $tz = -1$, so $t = -1/z$. Therefore, gnomonic projection is the map

$$G : H \longrightarrow Z, \quad G(x, y, z) = \left(-\frac{x}{z}, -\frac{y}{z}, -1\right).$$

Gnomonic projection has a useful property that distinguishes it from those previously considered: it maps great circles on S to lines in Z. For a great circle on the sphere is the intersection $P \cap S$, where P is a plane through the origin, and gnomonic projection maps $P \cap S$ to the intersection of P with the plane Z, which is a line.

So, gnomonic projections are useful for plotting efficient aerial trajectories between points on the earth. Here is part of a map created by gnomonic projection, when the earth is oriented upside down so that the North Pole is $(0, 0, -1)$.



EXERCISE 2.5.1. Gnomonic projection takes great circles to lines, while the Mercator projection and stereographic projection preserve angles. Is there any way to define a projection from part (say, a hemisphere) of S_r into $\mathbb{R}^2$ that has both properties?

We can now prove the following theorem.

THEOREM 2.5.2. *Any proper n-gon on S_r can be triangulated with $n - 2$ triangles.*

PROOF. Suppose that P is a proper spherical n-gon. Rotating it on S_r, we can assume that P lies in H. Since the gnomonic projection G takes great circle segments to line segments, the image $G(P)$ is a Euclidean n-gon in the plane Z. So, $G(P)$ can be triangulated with $n - 2$ Euclidean triangles. The preimages $G^{-1}(P)$ of these triangles are spherical triangles that triangulate P. □

EXERCISE 2.5.3. Come up with an example of a (non-proper) spherical quadrilateral that cannot be triangulated. *Remember, for us every triangle in a triangulation has vertices that are vertices of the original polygon.*

2.6. Area of spherical polygons and Euler characteristic

Girard's theorem gave an amazing formula for the area of a spherical polygon in terms of its angles. Our first goal in this section is to generalize this to an area formula for proper spherical polygons.

THEOREM 2.6.1. *If P is a proper n-gon on S_r with interior angle sum s, then*

$$\text{Area}(P) = r^2 \left(s - (n-2)\pi\right).$$

Recall that the interior angle sum of a Euclidean n-gon is $(n-2)\pi$. So, just as in Girard's theorem, the corollary is stating that area is r^2 times the angle excess.

PROOF. We'd like to use Girard's theorem, so it would be convenient to have a triangulation of P. Rotating P does not change its angles or its area, so we may assume that this is the southern hemisphere. Gnomonic projection then sends P to a Euclidean polygon, which can be triangulated. The gnomonic inverse of this triangulation is then a triangulation $T_1 \cup \cdots \cup T_{n-2} = P$. Let s_i be the interior angle sum of T_i. As $\sum_i s_i = s$, we have

$$\text{Area}(P) = \sum_{i=1}^{n-2} \text{Area}(T_i) = \sum_{i=1}^{n-2} r^2(s_i - \pi) = r^2 \left(s - (n-2)\pi\right). \quad \square$$

EXERCISE 2.6.2. Show directly that the conclusion of Theorem 2.6.1 holds for lunes and monogons.

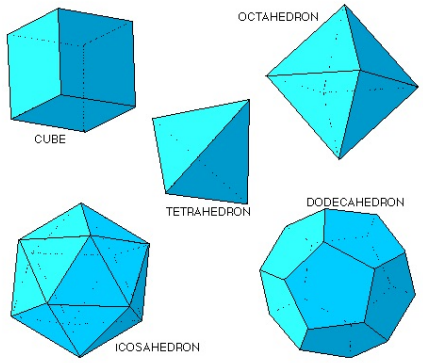
EXERCISE 2.6.3. Show that the conclusion of Theorem 2.6.1 holds for polygons P that are *complements* of proper polygons.

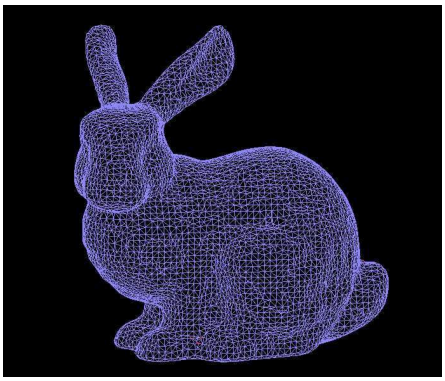
This is a surprising application to a certain invariant of polyhedra in $\mathbb{R}^3$. If P is a polyhedron, define the *Euler characteristic* of P to be the number

$$\chi(P) = V - E + F,$$

where V, E, F are the numbers of vertices, edges and faces of P, respectively. Let's compute the Euler characteristic of some of the polyhedra below, that you may remember from the section on scissors congruence.

The tetrahedron has 4 vertices, 6 edges and 4 faces, so $\chi = 4 - 6 + 4 = 2$. The cube has 8 vertices, 12 edges and 6 faces, so $\chi = 8 - 12 + 6 = 2$. In fact, you can compute by hand (although this will be hard for the rabbitohedron) that the Euler characteristic of any of the polyhedra pictured is 2!





For convex polyhedra P, we can explain this using spherical area. Position P so that it contains the origin, and let r be large enough so that P does not intersect S_r. Then radially projecting the vertices, edges and faces of P from the origin gives a *proper tiling* of S_r, by which we mean a collection of proper spherical polygons on S_r whose union in S_r, and where the intersection of two given polygons is either a vertex or an edge of each.

EXERCISE 2.6.4. Using convexity, explain why the radial projection from the boundary of P to S_r is a bijection, and conclude that the numbers of faces, edges and vertices of the spherical tiling are the same as those of P.

EXERCISE 2.6.5. If P is not convex, one can still project its edges and vertices onto S_r, but the projections of two edges may cross, so the induced tiling of S_r looks like it has more vertices than P does. Try to draw a picture of this happening.

Like a polyhedron, a tiling of S_r has an Euler characteristic $\chi = V - E + F$, where the Euclidean polygons and edges are replaced with their spherical analogues.

THEOREM 2.6.6. *The Euler characteristic of a proper tiling of S_r is* 2.

PROOF. The interior angles around a vertex of the tiling sum to 2π, so the sum of all interior angles in all the polygons is $2\pi V$. Also, every edge is contained in two polygons, so E is half the sum of the numbers of sides $n(P)$ in the polygons P. So, if $s(P)$ is the interior angle sum of P, we have

$$\begin{aligned}
4\pi r^2 &= \text{Area}(S_r) \\
&= \sum_{\text{polygons } P} \text{Area}(P) \\
&= \sum_{\text{polygons } P} r^2\Big(s(P) - \pi(n(P) - 2)\Big), \qquad \text{by Corollary 2.6.1} \\
&= r^2 \left(\sum_{\text{polygons } P} s(P) - \pi \sum_{\text{polygons } P} n(P) + \pi \sum_{\text{polygons } P} 2 \right)
\end{aligned}$$

$$= r^2(2\pi V - 2\pi E + 2\pi F)$$
$$= 2\pi r^2 \chi,$$

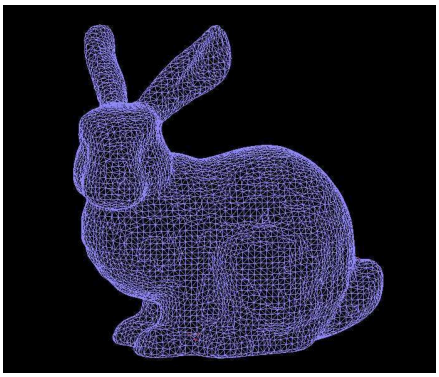
This implies that $\chi = 2$. □

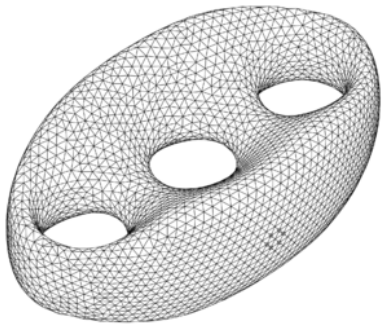
EXERCISE 2.6.7. Suppose that $P \subset S_r$ is a proper polygon. A *tiling* of P is a collection of proper[1] spherical polygons whose union is P, and where any two of the polygons intersect in either a vertex or an edge of both. Show that any tiling of P has Euler characteristic 1. *Hint: explain why it suffices to just repeat the proof above, but using Exercise 2.6.3 instead of Corollary 2.6.1.*

By Exercise 2.6.4, the numbers of vertices, edges and faces of a convex polyhedron are the same as those of its associated spherical tiling, so we obtain:

COROLLARY 2.6.8. *The Euler characteristic of a convex polyhedron is* 2.

In fact, convexity here is not really necessary. It suffices only to assume that the polyhedron does not have 'holes' – this is the case for the rabbit on the left, while the polyhedron on the right has three holes. Intuitively, the surface of a polyhedron with no holes can be smoothed out along the surface of a sphere, and the faces can then be straighten to spherical polygons, so that the preceding argument will still work.





The number of holes of a polyhedron is usually called its *genus*. In general, it turns out that the Euler characteristic of a genus g polyhedron is $2 - 2g$! This is a little tricky to prove (and state precisely) in general, but you can verify it in special cases.

EXERCISE 2.6.9. For each g, describe the construction of a specific polyhedron with g holes in which you can easily show that the Euler characteristic is $2 - 2g$.

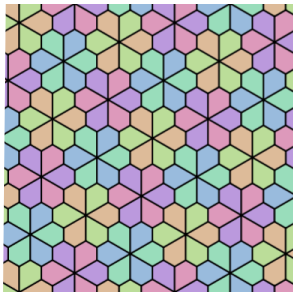
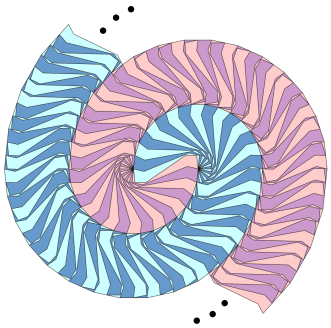
[1]It's not really necessary to say proper here, since all the polygons are subsets of P, which is proper, so they're automatically proper too.

2.7. Tilings of $\mathbb{R}^2$

We say that two subsets A, B of a metric space X are *congruent* if there is an isometry $f : X \longrightarrow X$ with $f(A) = B$.

EXERCISE 2.7.1. If two triangles T, T' in $\mathbb{R}^2$ have the same side lengths, they are congruent. *Hint: by Exercise 1.1.11, the two triangles must also have the same angles. You might try composing a translation with a suitable rotation and reflection.*

A *monohedral tiling* of $\mathbb{R}^2$ is a collection of congruent polygons $P_1, P_2, \ldots$ with disjoint interiors that union to $\mathbb{R}^2$. Some examples are pictured below. The colors and the bees are unimportant to the mathematics.

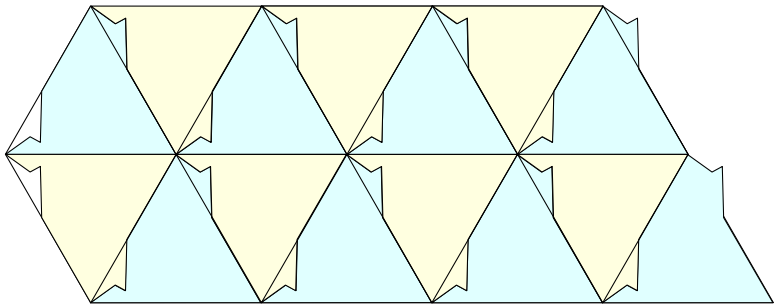

On the left is a tiling by congruent pentagons, the middle tiling is by congruent 9-gons, and the honeycomb in the last picture is an example in nature of part of a tiling of $\mathbb{R}^2$ by regular hexagons.

EXERCISE 2.7.2. Show that if T is *any* triangle, there is a monohedral tiling of $\mathbb{R}^2$ in which all the polygons are congruent to T. *Be careful here... you have to start with an arbitrary triangle and explain why the tiling exists. Don't just draw a few triangles and say they're all congruent.*

EXERCISE 2.7.3 (Harder). Show that for any quadrilateral Q, there is a monohedral tiling of $\mathbb{R}^2$ in which all the polygons are congruent to Q. *Hint: one way to do this is to combine two copies of your given quadrilateral into a hexagon that more obviously tiles the plane.*

EXERCISE 2.7.4. Show that there is a monohedral tiling of $\mathbb{R}^2$ in which all the polygons are regular n-gons if and only if $n = 3, 4, 6$. *Hint: look at the interior angles at a vertex in the tiling.*

For each n, there is a wealth of tilings of the plane by congruent n-gons. When $n = 3, 4$, this is clear from Exercises 2.7.2 and 2.7.3. In general, one can start with the tiling of the plane by equilateral triangles pictured below, cut out a polygonal piece from the one side of each triangle and glue it on one of the other sides.

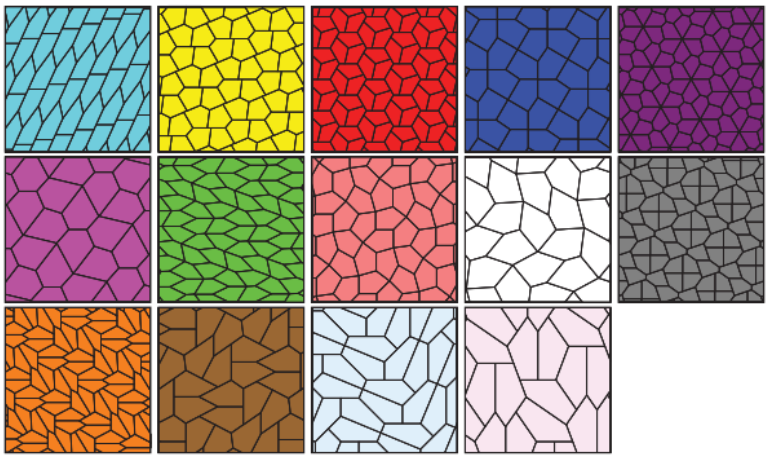


If this is done properly as above, the result is a monohedral tiling by $(2s+1)$-gons, where s is the number of sides with which we replaced one side of each equilateral triangle. For example, above $s = 4$ and the resulting tiling is by 9-gons. The only constraint in this construction is that $s \geqslant 2$; so this produces infinitely many 'distinct' tilings by congruent n-gons whenever $n = 2s + 1 \geqslant 5$ is odd.

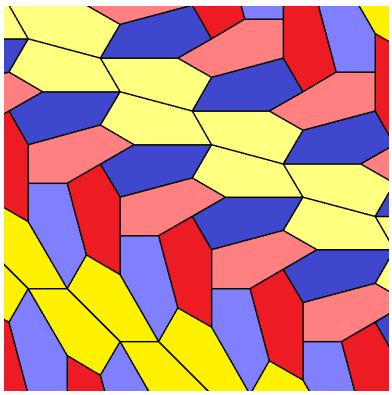
EXERCISE 2.7.5. With a similar construction, produce tilings by congruent n-gons whenever $n \geqslant 4$ is even.

What about monohedral tilings by *convex* n-gons? As mentioned above, any triangle or quadrilateral can be used to tile the plane, so the question is only interesting when $n \geqslant 5$. Rienhardt (1918) gave a complete classification of the convex hexagons that tile the plane: they fall into three families, defined using conditions on angles and side length. He also found five families of convex pentagonal tilings. This was the state-of-the-art for convex pentagonal tilings until 1968, when Kershner found some additional families and claimed that his list was complete.

In 1975, Martin Gardner wrote an expository article on convex pentagonal tilings in *Scientific American* that included Kershner's list. Soon afterwards, a reader named Richard James wrote in with a new tiling, showing Kershner's claim of completeness to be false! Even better, in 1977 a reader named Marjorie Rice, a stay-at-home mother and amateur mathematician living in San Diego, discovered 4 additional families of convex monohedral pentagonal tilings. After Rolf Stein discovered one additional family in 1985, the total number of known families of convex pentagonal tilings was 14. A representative of each family is pictured below.



This was the state of the art until 2015, when Mann, McLoud, and Von Derau found a fifteenth tiling, pictured below.

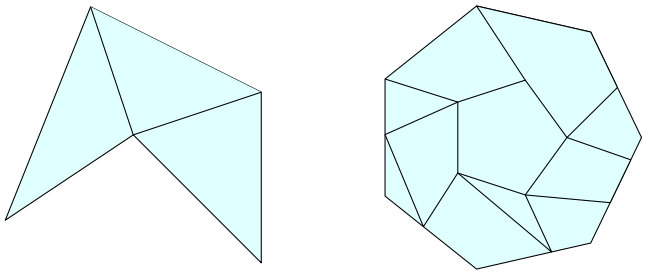


The problem was then settled in July 2017, when Michael Rao showed that the resulting list of fifteen families is complete!

Perhaps surprisingly, there are no other convex monohedral tilings.

THEOREM 2.7.6. *There is no monohedral tiling of* $\mathbb{R}^2$ *by convex* n*-gons when* $n > 6$.

The rest of the section is devoted to the proof of Theorem 2.7.6. Suppose that P is a polygon in $\mathbb{R}^2$. A *tiling* of P is what you would expect: it is a collection of polygons, that intersect in vertices or edges, and that union to P. Triangulations are examples, but in general a tiling may have vertices in the interior of P and its polygons may not be triangles, as in the second example pictured below.

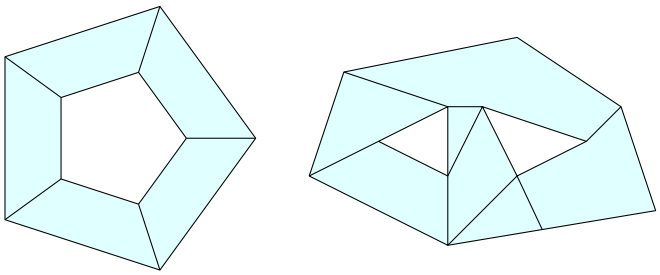


The definition of Euler characteristic $\chi = V - E + F$ makes perfect sense for a tiling of a Euclidean polygon.

LEMMA 2.7.7. *For any tiling of a Euclidean polygon P, $\chi(P) = 1$.*

PROOF. Use the inverse of gnomonic projection to transform a tiling of P into a tiling of a proper spherical polygon Q. Then Exercise 2.6.7 says that $\chi(P) = 1$. □

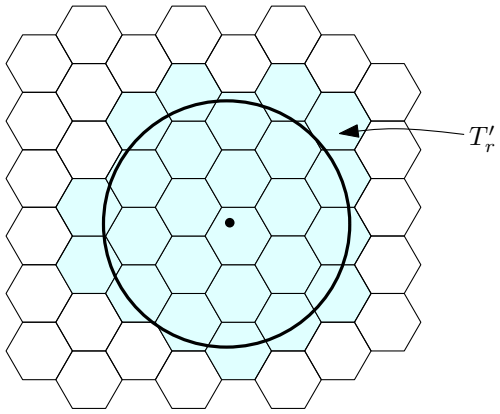
EXERCISE 2.7.8. Instead of polygons, one could tile more general regions of the plane. What do you think the Euler characteristic records? Try out the following examples to get some intuition.



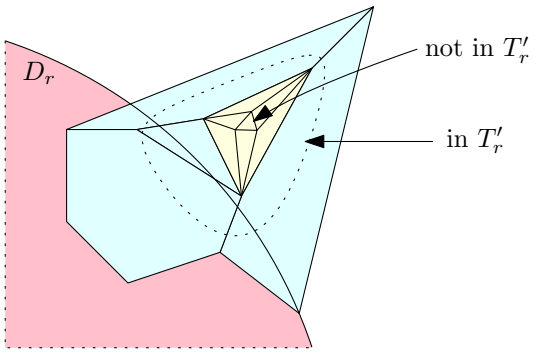
We are now ready to start the proof of the theorem above.

PROOF OF THEOREM 2.7.6. Suppose that we have a tiling of $\mathbb{R}^2$ by polygons that are all congruent to an n-gon P. We want to show that $n \leqslant 6$.

Fix $r > 0$, and let T'_r be the union of all polygons in the tiling that intersect the disc $D_r = \{x \in \mathbb{R}^2 \mid |x| \leqslant r\}$, as pictured below.



We'd like to say that T_r' is a polygon, and indeed, it looks like the polygon in the picture above. But what some of the polygons in T_r' enclose other polygons of the tiling that don't touch D_r, as in the following picture? Then T_r' won't be a polygon, as it will have a hole in the middle of it.



It's true that this picture doesn't look much like a monohedral tiling, but it's hard to say that something like this never happens. So instead, we set T_r to be the union of all polygons in the tiling that either lie in T_r' or are enclosed by a loop in T_r'. Then any 'holes' in T_r' are filled in when we look at T_r, so T_r is a polygon. (Try to write a more formal argument for this if you like.) Therefore, $\chi(T_r) = 1$.

Let V, F, E be the number of vertices, polygons (faces) and edges in the tiling of T_r induced by our tiling of $\mathbb{R}^2$. An *exterior vertex* is one that is a vertex of the polygon T_r, while an *interior vertex* is a vertex of one of the polygons in our tiling of T_r that lies in the interior of T_r. Let the number of interior/exterior vertices be V_{int} and V_{ext}, so that we have $V = V_{int} + V_{ext}$. The total angle sum of all the polygons in T_r is at most $2\pi V$, since each interior vertex contributes 2π while the other vertices contribute less than 2π. On the other hand, the angle sum of each polygon is $(n-2)\pi$, so the interior angle sum in T_r is $(n-2)\pi F$. Therefore,

$$(n-2)\pi F \leqslant 2\pi V \implies \frac{n-2}{2} F \leqslant V.$$

As the polygons in T_r are convex, their angles are less than π. So, each interior vertex in T_r is adjacent to at least three edges, implying

$$3V_{int} \leqslant 2E.$$

We now use these inequalities in the definition of the Euler characteristic of T_r, which we said above is 1, giving

$$1 \ = \chi(T_r) = \ V - E + F \ \leqslant \ V - \frac{3}{2}V_{int} + \frac{2}{5}V \ = \ \left(\frac{n}{n-2} - \frac{3}{2}\frac{V_{int}}{V}\right)V.$$

The key now is in the following claim:

Claim 2.7.9. *As $r \longrightarrow \infty$, we have $\frac{V_{int}}{V} \longrightarrow 1$.*

Assuming the claim for a minute, the inequality above says that when r is large, we have $0 \leqslant (\frac{n}{n-2} - \frac{3}{2})$, so $n/(n-2) \geqslant 3/2$, and solving for n we have $n \leqslant 6$. So, this finishes the proof of the theorem. □

It remains to prove the claim.

Proof of Claim 2.7.9. Let's call a polygon in our tiling of T_r an *exterior polygon* if it shares an edge with T_r, and an *interior polygon* otherwise. Let's indicate the number of interior polygons by F_{int}, and exterior polygons by F_{ext}. The intuition here is that V_{int} and F_{int} should both be proportional to the area πr^2 of D_r, while V_{ex} and F_{ext} should be proportional to the circumference $2\pi r$ of D_r, and for large r, we have $\pi r^2 >> 2\pi r$.

Let D be bigger than the diameter of the polygon P to which all our tiles are congruent. Then the union of all interior polygons of T_r must contain the disk of radius $R - D$ around the origin. Thus,

$$\text{Area}(P)F_{int} \geqslant \pi(R - D)^2 \ .$$

Similarly, every exterior polygon lies outside the disk of radius $R - D$, and within a disk of radius $R + D$, so we have

$$\text{Area}(P)F_{ext} \leqslant \pi\left((R + D)^2 - (R - D)^2\right) = 4\pi DR \ .$$

Summing the number of vertices from each tile over all of the interior tiles, we end up counting each vertex according to the number of edges touching it. Since the sum of the angles formed at each vertex is 2π, this number is at most 2π divided by the minimum angle of P. Denoting this last quantity by $\alpha(P)$, we have

$$n \cdot F_{int} \leqslant \frac{2\pi}{\alpha(P)}V_{int} \ .$$

We may do the same count for the exterior tiles, where we may use the simpler observation that each exterior vertex is counted at least once when we sum the number of vertices over the exterior tiles. Thus

$$n \cdot F_{ext} \geqslant V_{ext} \ .$$

Putting this info together, we have

$$V_{int} \geqslant \frac{n\ \alpha(P)}{2\pi} \cdot \frac{\pi(R-D)^2}{\text{Area}(P)} = \frac{n\ \alpha(P)}{2\,\text{Area}(P)}(R-D)^2$$
$$V_{ext} \leqslant n \cdot \frac{4\pi DR}{\text{Area}(P)} = \frac{4\pi nD}{\text{Area}(P)}\ R\ .$$

This means we have

$$\frac{V_{ext}}{V_{int}} \leqslant \frac{\frac{4\pi nD}{\text{Area}(P)}R}{\frac{n\alpha(P)}{2\,\text{Area}(P)}(R-D)^2} = \frac{8\pi D}{\alpha(P)}\frac{R}{(R-D)^2}$$

which goes to zero as $R \to \infty$. Thus we have

$$\frac{V_{int}}{V_{int}+V_{ext}} = \frac{1}{1+\frac{V_{ext}}{V_{int}}} \to 1$$

as $R \to \infty$, as desired.

□

2.8. Pick's Theorem

An *integer point* in $\mathbb{R}^2$ is a point (n, m) where $n, m \in \mathbb{Z}$. The set of all integer points is denoted by $\mathbb{Z}^2 \subset \mathbb{R}^2$. An *integer polygon* is a polygon all of whose vertices are integer points. This section is devoted to the following theorem, which was proven by Georg Pick in 1899.

THEOREM 2.8.1 (Pick's Theorem). *Let $P \subset \mathbb{R}^2$ be an integer polygon. Let I be the number of integer points that lie in the interior of P, and let B be the number of integer points on the boundary of P. Then* $\text{Area}(P) = I + B/2 - 1$.

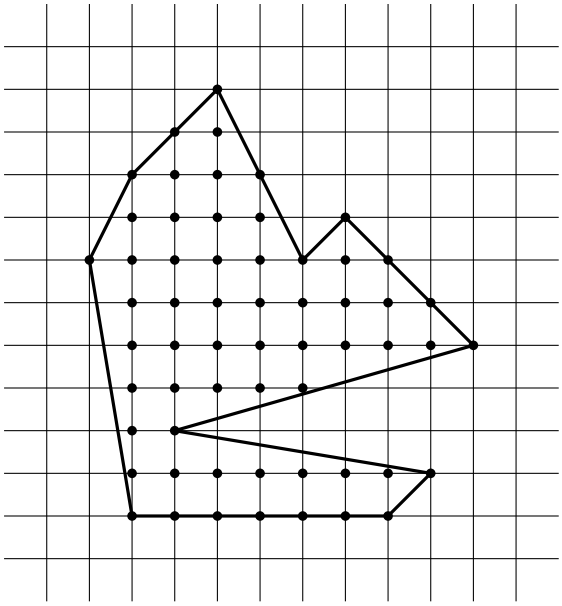
As an example, in the picture, we have $I = 40$ and $B = 19$, so the area is 48.5. This result is rather surprising! It's not too hard to believe that integer polygons large area should contain lots of integer points, perhaps even a number of integer points that's roughly proportional to the area. However, the precise formula in the theorem is a much more subtle result. vertices on the boundary As a warmup, try to draw some examples where you know how to compute area, using graph paper if you have it, and verify that the theorem holds for your examples! Once you've done that, let's work on the proof.

An integer polygon is *minimal* if the only integer points it contains are its vertices. For a minimal integer triangle, the right hand side in Pick's Theorem is

$$I + B/2 - 1 = 0 + 3/2 - 1 = \frac{1}{2}.$$

Our first step is to show that Pick's Theorem is true for such triangles.

LEMMA 2.8.2. *Suppose $\Delta \subset \mathbb{R}^2$ is a minimal integer triangle. Then* $\text{Area}(\Delta) = \frac{1}{2}$.



PROOF OF LEMMA 2.8.2. For notational convenience, let's translate Δ so that its vertices are $0, p, q$, where $p, q \in \mathbb{Z}^2$. Let $m = \frac{1}{2}(p+q)$ be the midpoint of the segment pq. Then the rotation by π around p has the formula

$$O_{m,\pi} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2, \quad O_{m,\pi}(x) = 2m - x = p + q - x.$$

Since $p, q \in \mathbb{Z}^2$, so is $p+q$. Therefore $x \in \mathbb{Z}^2$ if and only if $O_{m,\pi}(x) \in \mathbb{Z}^2$. It follows that $O_{m,\pi}(\Delta)$ is a minimal integer triangle. The union of Δ and $O_{m,\pi}(\Delta)$ is the parallelogram P with vertices $0, p, q, p+q$. Note that all vertices of P are integer points, and there are no other integer points in P.

It suffices to show that $\text{Area}(P) = 1$. Tile the plane with parallelograms congruent to P. The parallelograms in our tiling all have the form $T_{ip+jq}(P)$, where $i, j \in \mathbb{Z}$. Since $ip + jq$ is an integer point, each $T_{ip+jq}(P)$ has vertices that are integer points, and no other integer points. So, we have a tiling of the plane by parallelograms where the integer points in $\mathbb{R}^2$ are exactly the vertices of the tiling.

Since $(1, 0)$ is an integer point, there is a parallelogram $T_{ap+bq}(P)$ in our tiling that has $(1, 0)$ as a vertex, where here $a, b \in \mathbb{Z}$ play the role of i, j above. Moreover, we can assume that $(1, 0)$ is the translation by $ap + bq$ of the vertex 0 of P. So, $ap + bq = (1, 0)$. Similarly, there are $c, d \in \mathbb{Z}$ such that $cp + dq = (0, 1)$. Then

$$\begin{pmatrix} p & q \end{pmatrix} \begin{pmatrix} a & c \\ b & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

where $(p\ q)$ is the matrix with columns p and q. So,

$$1 = \det \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \det \begin{pmatrix} p & q \end{pmatrix} \det \begin{pmatrix} a & c \\ b & d \end{pmatrix},$$

and both determinants on the right hand side are integers, so they are both ± 1. In particular, Exercise 1.1.10 says that $\text{Area}(P) = |\det \begin{pmatrix} p & q \end{pmatrix}| = 1$. □

To deduce Pick's Theorem from the Lemma 2.8.2, we need to show:

LEMMA 2.8.3. *Let $P \subset \mathbb{R}^2$ be a polygon whose vertices are integer points. Then P can be tiled by minimal integer triangles.*

PROOF. Let's first do the proof when our polygon is a triangle Δ. Here, the argument is by strong induction on *badness* of Δ, which we define to be the number of integer points that lie in Δ but are not vertices of Δ. If the badness of Δ is zero, then Δ is a minimal integer triangle, and we are done. For the inductive case, suppose the claim holds for integer triangles P with badness less than n, and take an integer triangle Δ with badness n. Pick some integer point q in Δ that is not a vertex. If q lies on an edge of Δ, split Δ along the segment connecting q to the opposite vertex of Δ, into two integer triangles with smaller badness than Δ. By induction, these triangles can be tiled by minimal integer triangles. If q lies in the interior of Δ, split Δ along the line segments connected q to the vertices of Δ, giving three integer triangles with smaller badness than Δ. Applying the induction hypothesis again, these can be tiled by minimal integer triangles.

For general integer polygons P, triangulate P and apply the previous case. □

We can now prove Pick's Theorem. Let $P \subset \mathbb{R}^2$ be an integer polygon. Using Lemma 2.8.3, tile P by minimal integer triangles, and let V, E, F be the numbers of vertices, edges and faces of the tiling. Then we have $V - E + F = 1$ by Lemma 2.7.7. Since all the tiles have integer vertices, and no other integer points, the vertices of the tiling are exactly the integer points in P, so $V = I + B$.

The number of edges on the boundary of P is B, since it is the same as the number of vertices on the boundary of P. We claim

$$3F = 2(E - B) + B, \quad \implies E = \frac{1}{2}(3F + B)$$

To see this, note that each side counts the number of pairs (Δ, e), where Δ is a triangle in our tiling, and e is one of its edges. On the left side, we pick Δ first out of F possibilities, and then there are 3 choices for e. On the right, we pick e first. There are $E - B$ edges e in the interior of P, and each such edge has two adjacent triangles. And there are B edges on the boundary of P, each adjacent to one triangle.

Summing up, we get

$$1 = V - E + F = I + B - \frac{1}{2}(3F + B) + F = I + \frac{B}{2} - \frac{1}{2}F.$$

But by Lemma 2.8.2, we have $A = F/2$, implying that $F = 2A$, and hence

$$1 = I + B/2 - A, \quad \implies A = I + B/2 - 1.$$

2.8.1. Exercises.

EXERCISE 2.8.4 (Reeve tetrahedra, see [**3**]). Given $r \in \mathbb{N}$, show that the tetrahedron with vertices $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, and $(1,1,r)$ contains no integer points of $\mathbb{R}^3$ other than its vertices. Calculate the volume of this tetrahedron, and explain why no exact analogue of Pick's Theorem holds for integer polyhedra in $\mathbb{R}^3$.

EXERCISE 2.8.5. Figure out how to replace the last paragraph of the proof of Lemma 2.8.2 with an argument of the following form. Take r large, and look at the disc D_r of radius r around the origin. Let N_r be the number of parallelograms in the constructed tiling that intersect D_r, and let $I_r = |\mathbb{Z}^2 \cap D_r|$. Arguing as in proof of Claim 2.7.9, show that we have:
(a) $\lim_{r\to\infty} \text{Area}(D_r)/I_r = 1$,
(b) $\lim_{r\to\infty} \text{Area}(D_r)/N_r = \text{Area}(P)$,
(c) $\lim_{r\to\infty} I_r/N_r = 1$.
Then conclude that $\text{Area}(P) = 1$.

EXERCISE 2.8.6. Suppose that P is a (possibly non-integer) polygon in $\mathbb{R}^2$ with vertices (x_i, y_i), where $i = 1, \ldots, n$, and define $(x_{n+1}, y_{n+1}) := (x_0, y_0)$.

(a) Show that $\text{Area}(A) = \sum_{i=1}^{n} (y_i + y_{i+1})(x_i - x_{i+1})$, *Hint: compute the area of the trapezoid with vertices* $(x_i, 0), (x_{i+1}, 0), (x_{i+1}, y_{i+1}), (x_i, y_i)$.
(b) Using (a), show that $\text{Area}(A) = \sum_{i=1}^{n} \det \begin{pmatrix} x_i & x_{i+1} \\ y_i & y_{i+1} \end{pmatrix}$.

EXERCISE 2.8.7 (Blichfeldt's Theorem). Suppose that $X \subset \mathbb{R}^2$ is a subset with area A. Show that there is a translation $T_v : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ such that $T_v(X)$ contains at least A integer points. *Hint: first define* $[a,c] \times [c,d] := \{(x,y) \in \mathbb{R}^2 \mid x \in [a,b], y \in [c,d]\}$, *which is a rectangle in* $\mathbb{R}^2$. *For each pair of integers* i, j, *consider*

$$X_{ij} = X \cap ([i, i+1] \times [j, j+1]),$$

so in other words, X_{ij} *is just the part of* X *that lies in the unit square whose lower left corner is* (i,j). *Show that there is a point in the square* $[0,1] \times [0,1]$ *that is contained in* $T_{-(i,j)}(X_{i,j})$ *for at least* $\lceil A \rceil$ *different choices of* (i,j).

EXERCISE 2.8.8. Let $C \subset \mathbb{R}^2$. We say that C is *centrally symmetric* if whenever $x \in C$, then $-x \in C$. We say C is *convex* if whenever $x, y \in C$, the line segment $xy \subset C$. The point of this exercise is to prove (d) below, so go read (d), but before trying to prove it work on (a)–(c), which will help you understand the problem.
(a) Find a centrally symmetric, convex subset of $\mathbb{R}^2$ with area *exactly* 4 that doesn't contain any nonzero integer point. *Hint: the set doesn't need to include its boundary.*
(b) Find a convex subset of $\mathbb{R}^2$ with area bigger than 4 that doesn't contain any nonzero integer point.

(c) Draw a centrally symmetric subset of $\mathbb{R}^2$ with area bigger than 4 that doesn't contain any nonzero integer point.

(d) (Minkowski's Theorem) Show that any centrally symmetric, convex subset $C \subset \mathbb{R}^2$ with area bigger than 4 contains a nonzero integer point. *Hint: the set*

$$\frac{1}{2}C := \{(1/2)v \mid v \in C\}$$

has area bigger than 1*, since when you scale a shape by* r*, the area scales by* r^2*. Show that there are points* $p, q \in C$ *such that* $v = p/2 - q/2$ *is an integer point, and prove that* $v \in C$*.*

Exercise 2.8.9. Here's an approach to Lemma 2.8.2 using Minkowski's Theorem.

(a) If Δ is an integer triangle, show that $\text{Area}(\Delta) \geqslant \frac{1}{2}$, using an exercise from the first HW assignment.

(b) Suppose Δ is a minimal integer triangle. Construct an integer parallelogram P tiled by 8 copies of Δ such that P contains a single integer point in its interior. Using Exercise 2.8.8 (d), show that $\text{Area}(\Delta) \leqslant \frac{1}{2}$.

CHAPTER 3

Hyperbolic geometry

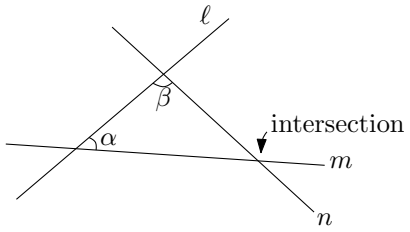
3.1. Euclid's axioms

Modern geometry, and in some sense modern mathematics, began with Euclid around 300 BC. Until Euclid, geometry operated on an intuitive level and the concept of a rigorous 'proof' was not codified. Euclid's book *The Elements* was a first attempt to set down an axiom-theorem framework for plane geometry. See [**2**] for an English translation, and Hartshorne [**1**] for a discussion of how Euclid's work relates to more modern work in geometry.

Euclid's book starts by defining common geometric terms like points, lines, right angles and circles, but in an abstract context without mentioning the Euclidean plane. He then introduces five axioms[1] that constrain how these objects behave.

1. "To draw a straight line from any point to any point."
2. "To extend a line segment continuously in a straight line."
3. "To describe a circle with any center and radius[2]."
4. "That all right angles are equal to one another."
5. The parallel postulate: "That, if a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles."

The meaning of the parallel postulate is that if the angles α, β below sum to less than π then the lines m and n intersect on the indicated side of ℓ.



[1]In most translations of *The Elements*, the five axioms listed above are called *postulates*. We'll use the more modern term 'axiom' here, except when referring to the parallel postulate, given the long history of that name and its alliterative appeal.

[2]Here, a *radius* for a circle is interpreted as a straight line segment starting at the center of the circle and terminating on the circle, rather than as a number.

A more succint alternative to the parallel postulate is *Playfair's axiom*, which is equivalent to it in the presence of the other four axioms. It was popularized by the mathematician John Playfair in a 1795 treatise on Euclidean geometry.

(P) "If a point p does not lie on a line ℓ, there is a unique line passing through p that is parallel to ℓ."

Euclid's axiom system above is not quite rigorous by modern standards. For instance, Proposition 1 in Book 1 is "To construct an equilateral triangle on a given finite straight line". In modern language, Euclid is saying that any line segment is a side of some equilateral triangle. To prove this, Euclid takes a line segment with endpoints p, q, and draws circles centered at p, q, respectively, where the line segment pq is a radius of each circle. He then sets z to be an intersection point of the two circles, and says that p, q, z is an equilateral triangle. However, there is no axiom above that says that guarantees that the circles above intersect.

EXERCISE 3.1.1. Imagine a version of plane geometry where the 'plane' is the subset $\mathbb{Q}^2 \subset \mathbb{R}^2$ of points with rational coordinates. Explain why all of Euclid's axioms hold in $\mathbb{Q}^2$, at least when suitably interpreted. Then show there is no equilateral triangle that has the segment from $(0, 0)$ to $(1, 0)$ as one of its sides.

Euclid's axioms were invented to characterize the geometry of the plane, so they should not hold for spherical geometry. However, since the axiom system isn't completely rigorous, it is a bit difficult to say which ones fail. If S_r is the sphere of radius r in $\mathbb{R}^3$, and we define a 'straight line' on S_r to be a great circle, then Axiom 1 holds. Axiom 2 holds in a certain sense, since any segment of a great circle can be extended in both directions to give the entire great circle, but it is not clear that Euclid would approve of this interpretation. Axiom 3 is true in the sense that given any p on the sphere and any $s > 0$, we can construct the metric circle $C(p, s)$ of radius s, as defined in §2.1. However, this circle degenerates to a point if $s = n\pi r$, $n \in \mathbb{N}$, which probably violates Euclid's earlier definition of a circle as a type of (non-straight) line. Alternatively, to Euclid a 'radius' of a circle should really be a line segment, and if one attempts to construct a circle on S_r using as a radius a line segment of length bigger than πr, the radius will intersect the circle at least twice, which Euclid may not want to allow. Some of the propositions in *The Elements* are indeed wrong on a sphere. For example, Proposition 32 in Book 1 states that the interior angles of a triangle sum to π. On a sphere, this contradicts Girard's Theorem, see §2.3. Propositions 16 is the first result in Euclid's work that fails for spherical geometry. The reason the proof fails is not related to the failure of axioms, though. Rather, it fails because Euclid makes an additional assumption about how lines should behave in a plane.

Historically, the parallel postulate was consider less self-evident than the first four axioms, and a great deal of effort was made to prove it using the other axioms. The first recorded such attempt was by Ptolemy (90-168). Proclus (410-485) explained why Ptolemy's proof was false, then gave a false proof of his own. Ibn al-Haytham

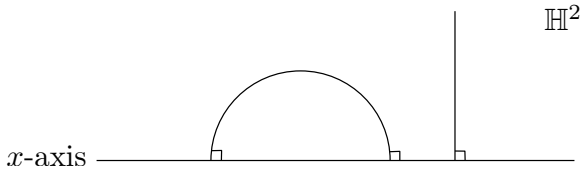
(965-1039) gave another false proof, but essentially invented the notion of an 'isometry' while doing so. Later false proofs were also given by Nasir al-Din al-Tusi (1201-1274), Giordano Vitale (1633-1711) and Girolamo Saccheri (1667-1733).

Around 1830 it was shown by Lobachevsky, Bolyai and Gauss (independently) that there is a geometry that satisfies all of Euclid's axioms except the parallel postulate, and unlike with the sphere, is intuitively a planar geometry. This geometry is now called the *hyperbolic plane*, and written $\mathbb{H}^2$. We'll build up hyperbolic geometry in stages – in this section, we will introduce it as a set and describe hyperbolic lines, noting that Playfair's axiom fails. Later, we'll see that $\mathbb{H}^2$ can be described as a metric space in which 'shortest paths' between points in $\mathbb{H}^2$ lie along hyperbolic lines.

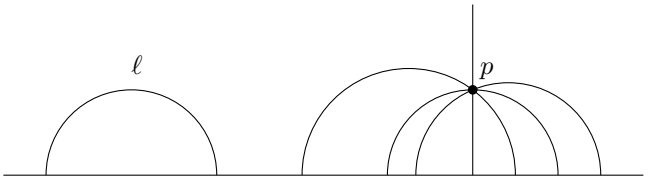
DEFINITION 3.1.2. The *hyperbolic plane* is defined to be the upper half plane

$$\mathbb{H}^2 = \{(x, y) \in \mathbb{R}^2 \mid y > 0\} \subset \mathbb{R}^2.$$

A *hyperbolic line* is the intersection $\ell \cap \mathbb{H}^2$, where ℓ is either a line or a circle in $\mathbb{R}^2$ that intersects the x-axis orthogonally.



As in Euclidean geometry, two hyperbolic lines are *parallel* if they don't intersect. Playfair's axiom fails for $\mathbb{H}^2$: if ℓ is a hyperbolic line and $x \in \mathbb{H}^2$ does not lie on ℓ, there are infinitely many hyperbolic lines through x that are parallel to ℓ. For instance, in the following picture many hyperbolic lines are drawn through p that do not intersect the given hyperbolic line ℓ.



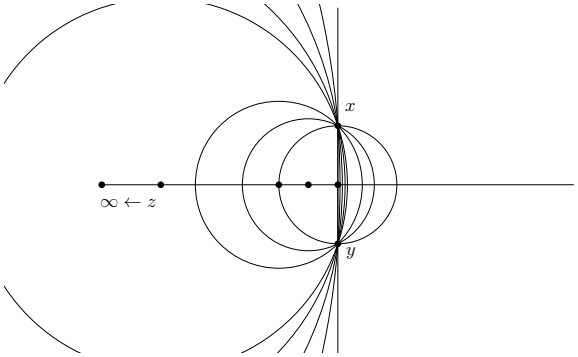
EXERCISE 3.1.3. Show with an example that the parallel postulate fails for $\mathbb{H}^2$. *Remember, 'straight line' now means hyperbolic line.*

3.2. Lircles

In §3.1, we introduced the hyperbolic plane, noting that hyperbolic lines are the intersections with $\mathbb{H}^2$ of lines or circles perpendicular to the x-axis. While it may seem strange to have this two-case definition of a hyperbolic line, we'll see in this section that a line can be considered as a degenerate version of a circle, where the center of the circle is at infinity. To this end, we define:

DEFINITION 3.2.1. A *lircle* in $\mathbb{R}^2$ is a subset that is either a line or a circle.

Here's one way to justify having a common term for lines and circles. Fix $x, y \in \mathbb{R}^2$. Any point z on the perpendicular bisector to xy is the center of a circle passing through both x and y. As $z \to \infty$ in either direction, the circle it determines converges to the line through x and y. So, lines are 'circles centered at infinity'.



Alternatively, consider quadratic equations of the form

$$ax^2 + ay^2 + bx + cy + d = 0, \tag{7}$$

where $a, b, c, d \in \mathbb{R}$. We call the equation *degenerate* if it has no solutions, a single solution, or if every $(x, y) \in \mathbb{R}^2$ is a solution. For example, the three equations

$$x^2 + y^2 + 1 = 0, \quad x^2 + y^2 = 0, \quad 0 = 0$$

are all degenerate, and have solution sets that are empty, a single point, and all of $\mathbb{R}^2$, respectively.

FACT 3.2.2. *Lircles are exactly the solution sets of nondegenerate equations*

$$ax^2 + ay^2 + bx + cy + d = 0. \tag{8}$$

The solution set is a line if $a = 0$, and a circle otherwise.

PROOF. First, note that any line is the solution set of a linear equation

$$bx + cy + d = 0,$$

while a circle with center (a, b) and radius r is the solution set of

$$(x - a)^2 + (y - b)^2 = r^2 \iff x^2 + y^2 - 2xa - 2by + (a^2 + b^2 - r^2) = 0,$$

which has the desired form.

Conversely, consider an equation $ax^2 + ay^2 + bx + cy + d = 0$ as above. Suppose first that $a = 0$. If $b = c = 0$, the equation is degenerate, as its solution set is either all of $\mathbb{R}^2$ or is empty, depending on whether $d = 0$ or not. Otherwise, the solution set is a line. So, we may now assume that $a \neq 0$. Dividing by a, the left side becomes

$$x^2 + y^2 + \frac{b}{a}x + \frac{c}{a}y + \frac{d}{a} = \left(x + \frac{b}{2a}\right)^2 - \left(\frac{b}{2a}\right)^2 + \left(y + \frac{c}{2a}\right)^2 - \left(\frac{c}{2a}\right)^2 + \frac{d}{a},$$

so the equation $ax^2 + ay^2 + bx + cy + d = 0$ is equivalent to

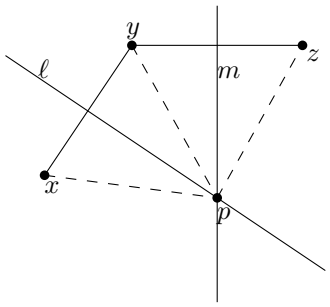
$$\left(x + \frac{b}{2a}\right)^2 + \left(y + \frac{c}{2a}\right)^2 = \left(\frac{b}{2a}\right)^2 + \left(\frac{c}{2a}\right)^2 - \frac{d}{a}.$$

If the right side is zero, this equation has a single solution, so our original equation is degenerate. Otherwise, the solution set is a circle centered at $q = \left(-\frac{b}{2a}, -\frac{c}{2a}\right)$ whose radius is the square root of the right hand side. □

Finally, having a blanket term for both also allows for simpler statements of some geometric facts.

LEMMA 3.2.3. *If $x, y, z \in \mathbb{R}^2$ are distinct, there is a unique lircle through x, y, z.*

PROOF. If x, y, z are collinear, then the line containing them is the unique lircle containing them. Otherwise, let ℓ and m be the perpendicular bisectors of xy and yz. Since xy and yz are not collinear, ℓ and m are not parallel, so they intersect at a point $p \in \mathbb{R}^2$.



The point p is equidistant from x, y, z, so there is a circle with center p through x, y, z. The center of a circle containing x, y, z must lie on both the perpendicular bisector of xy and the perpendicular bisector of xz, so the circle above is the unique circle containing all three points. □

We can now show that the hyperbolic plane satisfies the first of Euclid's axioms.

PROPOSITION 3.2.4. *If $x, y \in \mathbb{H}^2$, there is a unique hyperbolic line through x, y.*

PROOF. Let z be the reflection of x over the x-axis. By Exercise 3.2.4, there is a unique lircle C through x, y, z. If C is a line, it contains x, z, while if C is a circle, its center lies on the perpendicular bisector of xz, i.e. the x-axis. So, in both cases C is perpendicular to the x-axis, and intersects $\mathbb{H}^2$ in a hyperbolic line.

The hyperbolic line through x, y is unique: any lircle perpendicular to the x-axis is preserved by the reflection over the x-axis, so contains z, and therefore is C above. □

EXERCISE 3.2.5. Suppose that lircles ℓ, ℓ' intersect at two points $x, y \in \mathbb{R}^2$. Show that the angles of intersection at x, y are equal. In particular, ℓ, ℓ' are perpendicular at x if and only if they are perpendicular at y.

3.3. Incidence geometry

Although it was revolutionary for its time, from a modern perspective Euclid's book still lacks some rigor. It wasn't until the early 20th century that a more rigorous axiom system for plane geometry was developed by David Hilbert. His full system includes 21 axioms, and some definitions along the way, so instead of presenting the complete list we'll just discuss *incidence geometries*, which satisfy the first three of Hilbert's axioms. This should give a good feel of a rigorous development of plane geometry.

We start with a set P called the *plane*. Elements of this set are called *points* and there are certain subsets of the plane called *lines*. Here are three axioms:

I1. Given points $a, b \in P$, $a \neq b$, there is a unique line ℓ containing both a and b.
I2. Every line contains at least two points.
I3. There exist three noncollinear points, i.e. there are three distinct points in the plane that are not all contained in a line.

Any set P with a collection of subsets called lines that satisfies these three axioms is called an *incidence geometry*. The prototypical example is the *Euclidean plane* $\mathbb{R}^2$, where 'lines' are exactly the (bi-infinite) straight lines in $\mathbb{R}^2$. The hyperbolic plane $\mathbb{H}^2$, considered with its hyperbolic lines, is another example.

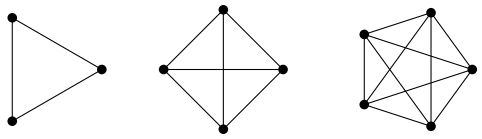
Here is a stranger example:

EXAMPLE 3.3.1. Suppose that $n \geqslant 3$ and $K_n = \{a_1, \ldots, a_n\}$ is a set consisting of n points. A subset of K_n is called a 'line' if it has two elements: that is, the lines are

$$\{a_i, a_j\}, \text{ where } i \neq j.$$

It is then clear that any two distinct points are contained in a unique line, which is just the set containing them. Certainly every line contains at least two points and the points a_1, a_2, a_3 are noncollinear.

Here are schematic representations of K_3, K_4 and K_5. There is a line segment drawn for every line in the geometry, but remember that the 'line' consists of just the endpoints of the segment, and we're only drawing the segment as a visual aid. Similarly, two lines intersect when they share an endpoint, not when the segments 'cross' somewhere in the middle of the polygon.



K_3 K_4 K_5

Here's a first result about incidence geometry.

PROPOSITION 3.3.2. *In an incidence geometry, two distinct lines can have at most one point in common.*

PROOF. Suppose lines ℓ and ℓ' intersect in more than one point, say at both a and b. As I1 states that there is a *unique* line containing both a and b, we have $\ell = \ell'$. $\square$

EXERCISE 3.3.3. Show that if p is a point in an incidence geometry, there are at least two distinct lines passing through p.

A much harder theorem is the following:

THEOREM 3.3.4 (De Bruijn – Erdös). *In a finite incidence geometry, the number of lines is greater than or equal to the number of points.*

PROOF. If a is a point, let $L(a)$ be the number of lines containing a. The number of points in a line ℓ is written $|\ell|$, as usual. Our first observation is:

$$\text{if } a \notin \ell, \text{ then } |\ell| \leqslant L(a).$$

Indeed, any point b in ℓ lies on a line m_b with a, by I1, and the lines m_b are all distinct, since if $m_b = m_{b'}$, we'd have that the two points b, b' are contained in both ℓ and the line $m_b = m_{b'}$, contradicting Proposition 3.3.2.

The total number of pairs (a, ℓ), where a is a point contained in the line ℓ, can be counted two ways, as on the two sides of the equality below.

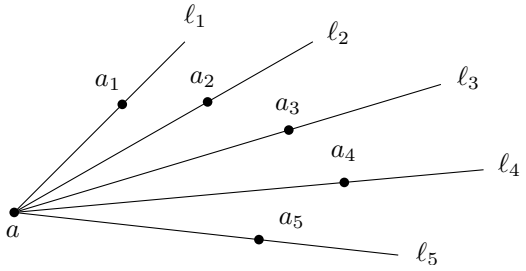
$$\sum_{\text{lines } \ell} |\ell| = \sum_{\text{points } a} L(a). \tag{9}$$

The idea is to show that the terms on the right can be paired up with terms on the left that are at least as big, so since we have equality there must be at least as many terms total on the right, i.e. as many lines as points.

So, let a be a point such that $L(a)$ is minimal. Write the lines through a as

$$\ell_1, \ldots, \ell_{L(a)},$$

and pick a point a_i on ℓ_i for each i. Note that again by Proposition 3.3.2, the points a_i are all distinct.



We now apply the central observation. As $a_{i+1} \notin \ell_i$ and $a_1 \notin \ell_L(a)$, we have

$$|\ell_i| \leqslant L(a_{i+1}), \text{ for } i = 1, \ldots, L(a), \quad \text{and} \quad |\ell_{L(a)}| \leqslant L(a_1).$$

So, from this and the central observation,

$$\begin{aligned}\sum_{\text{lines } \ell} |\ell| &= \sum_{i=1}^{L(a)} |\ell_i| \quad + \sum_{\text{lines } \ell,\ a\notin\ell} |\ell| \\ &\leqslant \sum_{i=1}^{L(a)} L(a_i) \quad + \sum_{\text{lines } \ell,\ a\notin\ell} L(a).\end{aligned}$$

Now, we also have that

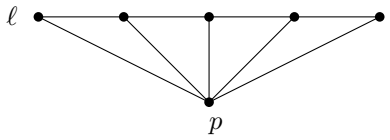
$$\begin{aligned}\sum_{\text{points } b} L(b) &= \sum_{i=1}^{L(a)} L(a_i) \quad + \sum_{b\notin\{a_1,\ldots,a_{L}(a)\}} L(b) \\ &\geqslant \sum_{i=1}^{L(a)} L(a_i) \quad + \sum_{b\notin\{a_1,\ldots,a_{L(a)}\}} L(a),\end{aligned}$$

by minimality of $L(a)$. So, by Equation (4) we have

$$\sum_{b\notin\{a_1,\ldots,a_{L(a)}\}} L(a) \leqslant \sum_{\text{lines } \ell,\ a\notin\ell} L(a),$$

and as the number of terms on the left is the total number of points minus $L(a)$, while the number of terms on the right is the total number of lines minus $L(a)$, there must be at least as many lines as points. □

Exercise 3.3.5. Show that if an incidence geometry has both n points and n lines, then there is a point p such that the rest of the points form a line ℓ with $n-1$ points, and every other line is a pair $\{p, x\}$, where $x \in \ell$.



Exercise 3.3.6. Prove the 'Sylvester-Gallai Theorem': for any finite set S of non-collinear points in $\mathbb{R}^2$, there is a line passing through exactly two of the points in S. *Hint: look at pairs (p, ℓ), where $p \in S$ and ℓ passes through at least two points of S, but not through p. Take such a pair where the distance from p to ℓ is minimal, and show that ℓ only passes through two points of S.*

Exercise 3.3.7. If $S \subset \mathbb{R}^2$ has $n \geqslant 3$ points and S is not contained in a single line, show that there are at least n lines in $\mathbb{R}^2$ that pass through at least two points of S. *This is exactly Theorem 3.3.4 for incidence geometries that are subsets of the plane,*

but you can give a much shorter proof using the Sylvester-Gallai theorem (Exercise 3.3.6) and induction on n.

If P, Q are incidence geometries, $f : P \longrightarrow Q$ is an *isomorphism* if it is a bijection and if $\ell \subset P$ is a line if and only if $f(\ell) \subset Q$ is a line, and we say that P, Q are *isomorphic* if such a map exists. Exercise 3.3.5 then says that any incidence geometry with n points and n lines is isomorphic to that described by the picture above.

Exercise 3.3.8. Show that an incidence geometry with four points is either isomorphic to K_4 or to the four element version of the geometry in Exercise 3.3.5.

Exercise 3.3.9. (Harder) Classify all five element incidence geometries. There are four, up to isomorphism, including K_5 and the geometry of Exercise 3.3.5.

Incidence geometry is really more combinatorics than geometry. However, we can add additional axioms to make it conform better to our geometric intuition. Euclid's axioms are natural candidates; but the first is exactly I1, and the rest of them do not make sense to state because we have not defined the terms 'line segment', 'circle' and 'angle' in an incidence geometry. On the other hand, Playfair's axiom only references lines, points and parallelism, so does make sense in this setting! Namely, we say that two lines ℓ, ℓ' in an incidence geometry are *parallel*, written $\ell \,||\, \ell'$, if they are disjoint, or if they are the same line. Playfair's axiom (P) can then be stated as follows:

(P) If a point p does not lie on a line ℓ, there is a unique line ℓ' with $p \in \ell'$ and $\ell \,||\, \ell'$.

Incidence geometries that satisfy Playfair's axiom (P) are called *affine planes*. The Euclidean plane $\mathbb{R}^2$ is an example, as are K_3 and K_4. On the other hand, if $n > 4$ then the geometry K_n does not satisfy (P): for then both $\{a_1, a_2\}$ and $\{a_1, a_3\}$ are lines containing a_1 that are parallel to $\{a_4, a_5\}$.

Proposition 3.3.10 (Parallelism is transitive). *Suppose ℓ_1, ℓ_2, ℓ_3 are lines in an affine plane. If $\ell_1 \,||\, \ell_2$ and $\ell_2 \,||\, \ell_3$, then $\ell_1 \,||\, \ell_3$.*

Proof. If not, then ℓ_1 and ℓ_3 are not the same and intersect at a single point p. They are then both lines through p that are parallel to ℓ_2, which contradicts (P). □

Parallelism is also *reflexive*, i.e. every line is parallel to itself, and *symmetric*, i.e. $\ell_1 \,||\, \ell_2 \iff \ell_2 \,||\, \ell_1$. So, in affine planes parallelism is an equivalence relation.

Exercise 3.3.11. Conversely, show that any incidence geometry in which parallelism is an equivalence relation (that is, where if $\ell_1 \,||\, \ell_2$ and $\ell_2 \,||\, \ell_3$, then $\ell_1 \,||\, \ell_3$) satisfies (P), so is an affine plane.

Exercise 3.3.12. Show that any two lines in a affine plane have the same number of points. *Hint: find a bijection between the points on one line and those on the other.*

Exercise 3.3.13. Show that if an affine plane P has a line with exactly n points, then the total number of points in P is n^2.

EXERCISE 3.3.14. Draw affine planes with 4 and 9 points, respectively, in the same way that we drew pictures of the geometries K_n.

A *field* is a set F together with addition and multiplication operations $+, \cdot$ satisfying all the usual commutativity and distributivity properties. Fields are assumed to contain elements called 0 and 1 that act as additive and multiplicative identity is, i.e. $0 + x = x$ and $1 \cdot x = x$ for all $x \in F$. Also, additive and multiplicative inverses are assumed to exist: if $x \in F$, there is always some element $-x$ with $x + (-x) = 0$, and unless $x = 0$, there is an element $\frac{1}{x}$ with $x \cdot \frac{1}{x} = 1$. The existence of such inverses allows us to define subtraction and division in the usual way. Examples include the rational numbers $\mathbb{Q}$, the real numbers $\mathbb{R}$, and the complex numbers $\mathbb{C}$.

EXAMPLE 3.3.15. Let p be a prime number and set $F_p = \{0, \ldots, p-1\}$, considered with the operations of addition and multiplication modulo p.

We claim that F_p is a field. Most of the properties are easy to check; the only non-trivial one, and the only one that requires primality, is the existence of multiplicative inverses. So, if $x \in F_p$ is nonzero, we want some $y \in F_p$ with

$$xy = 1 \pmod{p}, \iff xy - pn = 1 \text{ for some } n \in \mathbb{Z}.$$

By the Euclidean algorithm, this can be solved exactly when the greatest common divisor of x and p is 1, which is always the case if p is prime and $x \in F_p$.

If F is a field and $n \in \mathbb{N}$, let F^n the set of all n-tuples $(x_1, \ldots, x_n)$ of elements of F. Elements of F^n can be added, and multiplied by elements of F:

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n),$$
$$t(x_1, \ldots, x_n) = (tx_1, \ldots, tx_n).$$

We can make F^n into an incidence geometry by declaring *lines* to be subsets

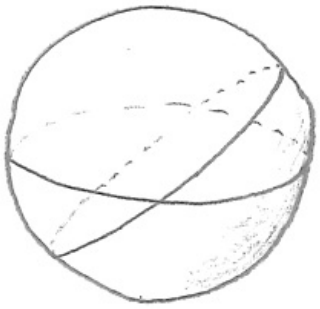
$$\ell = \{tx + b \mid t \in F\}, \quad x, b \in F^n.$$

EXERCISE 3.3.16. Check that F^n satisfies I1-I3, and that F^n is an affine plane if and only if $n = 2$.

EXERCISE 3.3.17. Show that there exists a field with four elements. *Call the elements* $0, 1, a, b$. *The ways* $0, 1$ *behave under addition and multiplication are determined, so you just have to define* $a + a, b + b, a + b, a \cdot b, a \cdot a, a \cdot b$.

By Example 3.3.15 and Exercise 3.3.17, there are fields with $2, 3, 4$ and 5 elements. So, there are affine planes with $4, 9, 16$ and 25 elements. In fact, the pattern stops at 25, and there is no affine plane with 36 elements. The proof of this is difficult and is due to Euler.

Suppose that S is the unit sphere. If we view the great circles on S as 'lines', do we get an incidence geometry?



The answer is no, since if $a, b \in S$ are antipodal then there are infinitely many great circles containing both a and b, so axiom I1 fails. However, this deficiency can be circumvented by 'identifying' antipodal points. Namely, the *projective plane* is

$$P = \big\{\, \{p, -p\} \;\big|\; p \in S \,\big\},$$

which we make into an incidence geometry by declaring that for every great circle C on S, the following subset is a line in the projective plane:

$$\big\{\, \{p, -p\} \;\big|\; p \in C \,\big\} \subset P.$$

This is an incidence geometry: it satisfies I1 since if $\{p, -p\}$ and $\{q, -q\}$ are not equal, then p, q are not antipodal, so the unique great circle containing p, q them gives the unique line in P through $\{p, -p\}$ and $\{q, -q\}$. Lines in the projective plane are infinite, so I2 is satisfied, and if p, q, r are points of S that do not lie on a great circle, then the corresponding points of P are not collinear.

Recall that any two great circles on S intersect. This property is inherited by the projective plane, which motivates the following definition.

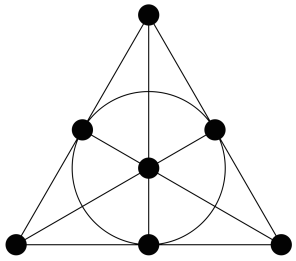
Definition 3.3.18. An incidence geometry is called a *generalized projective plane* if any two lines intersect and every line contains at least three points.

Note that the condition that every line contains at least three points excludes the geometries depicted in Exercise 3.3.5.

Exercise 3.3.19. Let $\mathbb{R}\mathrm{P}^n$ be the set of all lines through the origin in $\mathbb{R}^{n+1}$. We will call a straight line through the origin a POINT, using capital letters to distinguish it from a usual point in $\mathbb{R}^{n+1}$. A plane through the origin in $\mathbb{R}^{n+1}$ gives rise to a LINE in $\mathbb{R}\mathrm{P}^n$, consisting of all POINTS that are contained in that plane.

(a) Show that $\mathbb{R}\mathrm{P}^n$ is a generalized projective plane.
(b) Show that $\mathbb{R}\mathrm{P}^2$ is isomorphic to the projective plane P described above.
(c) If F is a field, let FP^n be the set of lines through the origin in F^{n+1}, as described before Exercise 3.3.16. Generalizing the case $F = \mathbb{R}$, define a set of *lines* in FP^n that gives it the structure of a generalized projective plane.
(d) Show that FP^n has $(p^n - 1)/(p - 1)$ elements.

EXERCISE 3.3.20. The following picture describes the *Fano plane*: its lines are the seven 3-element subsets represented by the straight-line segments and the circle.



(a) Show that the Fano plane is a generalized projective plane.
(b) Show that any generalized projective plane has at least 7 points.
(c) Prove that any generalized projective plane that has exactly 7 points is isomorphic to the Fano plane.
(d) With Exercise 3.3.19 (c) in mind, show that the Fano plane is isomorphic to F_2P^2, the set of lines through the origin in F_2^3.

EXERCISE 3.3.21 (Gambling!). Let's suppose that ping-pong balls labeled with the numbers 1-7 are circulating around in a box attached to a powerful fan. Players buy \$.50 lottery tickets labeled with three numbers between 1 and 7 – the order in which the three numbers is listed does not matter. The game show host draws three ping-pong balls from the box, and a player receives \$2 whenever two of the drawn numbers appear on his ticket, and \$6 if the drawn numbers all appear on his ticket.

Show that if a player labels the vertices of the Fano plane with the numbers 1-7, and buys seven tickets corresponding to the lines in the Fano plane, he or she is guaranteed exactly a \$2.50 net gain when the three lottery numbers are drawn.

As an aside, if you know some probability, you can show that \$2.50 is the expected net gain if you purchase 7 random lottery tickets. However, if you are a conservative player, or are playing on someone else's dime, you might not want to leave things to chance, in which case the Fano plane strategy is advantageous. The property that makes the Fano plane desirable here is that it is a Steiner system*; these may have been used by a group of MIT students who made millions over the years taking advantage of a Massachusetts state lottery with an unusually positive expected payout. See Ellenberg's book "How not to be wrong: the power of mathematical thinking."*

EXERCISE 3.3.22. Suppose that P is a generalized projective plane and ℓ is a line in P. Define a notion of 'line' in $P\backslash\ell$ that makes $P\backslash\ell$ into an affine plane.

3.4. Inversions

In order to say anything useful about the hyperbolic plane, we'll need to better understand lircles in $\mathbb{R}^2$. The crucial ingredient is a transformation of the plane called an *inversion*, which one can think of as a 'reflection' through a circle.

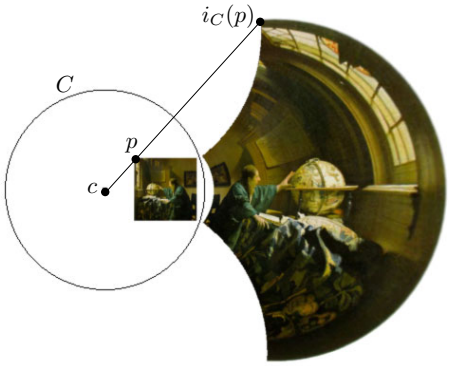
DEFINITION 3.4.1. If C is a circle centered at $q \in \mathbb{R}^2$, the *inversion* through C is

$$i_C : \mathbb{R}^2 \backslash \{q\} \longrightarrow \mathbb{R}^2 \backslash \{q\},$$

where if C has radius r, we define $i_C(p)$ to be the point on the ray from q in the direction of p such that $|i_C(p) - q||p - q| = r^2$. In coordinates,

$$i_C(p) = \frac{r^2}{|p-q|^2}(p-q) + q.$$

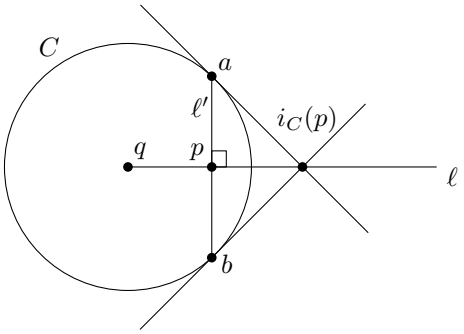
To derive the last formula, note that the vector $\frac{r^2}{|p-q|^2}(p-q)$ has length $r^2/|p-q|$ and points in the same direction as $p - q$, so adding it to q gives $i_C(p)$ as desired. Here is the effect of inverting Vermeer's *The astronomer* through a circle C.



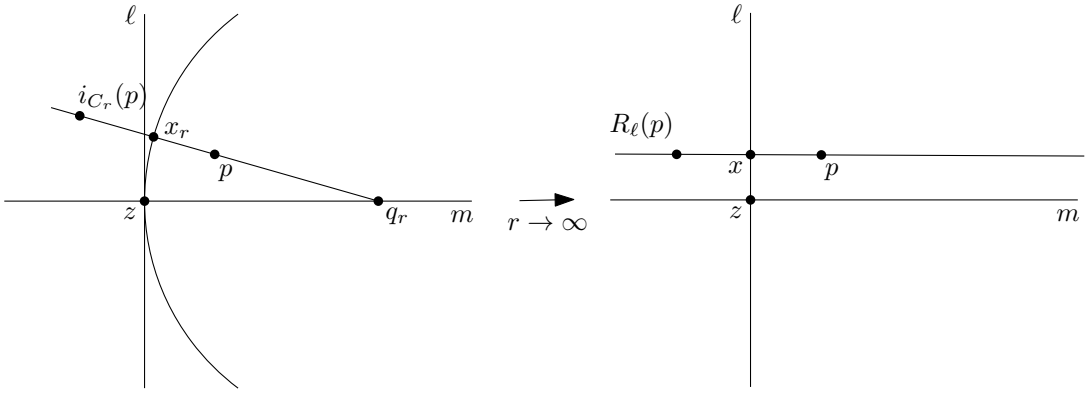
Note that $i_C \circ i_C(p) = p$ for all p, as the equation $|i_C(p)-q||p-q| = r^2$ is symmetric in p and $i_C(p)$. If $p \in C$, it follows from the definition that $i_C(p) = p$. Although $i_C(q)$ is not defined, one should imagine that $i_C(q) = \infty$ and $i_C(\infty) = q$.

Here is a geometric way to construct $i_C(p)$ from p.

EXERCISE 3.4.2. Suppose that C is a circle with center q and $p \in \mathbb{R}^2 \backslash \{q\}$ be a point that lies inside C. Let ℓ be the line through q, p and let ℓ' be the line perpendicular to ℓ through p. Suppose that ℓ' intersects C at a and b. Show that the point d on ℓ where the lines through a and b tangent to C intersect is equal to $i_C(p)$.



In §3.3, we saw that lines can be considered as circles centered at infinity. From this perspective, one can consider a reflection through a line as a degenerate version of a inversion in a circle. Namely, let ℓ be a line and let m be a line that intersectsℓ perpendicularly at a point z. Given $r > 0$, let C_r be a circle with radius r that passes through z and whose center q_r lies on m, as on the left in the following picture.



As $r \to \infty$, the circle C_r limits onto the line ℓ. We claim that for any $p \in \mathbb{R}^2$,

$$\lim_{r\to\infty} i_{C_r}(p) = R_\ell(p),$$

where R_ℓ is the reflection through ℓ. So in other words, as C_r approaches ℓ, the inversion through C_r approaches the reflection through ℓ. To see this, note first that as $r \to \infty$, the line through p, q_r limits onto the line through p that is parallel to m, as in the right side of the picture above. Let x_r be the point at which the line through p, q_r hits C_r. As $r \to \infty$, the point x_r limits to the closest point $x \in \ell$ to p. Moreover,

$$d(i_{C_r}(p), q_r) = r^2/d(p, q_r), \implies d(i_C(p), x_r) = \frac{r^2}{r - d(x_r, p)} - r,$$

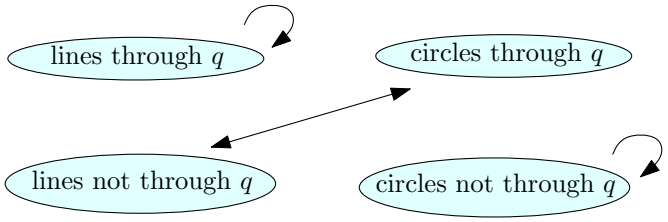
so as $r \to \infty$, we have

$$\lim_{r\to\infty} \frac{r^2}{r - d(x_r, p)} - r = \lim_{r\to\infty} \frac{r \cdot d(x_r, p)}{r - d(x_r, p)} = d(x, p)$$

where the last equality is L'Hôpital's rule, using that $x_r \to x$. So, as $r \to \infty$ we have that $d(i_{C_r}(p), x_r)) \to d(x, p)$. This implies that $i_{C_r}(p) \to R_\ell(p)$ as desired.

The above discussion shows that inversions through circles are analogous to reflections through lines. Reflecting a circle through a line gives a circle, and the reflection of a line through a line is another line. Here is the analogous statement for inversions.

THEOREM 3.4.3 (Inversions send lircles to lircles). *If C is a circle and C' is a lircles, then $i_C(C')$ is also a lircle. More specifically, if q is the center of C then i_C sends lines through q to lines through q, circles not through q to circles not through q, circles through q to lines not through q, and lines not through q to circles through q.*



PROOF. We saw in §3.1 that lircles are solution sets of nondegenerate equations

$$ax^2 + ay^2 + bx + cy + d = 0. \tag{10}$$

So, to show that inversions send lircles to lircles, we just have to check that an equation of the form in (10) becomes another such equation when (x, y) is replaced by $i_C(x, y)$. Suppose for convenience that C is a circle centered at the origin. If C has radius r, the inversion through C can be written as

$$i_C(x, y) = \frac{r^2}{|(x, y)|^2}(x, y) = \left(\frac{r^2x}{x^2 + y^2}, \frac{r^2y}{x^2 + y^2}\right).$$

So, plugging in the output into Equation (10) gives

$$a\left(\frac{r^2x}{x^2 + y^2}\right)^2 + a\left(\frac{r^2y}{x^2 + y^2}\right)^2 + b\left(\frac{r^2x}{x^2 + y^2}\right) + c\left(\frac{r^2y}{x^2 + y^2}\right) + d = 0.$$

The numerators of the first two terms combined to give a $x^2 + y^2$, which cancels part of the $(x^2 + y^2)^2$ denominator. So, after simplification this becomes

$$ar^4 + br^2x + cr^2y + d(x^2 + y^2) = 0, \tag{11}$$

which is a quadratic equation exactly of the form in Equation (10), except that the coefficients have been rearranged and modified. Since the original equation (10) is nondegenerate, its solution set C' is a proper subset of $\mathbb{R}^2$ with at least 2 points, so the same is true of $i_C(C')$, which is the solution set of (11), so the latter equation is nondegenerate. Hence, i_C takes lircles to lircles.

The permutation of the four types of lircles described in the figure can be easily checked. As an example, Equation (10) describes a line that *does not* pass through 0 when $a = 0$ and $d \neq 0$. In this case, Equation (11) describes a circle (since the squared terms have nonzero coefficients) that goes through zero (as there is no constant term). The other cases are checked similarly.

If C is centered instead at some $q \neq 0$, let C' be a circle with the same radius centered at the origin. Then $i_C = T_q \circ i_{C'} \circ T_{-q}$, so i_C takes lircles to lircles as both $i_{C'}$ and the two translations do. Moreover, the two translations convert between lircles passing through q and lircles passing through the origin, so the description of the permutation of the four types of lircles follows for i_C from the description for $i_{C'}$. □

If $\alpha : [a, b] \longrightarrow \mathbb{R}^2$, $\alpha(t) = (\alpha_1(t), \alpha_2(t))$ is a differentiable path, recall that the *velocity vector* of α at time t is the vector $\alpha'(t) := (\alpha_1'(t), \alpha_2'(t))$ of derivatives of the

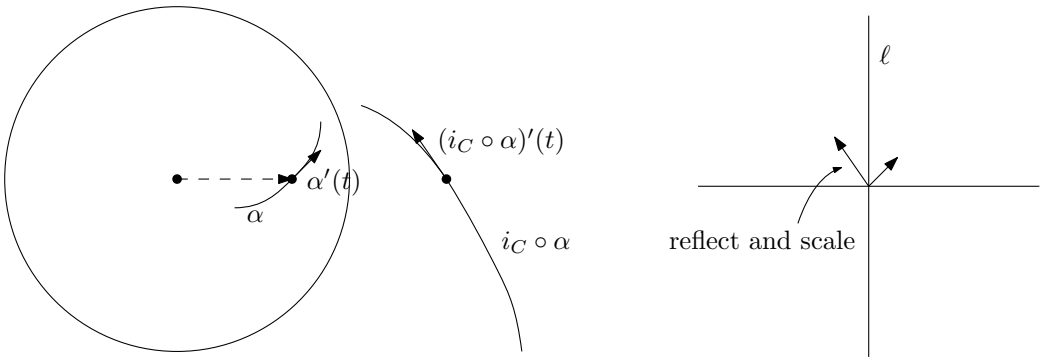
coordinates of α. The *speed* of α at time t is the length $|\alpha'(t)|$ of the velocity vector. The following theorem compares the velocity of a path α to the velocity of its image $i_C \circ \alpha$ under an inversion.

THEOREM 3.4.4. *Suppose C is a circle in $\mathbb{R}^2$ with center q and radius r, and that $\alpha : [a, b] \longrightarrow \mathbb{R}^2 \backslash \{q\}$ is a path. Then for each t, the velocity vector*

$$(i_C \circ \alpha)'(t) = \frac{r^2}{|\alpha(t) - q|^2} R_\ell(\alpha'(t)),$$

where ℓ is the line through the origin perpendicular to the vector $\alpha(t) - q$.

Here, remember that $\alpha'(t)$ and $(i_C \circ \alpha)'(t)$ are regarded as vectors. In the equation above, both vectors should be considered based at the origin. See the picture below, where the two vectors are drawn on the left as velocity vectors, and then on the right they are translated to be based at the origin. In the picture, $\alpha(t) - q$ is the dotted vector on the left, which is horizontal, so ℓ is the y-axis. Note that the scaling factor $r^2/|\alpha(t) - q|^2$ is bigger than one when $\alpha(t)$ lies inside the circle C, and is less than 1 when $\alpha(t)$ lies outside C. So in the picture, $\alpha'(t)$ is reflected over the vertical axis, and then scaled up by $r^2/|\alpha(t) - q|^2$.



We'll prove Theorem 3.4.4 rigorously in a moment, but first let's check that it makes sense in a couple examples. For simplicty, suppose C is a circle of radius r centered at the origin. If α is the path $\alpha(t) = (t, 0)$, then $i_C(\alpha(t)) = (r^2/t, 0)$, so

$$(i_C \circ \alpha)'(t) = (-\frac{r^2}{t^2}, 0) = \frac{r^2}{t^2} \cdot (-1, 0) = \frac{r^2}{|(t, 0) - (0, 0)|^2} \cdot R_\ell(\alpha'(t)),$$

where ℓ is the y-axis, which is perpendicular to $\alpha(t) - (0, 0) = (t, 0)$ as required. For another example, suppose that $\beta(t) : [0, 1] \longrightarrow \mathbb{R}^2$ is a constant speed parametrization of a radius s circle around the origin. Then $i_C \circ \beta$ is a constant speed parametrization of a circle of radius r^2/s around the origin, so

$$(i_C \circ \beta)'(t) = \frac{r^2/s}{s} \beta'(t),$$

i.e. the velocity is scaled by the ratio of the two radii. Here, β is perpendicular to the vector $\beta(t) - (0,0)$, so reflection through the line ℓ perpendicular to $\beta(t)$ fixes $\beta'(t)$.

PROOF. Translating, we may assume that $q = 0$, in which case the inversion can be written as $i_C(p) = r^2 \frac{p}{|p|^2}$, and $d(q, \alpha(t)) = |\alpha(t)|$.

$$(i_C \circ \alpha)'(t) = \frac{d}{dt}\, r^2 \frac{\alpha(t)}{|\alpha(t)|^2} = r^2\, \frac{\alpha'(t)\,|\alpha(t)|^2 - \alpha(t)\,\frac{d}{dt}|\alpha(t)|^2}{|\alpha(t)|^4}.$$

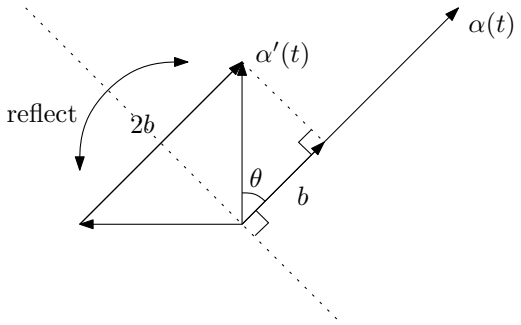
The mysterious term in the quotient on the right is

$$\begin{aligned}\frac{d}{dt}|\alpha(t)|^2 &= \frac{d}{dt}\alpha_1(t)^2 + \alpha_2(t)^2 \\ &= 2\alpha_1(t)\alpha_1'(t) + 2\alpha_2(t)\alpha_2'(t) \\ &= 2\,\alpha(t)\cdot\alpha'(t),\end{aligned}$$

which is no longer so mysterious. Plugging it in,

$$(i_C \circ \alpha)'(t) = \frac{r^2}{|\alpha(t)|^2}\left(\alpha'(t) - 2\alpha(t)\,\frac{\alpha(t)\cdot\alpha'(t)}{|\alpha(t)|^2}\right).$$

Interpreting everything as vectors based at the origin, the vector $b = \alpha(t)\,\frac{\alpha(t)\cdot\alpha'(t)}{|\alpha(t)|^2}$ is the *projection* of $\alpha'(t)$ onto $\alpha(t)$, as pictured below. So, $\alpha'(t) - 2b$ is the reflection of $\alpha'(t)$ over the line perpendicular to the vector $\alpha(t)$.



This means that $(i_C \circ \alpha)'(t)$ is obtained from $\alpha'(t)$ by first reflecting over the line perpendicular to $\alpha(t)$, then scaling by the factor $\frac{r^2}{|\alpha(t)|^2}$, as desired. □

As a consequence of Theorem 3.4.4, we have:

COROLLARY 3.4.5. *Suppose C is a circle in $\mathbb{R}^2$ with center q. Then i_C is* conformal*, or angle preserving: if two paths α and β intersect at an angle θ at $p \neq q$, then the paths $i_C \circ \alpha$ and $i_C \circ \beta$ meet with angle θ at $i_C(p)$.*

PROOF. Theorem 3.4.4 says that $(i_C \circ \alpha)'(t)$ and $(i_C \circ \beta)'(t)$ are obtained from $\alpha'(t)$ and $\beta'(t)$ by reflecting over the line perpendicular to $p - q$ and scaling by $r^2/|p-q|^2$. Reflecting and scaling vectors does not change the angle between them. □

Applying this corollary with one of the paths a parametrization of C, we see that if ℓ is a lircle that intersects C orthogonally, then $i_C(\ell)$ also intersects C orthogonally, at the same points. In fact, we can show:

THEOREM 3.4.6. *Suppose C is a circle in $\mathbb{R}^2$, and ℓ is a lircle. Then ℓ is orthogonal to C if and only if $i_C(\ell) = \ell$. Moreover, ℓ is orthogonal to C whenever ℓ contains a single point $p \notin C$ and its inversion $i_C(p)$.*

To prove this, we need a brief lemma.

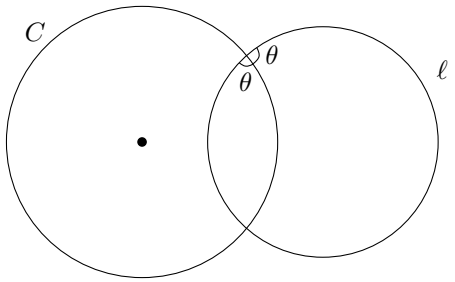
LEMMA 3.4.7. *If ℓ is a lircle and $x, y \in \ell$, there is a unique lircle ℓ' in $\mathbb{R}^2$ such that ℓ' intersects ℓ orthogonally, exactly at the points x, y.*

PROOF. Let's first prove the lemma when ℓ is a line. If $x = y$, then the perpendicular line to ℓ at $x = y$ is the unique such ℓ'. If $x \neq y$, the unique such ℓ' is the circle centered at the midpoint of the segment $xy \subset \ell$.

Now suppose ℓ is a circle. Take a circle C centered at some point q that lies on ℓ. Then Theorem 3.5.3 says that $i_C(\ell)$ is a line in $\mathbb{R}^2$. The points $i_C(x), i_C(y)$ lie on $i_C(\ell)$, so by the previous case, there is a unique lircle m that goes through $i_C(x), i_C(y)$ and intersects $i_C(\ell)$ orthogonally. Inverting back and using Corollary 3.4.5, $i_C(m)$ is the unique lircle intersects ℓ orthogonally in the points x, y. □

PROOF OF THEOREM 3.4.6. Suppose ℓ intersects C orthogonally at x, y. As noted above, $i_C(\ell)$ also intersects C orthogonally at x, y, and Theorem 3.5.3 says that $i_C(\ell)$ is a lircle, so Lemma 3.4.7 says that $i_C(\ell) = \ell$.

For the reverse direction, suppose ℓ contains $p \notin C$ and $i_C(p)$. If we pick some $x \in C$, then the lircles ℓ and $i_C(\ell)$ both contain the three points $p, x, i_C(p)$. Hence, $i_C(\ell) = \ell$ by Exercise 3.2.3. Corollary 3.4.5 then says that the two angles θ below are equal, and therefore are $\pi/2$. So, ℓ is orthogonal to C.



□

COROLLARY 3.4.8. *If C is a circle, the inversion i_C is the unique continuous map on its domain of definition, other than the identity, such that $i_C(\ell) = \ell$ whenever ℓ is a lircle orthogonal to C.*

PROOF. Suppose that q is the center of C and $f : \mathbb{R}^2\backslash\{q\} \longrightarrow \mathbb{R}^2\backslash\{q\}$ is a map such that $i_C(\ell) = \ell$ whenever ℓ is a lircle orthogonal to C.

Pick some $p \in \mathbb{R}^2 \backslash \{q\}$. We claim that either $f(p) = p$ or $f(p) = i_C(p)$. If $p \in \ell$, let ℓ, m be two lircles that intersect C orthogonally at p. Then ℓ, m are tangent to each other at p, so $\ell \cap m = \{p\}$, and since $f(\ell) = \ell$ and $f(m) = m$, we must have $f(p) = p$ as desired. So, we may assume $p \notin C$. Pick a point $x \in C$, let ℓ be the lircles through $p, i_C(p), x$ given by Exercise 3.2.3. Then pick $y \in \ell$ with $y \notin C$, and let m be the lircle through $p, i_C(p), y$, so that $\ell \neq m$. By Theorem 3.4.6, both ℓ and m are orthogonal to C, so we have $f(\ell) = \ell$ and $f(m) = m$. So, $f(p) \in \ell \cap m = \{p, i_C(p)\}$.

Finally, by continuity, either $f(p) = p$ for all p or $f(p) = i_C(p)$ for all p. □

3.4.1. Exercises.

EXERCISE 3.4.9. Suppose that C and C' are two circles with the same center q and radii r, r'. Show that the composition $i_{C'} \circ i_C$ is the dilation $D_{q,\lambda}$ around q by a factor of λ, where $\lambda = (\frac{r'}{r})^2$. (See the end of Section 1.2.)

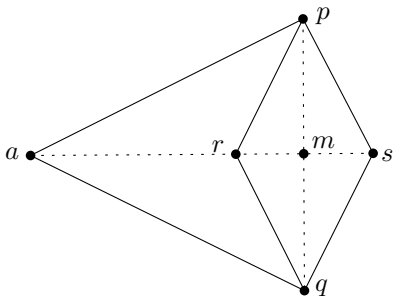
EXERCISE 3.4.10. If ℓ is a line through a point p and $q \neq p$, show that there is a unique lircle C that is tangent to ℓ at p and passes through q.

The following exercise should not be a surprise if you remember the geometric interpretation of conjugation given in Section 1.2. In it, let's interpret the 'inversion' through a line to mean the reflection through that line.

EXERCISE 3.4.11. If C, E are two lircles in $\mathbb{R}^2$, then $i_C \circ i_E \circ i_C = i_{i_C(E)}$. *Hint: you may find Corollary 3.4.8 and its analogue for reflections useful.*

A hot pursuit in the first half of the 19th century was to construct machines to convert between rotational motion and linear motion. This interest should make sense if you think about the relationship between a piston and a train wheel. The goal was to perform this transformation with a 'mechanical linkage' consisting of metal rods connected together at rotating joints.

The picture below illustrates a linkage formed from 6 metal rods connected at the joints a, r, p, q, s. Imagine that the position of a is forever anchored. The position of r then determines the positions of the rest of the joints, and we consider the path that s makes as we move r around. The way the picture is drawn is supposed to indicate the following conditions: $ap = aq$ and $pr = qr = ps = sq$, where for brevity these are the lengths of the rods with the indicated endpoints.



EXERCISE 3.4.12. Show that s is the inversion of r through a circle with center a and radius $\sqrt{ap^2 - pr^2}$.

Therefore, one can construct a 'machine' that performs inversion in a circle. For fun, you might try constructing one of these out of sticks.

EXERCISE 3.4.13. Augment the linkage above with an additional bar and anchor to create a new linkage in which r is constrained to circular motion, and s is constrained to linear motion. Such a linkage was first invented in 1864 by Charles-Nicolas Peaucellier, a captain in the French army.

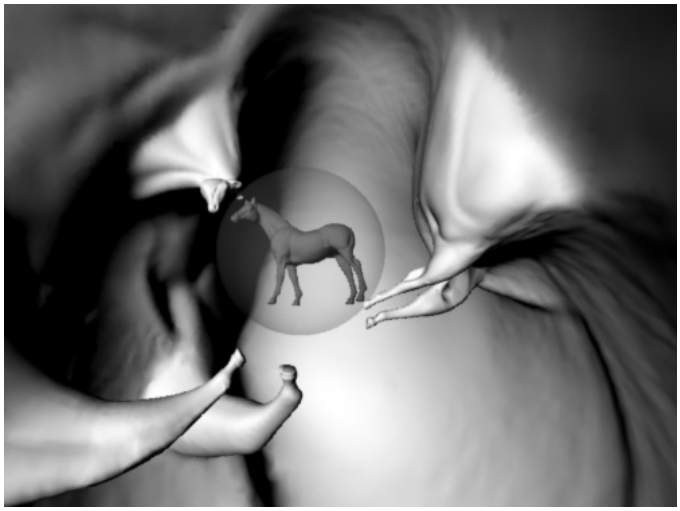
Inversions can be defined in $\mathbb{R}^n$, for any n. If

$$C = \{p \in \mathbb{R}^n \mid |p - q| = r\}$$

is a sphere in $\mathbb{R}^n$ with center q and radius r, the inversion through C is again

$$i_C(p) = \frac{r^2}{|p - q|^2}(p - q) + q.$$

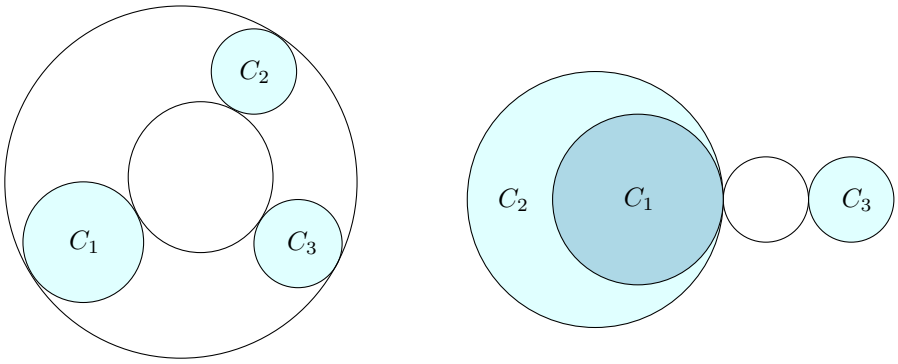
The geometric interpretation is the same: $i_C(p)$ is just the point on the ray through q, p whose distance to q is $r^2/d(p, q)$. Here is a picture of an inversion in $\mathbb{R}^3$.



In the picture, the center q of the sphere is inside the small horse pictured. The inversion takes the skin of the horse to the surface shown that is enclosing the camera outside of the camera, which separates everything shown from the interior of the horse, which extends off to infinity.

EXERCISE 3.4.14. Show that stereographic projection $\pi : S\backslash\{n\} \longrightarrow \mathbb{R}^2$ of the unit sphere $S \subset \mathbb{R}^3$ onto $\mathbb{R}^2$ is the restriction of an inversion i_C through some sphere $C \subset \mathbb{R}^3$. *So, the fact that stereographic projection is conformal and sends circles to lircles parallels the corresponding properties for inversions, which it turns out also are true in higher dimensions.*

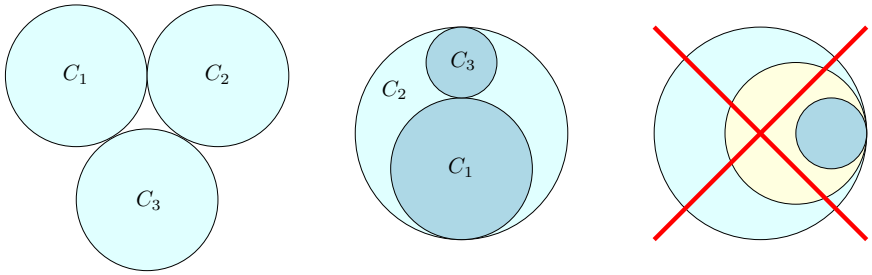
Suppose C_1, C_2, C_3 are circles in the plane. An *Apollonian circle* is a circle C that is tangent (but not equal) to all three of C_1, C_2, C_3.



The existence of such circles was of great interest to the ancient Greeks, and they are named after the Greek mathematician Apollonius of Perga.

EXERCISE 3.4.15. Give an example of circles C_1, C_2, C_3 for which there are no associated Apollonian circles. Then give an example where there are infinitely many.

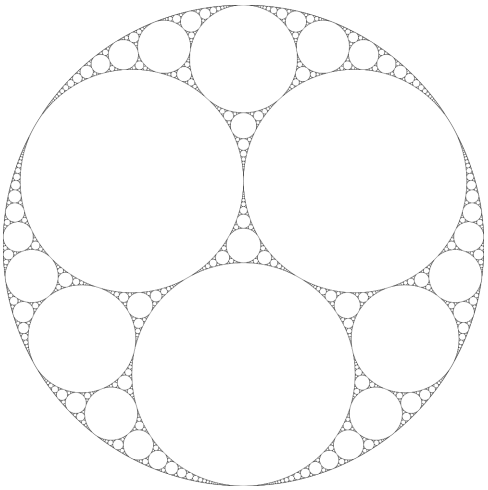
EXERCISE 3.4.16. Suppose now that C_1, C_2, C_3 are all tangent, but not all at the same point, as pictured below. Show that there are exactly two Apollonian circles D_1, D_2 associated to C_1, C_2, C_3, and that each D_i is tangent to the circles C_1, C_2, C_3 at three different points. *Hint: find an inversion that takes two of the circles to parallel lines.*



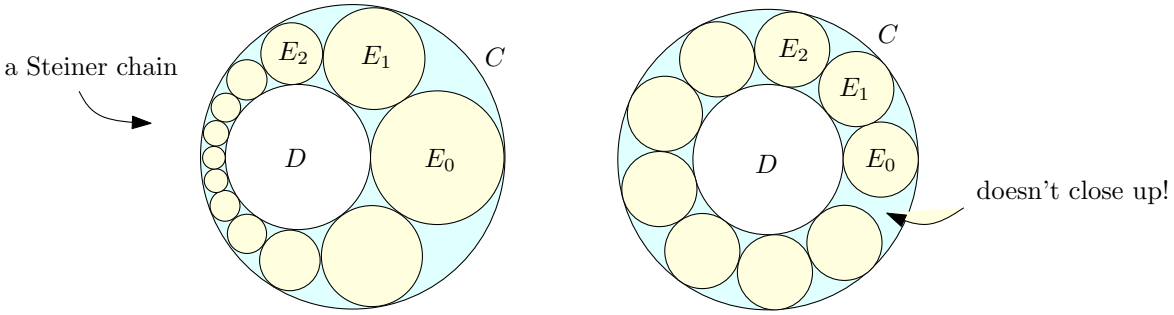
There are some amazing fractals that can be generated using this result. Starting with the circles C_1, C_2, C_3 above, let D_1, D_2 be the associated Apollonian circles. Then we have six new triples of mutually tangent circles:

$$D_1, C_1, C_2, \quad D_2, C_1, C_2, \quad D_1, C_2, C_3, \quad D_2, C_2, C_3, \quad D_1, C_1, C_3, \quad D_2, C_1, C_3.$$

Each of these has two associated Apollonian circles. Continue this process, drawing the new Apollonian circles every time a new triple of mutually tangent circles is created. The resulting fractal is called an *Apollonian gasket*. Here is an example.



Here is a related problem. Suppose we have two non-intersecting circles C and D, with D contained inside C. Start with a circle E_0 that is tangent to both, and is contained within C but does not contain D. We let E_1 be one of the two Apollonian circles tangent to C, D, E_0, and inductively define E_{i+1} to be the circle tangent to C, D, E_i that is not E_{i-1}. If after one revolution around D, the chain closes up with some $E_n = E_0$, we say that (E_i) is a *Steiner chain*.



EXERCISE 3.4.17. Assume the circles D and C are concentric with radii r and s, respectively. Show that a chain (E_i) as above closes up with $E_n = E_0$ if and only if

$$(s-r)^2 = 2\left(\frac{s+r}{2}\right)^2\left(1-\cos\left(\frac{2\pi}{n}\right)\right).$$

Hint: this is exactly the law of cosines for an appropriate triangle.

In particular, for concentric C, D either you get a Steiner chain for *every* starting circle E_0, or you never do. This shouldn't be such a surprise, since in this case a Steiner chain can be rotated to start at any circle tangent to C and D desired. Surprisingly, the same duality persists when C and D are not concentric!

THEOREM 3.4.18 (Steiner's porism). *Suppose $C, D \subset \mathbb{R}^2$ are circles and D is contained inside C. If there is a single Steiner chain of circles as above,* any *circle E_0 that lies between C and D and is tangent to both is part of a Steiner chain.*

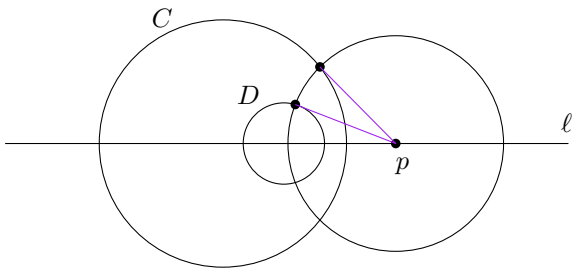
The proof relies on the following lemma:

LEMMA 3.4.19. *If C, D are non-intersecting circles in $\mathbb{R}^2$, there is some circle $S \subset \mathbb{R}^2$ such that $i_S(C)$ and $i_S(D)$ are concentric circles.*

Assuming the lemma, since i_S maps circles to circles, any Steiner chain for C, D maps under i_S to a Steiner chain for $i_S(C)$ and $i_S(D)$, and vice versa. So as the theorem is true for concentric circles, it must also be true for the circles C, D.

So, let's prove the lemma.

EXERCISE 3.4.20. Let C, D be non-intersecting circles, and ℓ be a line through their centers. Show that there is a circle E centered on ℓ that is orthogonal to C, D. *Hint: there is a circle centered at $p \in \ell$ that is orthogonal to C, D exactly when the line segments joining p to points of tangency on C, D have the same length. Sliding p along ℓ, argue using continuity that such a p must exist.*

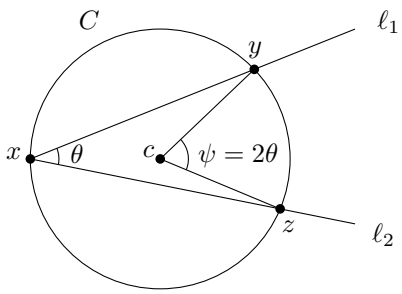


EXERCISE 3.4.21. Prove the lemma, using the previous exercise. *Hint: invert in a circle centered at a point of intersection of E and ℓ.*

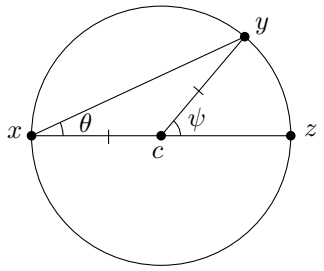
3.5. A alternative geometric approach to circle inversion

We'll discuss here a sequence of elementary theorems in Euclidean geometry that lead up to an alternative proof of Theorem 3.5.3. Here's the first result.

The inscribed angle theorem. *Suppose that two lines ℓ_1 and ℓ_2 intersect a circle C at x, y and x, z, respectively. Then their angle θ of intersection is half the angle ψ made by the radii connecting the center c of C to y and z.*

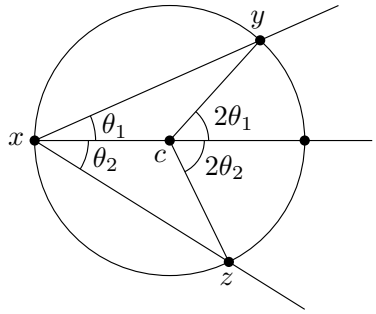


PROOF. As a warm-up, let's first prove the theorem when x, c, z are collinear.



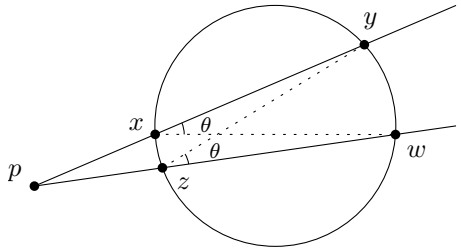
In that case, the triangle x, c, z is isosceles, so the angles at x and z are equal, implying that the third angle is $\pi - 2\theta$. However, then $\psi = \pi - (\pi - 2\theta) = 2\theta$.

In the general case, add in to the picture the line through x and c.



Two applications of the warm-up prove the theorem. □

The secant theorem. *Suppose that two lines intersect at a point p and intersect a circle at x, y and z, w, respectively. Then $d(p, x)d(p, y) = d(p, z)d(p, w)$.*



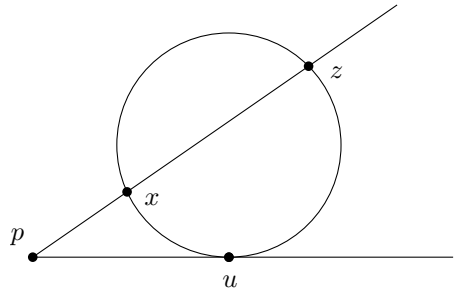
PROOF. The angles $\angle wxy$ at x is the same as the angle $\angle wzy$ at z, by the inscribed angle theorem, so we will call that angle θ. The triangles pyz and pxw then both have angles $\pi - \theta, \psi, \theta - \psi$, where ψ is the angle at p, so are similar. Thus,

$$\frac{d(p,w)}{d(p,x)} = \frac{d(p,y)}{d(p,z)},$$

which proves the theorem. □

COROLLARY 3.5.1 (Secant-tangent theorem). *Suppose that two lines ℓ, ℓ' intersect at p, and that ℓ intersects a circle C at x, y, while ℓ' intersects C at a point u, and possibly elsewhere. Then ℓ' is tangent to C if and only if*

$$d(p,x)d(p,y) = d(p,u)^2.$$



PROOF. If ℓ' is tangent to C at u, this a limit of the secant theorem: if we consider lines ℓ'' that intersect C at points z, w approaching u, then $d(p,z)d(p,w) \longrightarrow d(p,u)^2$.

For the converse, if there is another point v at which ℓ' intersects C, then $d(p,u) \neq d(p,v)$ and $d(p,u)d(p,v) = d(p,x)d(p,y)$, so $d(p,x)d(p,y) \neq d(p,u)^2$. □

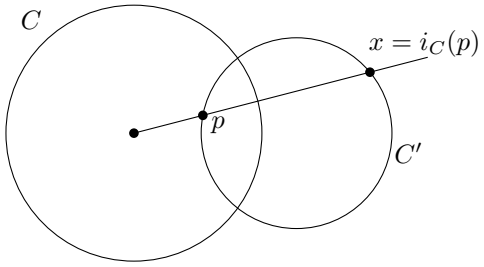
To show geometrically that inversions take lircles to lircles, we first show that lircles orthogonal to the circle of inversion are taken to themselves.

THEOREM 3.5.2. *Suppose that C is a circle in $\mathbb{R}^2$ and C' is a lircle.*

(a) If C' is orthogonal to C, then $i_C(C') = C'$.

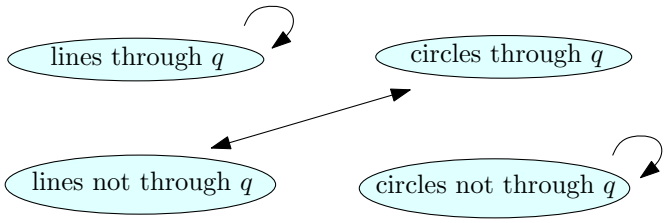
(b) Conversely, if C' contains both a single point $p \notin C$ and its inversion $i_C(p)$, then C, C' are orthogonal.

PROOF. If C' is a circle that intersects C orthogonally at u, then by Corollary 3.5.1, for any $p \in C'$ we have $d(c,p)d(c,x) = r^2$, where x is the other point of C lying on the line through c, p. So, $x = i_C(p)$ by definition of an inversion.



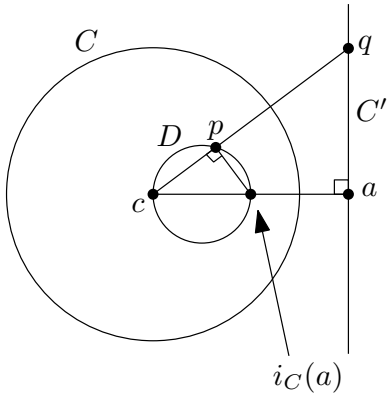
Conversely, suppose that C' is a circle that contains both a point p and its inversion $x = i_C(p)$. Let u be an intersection point of C and C', and let ℓ be the line containing c, u. By definition of inversion, $d(c,p)d(c,x) = r^2 = d(c,u)^2$. Corollary 3.5.1 shows that ℓ is tangent to C', implying that C, C' are orthogonal. □

THEOREM 3.5.3. *Inversions send lircles to lircles: if C, C' are lircles, so is $i_C(C')$. More specifically, if c is the center of C then we have:*



PROOF. We have already seen that lines through c invert to lines through c. This leaves two cases: we must show that inversions interchange lines not through c with circles through c, and that inversion sends circles not through c to circles not through c.

Case 1. (Lines not through c ↔ circles through c) Let C' be a line not through c and suppose that $a \in C'$ is the point such that ca is perpendicular to C'. Let D be the circle through c that is centered at the midpoint of the segment from c to $i_C(a)$.

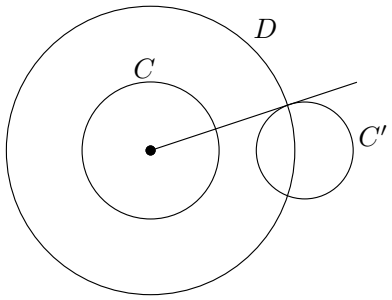


We claim that $i_C(C') = D$. So, suppose that c, p, q are collinear and p lies on D while q lies on C. The right triangles $c\,p\,i_C(a)$ and $c\,a\,q$ share an angle at c, so they are similar. Therefore,

$$\frac{d(c,p)}{d(c,i_C(a))} = \frac{d(c,a)}{d(c,q)} \implies d(c,p)d(c,q) = d(c,a)d(c,i_C(a)) = r^2,$$

where r is the radius of C. This shows that $i_C(q) = p$, so $i_C(C') = D$. It follows as well that $i_C(D) = C'$.

Case 1. (circles not through c $\leftrightarrow$ circles not through c) Suppose that C' is a circle that does not go through c. Then there is a circle D with center c that is orthogonal to C'. As $i_D(C') = C'$, we have $i_C(C') = i_C(i_D(C')) = (i_C \circ i_D)(C')$.

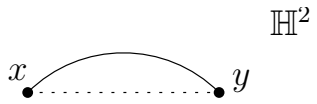


But as $i_C \circ i_D$ is a dilation around c, $(i_C \circ i_D)(C')$ is a circle. It doesn't pass through c since a dilation around c fixes c and C' does not pass through c. □

3.6. The hyperbolic metric

In Section 3.1, we defined the *hyperbolic plane* $\mathbb{H}^2$ as the open upper half plane in $\mathbb{R}^2$, and *hyperbolic lines* to be vertical half lines and semicircles orthogonal to the x-axis. We'll show in this section that there a metric on $\mathbb{H}^2$ such that the shortest paths between points lie along hyperbolic lines.

The shape of hyperbolic lines suggests what form this metric must take. Suppose, for instance, that $x, y \in \mathbb{H}^2$ are points with the same height. The hyperbolic line segment between them is part of a semicircle orthogonal to the x-axis. If this is to be the shortest path from x to y, then there must be some reason why taking a detour upwards is more efficient then taking the horizontal path.



We saw similar phenomena when discussing map projections in Section 2.4. For instance, the shortest path from New York to London curves strangely upward toward Greenland when viewed in the Mercator projection, to take advantage of the fact that distances near the poles are much smaller than they appear in the projection.

To define a distance on $\mathbb{H}^2$, we will require that near a point $(x, y) \in \mathbb{H}^2$, distances should be distorted by a factor of $\frac{1}{y}$. That is, the actual (hyperbolic) size of an object near (x, y) should be $\frac{1}{y}$ times its apparent (Euclidean) size. To make this rigorous,

Definition 3.6.1 (Hyperbolic length). Suppose that $\gamma : [a, b] \longrightarrow \mathbb{H}^2$ is a path, where $\gamma(t) = (\gamma_1(t), \gamma_2(t))$. The *hyperbolic length* of γ is defined to be

$$\text{length}_{\mathbb{H}^2}(\gamma) = \int_a^b |\gamma'(t)| \frac{1}{\gamma_2(t)}\, dt.$$

Recall that $|\gamma'(t)|$ is the Euclidean *speed* of γ, which integrates to the Euclidean length of γ. The integrand $|\gamma'(t)|\frac{1}{\gamma_2(t)}$ is called the *hyperbolic speed* of γ. If hyperbolic distances near $\gamma(t)$ are $1/\gamma_2(t)$-times the corresponding Euclidean distances, then it should make sense that the hyperbolic speed of γ is the same factor times the corresponding Euclidean speed. In the same way that Euclidean speed integrates to length, the hyperbolic length as the integral of the hyperbolic speed.

Example 3.6.2. The paths for which it is easiest to compute hyperbolic length are horizontal paths. For if $\alpha : [a, b] \longrightarrow \mathbb{H}^2$ has constant second coordinate $\alpha_2(t) = y$,

$$\text{length}_{\mathbb{H}^2}(\alpha) = \int_a^b |\alpha'(t)| \frac{1}{y}\, dt = \int_a^b |\alpha'(t)| \frac{1}{y}\, dt = \frac{1}{y} \text{length}(\alpha),$$

so hyperbolic length is just Euclidean length divided by height.

Example 3.6.3. Let's now compute the length of a vertical line segment from (x, y_1) to (x, y_2) in $\mathbb{H}^2$. Parametrically, $\gamma : [y_1, y_2] \longrightarrow \mathbb{H}^2$, where $\gamma(t) = (x, t)$. Well,

$$\text{length}_{\mathbb{H}^2}(\gamma) = \int_{y_1}^{y_2} |\gamma'(t)| \frac{1}{\gamma_2(t)}\, dt$$

$$\begin{aligned} &= \int_{y_1}^{y'} 1 \cdot \frac{1}{t}\, dt \\ &= \ln(y_2) - \ln(y_1) \\ &= \ln\left(\frac{y_2}{y_1}\right). \end{aligned}$$

In particular, the hyperbolic length of the line segment joining $(0, 1/n)$ to $(0, 1)$ is the same as that joining $(0, 1)$ to $(0, n)$! This is a reflection of the fact that hyperbolic distances high up in the half plane are much smaller than they appear, while hyperbolic distances near the x-axis are much larger than they appear.

Armed with a notion of hyperbolic length, we now define distance in $\mathbb{H}^2$. Just as on the sphere, distance is defined as the infimum of length of paths.

DEFINITION 3.6.4. The *hyperbolic distance* between two points $p, q \in \mathbb{H}^2$ is

$$d_{\mathbb{H}^2}(p, q) = \inf\left\{ \text{length}_{\mathbb{H}^2}(\gamma) \mid \gamma : [a, b] \longrightarrow \mathbb{H}^2,\ \gamma(a) = p,\ \gamma(b) = q \right\}.$$

Hyperbolic distance defines a metric on $\mathbb{H}^2$. We'll leave verifying the relevant properties as an exercise. Mostly, the proof is the same as that in the spherical case, except that there is a little bit more subtlety in proving that if $p \neq q$ then $d(p, q) > 0$.

EXAMPLE 3.6.5. What's the hyperbolic distance between the points $(a, 1)$ and $(b, 1)$ in $\mathbb{H}^2$, say when $b > a$? We'll be able to compute it on the nose by the end of the section, but it's easy to give an interesting upper bound.

Create a path γ by starting at $(a, 1)$, moving vertically to $(a, b-a)$, then horizontally to $(b, b-a)$, and vertically back down to $(b, 1)$. From Examples 3.6.2 and 3.6.3, these three segments have length $\ln(b-a)$, 1 and $\ln(b-a)$, respectively, so

$$d_{\mathbb{H}^2}\big((a, 1), (b, 1)\big) \leqslant \text{length}_{\mathbb{H}^2}(\gamma) = 2\ln(b-a) + 1.$$

So for example, the hyperbolic distance between $(0, 1)$ and $(1000000, 1)$ is a bit less than 30, which is *much* shorter than the length of the horizontal segment joining them! This illustrates that it is a tremendous advantage for a path to detour upwards and exploit the smaller distances at higher altitudes.

So, what paths use the distortion of the hyperbolic metric optimally? That is, we know that it is efficient to bend upwards, but certainly there is a limit to how much of an upwards detour a path should make in order to minimize length. As you might be guessing, these most efficient paths are exactly the hyperbolic line segments.

Here is a first case where we can verify this.

LEMMA 3.6.6. *The path with the shortest hyperbolic length between points $p = (x, y_1)$ and $q = (x, y_2)$ in $\mathbb{H}^2$ with the same first coordinate is the vertical line segment.*

PROOF. Assume $y_2 > y_1$. We've seen that the hyperbolic length of the line segment joining p and q is $\ln(y_2/y_1)$. So if $\gamma : [a, b] \longrightarrow \mathbb{H}^2$ is a path from p to q,

we claim that $\text{length}_{\mathbb{H}^2}(\gamma) \geqslant \ln(y_2/y_1)$, with equality if and only if γ is a vertical line segment.

Writing $\gamma(t) = (\gamma_1(t), \gamma_2(t))$, we compute:

$$\begin{aligned}
\text{length}_{\mathbb{H}^2}(\gamma) &= \int_a^b |\gamma'(t)| \frac{1}{\gamma_2(t)}\, dt \\
&= \int_a^b \sqrt{\gamma_1'(t)^2 + \gamma_2'(t)^2} \frac{1}{\gamma_2(t)}\, dt \\
&\geqslant \int_a^b \sqrt{0^2 + \gamma_2'(t)^2} \frac{1}{\gamma_2(t)}\, dt, \\
&\geqslant \int_a^b \gamma_2'(t) \frac{1}{\gamma_2(t)}\, dt, \\
&= \ln(\gamma_2(b)) - \ln(\gamma_2(a)) \\
&= \ln(y_2/y_1)
\end{aligned}$$

as desired. Equality occurs exactly when $\gamma_1'(t) = 0$ and $\gamma_2'(t) > 0$ for all t, which means that γ moves straight up the vertical line segment from p to q. □

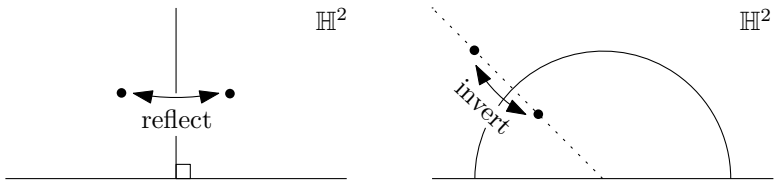
To prove in general that the shortest path between two points in $\mathbb{H}^2$ lies along the hyperbolic line joining them, we'll transform the general case into the special case above using a 'hyperbolic reflection'.

DEFINITION 3.6.7. If ℓ is a hyperbolic line, it is the intersection with $\mathbb{H}^2$ of either a vertical line L or a circle C orthogonal to the x-axis. The *hyperbolic reflection*

$$R_\ell : \mathbb{H}^2 \longrightarrow \mathbb{H}^2$$

through ℓ is the restriction to $\mathbb{H}^2$ of either the Euclidean reflection through L or the inversion through C, depending on which case we're in.

Note that since L and C are orthogonal to the x-axis, in both cases the reflection/inversion preserves the x-axis and does indeed send points in $\mathbb{H}^2$ into $\mathbb{H}^2$.



THEOREM 3.6.8. *If ℓ is a hyperbolic line and γ is a path in $\mathbb{H}^2$, then*

$$\text{length}_{\mathbb{H}^2}(R_\ell \circ \gamma) = \text{length}_{\mathbb{H}^2}(\gamma).$$

So, R_ℓ is a hyperbolic isometry: $d_{\mathbb{H}^2}(R_\ell(p), R_\ell(q)) = d_{\mathbb{H}^2}(p, q)$ for all $p, q \in \mathbb{H}^2$.

PROOF. If $\gamma : [a, b] \longrightarrow \mathbb{H}^2$ is a path, we will show that for all $t \in [a, b]$,

$$|\gamma'(t)|\frac{1}{\gamma_2(t)} = |(R_\ell \circ \gamma)'(t)|\frac{1}{(R_\ell \circ \gamma)_2(t).} \tag{12}$$

That is, the hyperbolic speeds of γ and $R_\ell \circ \gamma$ agree at time t. Integrating, this implies that $\text{length}_{\mathbb{H}^2}(\gamma) = \text{length}_{\mathbb{H}^2}(R_\ell \circ \gamma)$, so R_ℓ preserves path lengths. As hyperbolic distance is defined in terms of path lengths, R_ℓ must be an isometry.

Case 1. If ℓ is a vertical half line, then R_ℓ is a Euclidean isometry. Therefore, the Euclidean speeds of γ and $R_\ell \circ \gamma$ must agree for all t. Furthermore, since ℓ is vertical, the heights of $\gamma(t)$ and $R_\ell \circ \gamma(t)$ are the same. So, we have

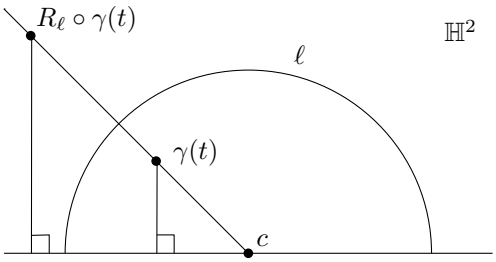
$$|\gamma'(t)| = |(R_\ell \circ \gamma)'(t)| \quad \text{and} \quad \gamma_2(t) = (R_\ell \circ \gamma)_2(t),$$

from which Equation (12) follows.

Case 2. Suppose now that ℓ is a semicircle with center c and radius r. In this case, R_ℓ preserves neither height nor Euclidean speed, and the goal is to show that the height distortion exactly matches the speed distortion, i.e.

$$\frac{|(R_\ell \circ \gamma)'(t)|}{|\gamma'(t)|} = \frac{(R_\ell \circ \gamma)_2(t)}{\gamma_2(t)}.$$

For each t the points $\gamma(t)$ and $R_\ell \circ \gamma(t)$ determine right triangles with bases on the x-axis and one vertex equal to c.



Since these right triangles share an angle at c, they are similar. Therefore, if r is the radius of the circle, the height ratio of $\gamma(t)$ and $R_\ell \circ \gamma(t)$ is

$$\frac{(R_\ell \circ \gamma)_2(t)}{\gamma_2(t)} = \frac{|R_\ell \circ \gamma(t) - c|}{|\gamma(t) - c|} = \frac{\frac{r^2}{|\gamma(t)-c|}}{|\gamma(t) - c|} = \frac{r^2}{|\gamma(t) - c|^2}. \tag{13}$$

This matches the speed distortion, by Theorem 3.4.4, so Equation (12) follows. □

COROLLARY 3.6.9. *If $p, q \in \mathbb{H}^2$, the (hyperbolically) shortest path from p to q is the hyperbolic line segment joining them.*

PROOF. Let ℓ be the hyperbolic line through p, q, and let z be one of its endpoints on the x-axis. Take any hyperbolic line m that is a semicircle centered at z. As the reflection R_m preserves path lengths, it must take shortest paths joining p, q to

shortest paths joining $R_m(p)$ and $R_m(q)$. But secretly, R_m is an inversion! So, as ℓ passes through the center of the circle of inversion, $R_m(\ell)$ is a line, which must again be orthogonal to the x-axis since R_m is conformal.

This means that $R_m(\ell)$ is a vertical line, on which lie $R_m(p)$ and $R_m(q)$. By Lemma 3.6.6, the shortest path from $R_m(p)$ to $R_m(q)$ is the segment of $R_m(\ell)$ joining them, which implies that the shortest path joining p, q is the corresponding segment of ℓ. □

In her book *The Universe in Zero Words: The Story of Mathematics as Told Through Equations*, Dr. Dana Mackenzie explains how hyperbolic geometry is the 'geometry of whales'. Turning the hyperbolic plane upside down, imagine the x-axis as the surface of the ocean, the depths of which are populated by whales.

Deep in the ocean, there is not so much light, and whales communicate by sonar. Sound travels in deep water at a rate proportional to the inverse of depth, so from the perspective of sound distances in the ocean are scaled by $1/y$, the hyperbolic scaling factor. This means that the most efficient way for sound to travel from one whale to another is to bend downwards along a hyperbolic line.

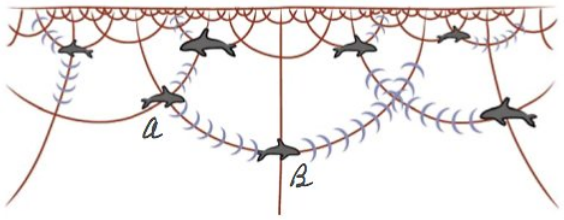
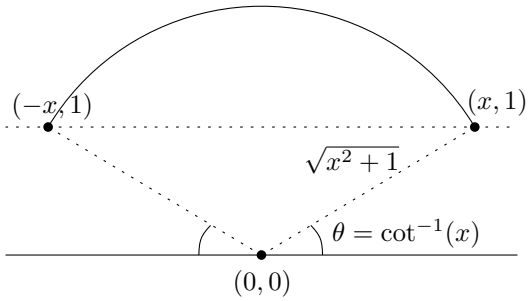


FIGURE 1. Taken from *The Universe in Zero Words*, by Dana Mackenzie.

This picture isn't completely accurate, as the surface of the ocean isn't infinitely far away from the whale, as is the x-axis from any point in $\mathbb{H}^2$, but the idea is beautiful.

EXERCISE 3.6.10. In Example 3.6.5, we estimated the hyperbolic distance between points in $\mathbb{H}^2$ with second coordinate 1. You can now perform an exact calculation; for simplicity, we will consider the points $(-x, 1)$ and $(x, 1)$.



Using the picture above as a guide, show that the hyperbolic distance

$$d_{\mathbb{H}^2}\big((-x,1),(x,1)\big) = 2\ln\cot\left(\frac{\cot^{-1}(x)}{2}\right). \tag{14}$$

You might need that the anti-derivative of $\csc(t)$ *is* $\ln\tan(t/2)$.

This expression in the exercise above is a bit ugly, but you can compare it to the estimate we gave in Example 3.6.5, which in this case is $2\ln(2x)$. It turns out that

$$\lim_{x\to\infty} \frac{2\ln\cot\left(\frac{\cot^{-1}(x)}{2}\right)}{2\ln(2x)} = 1,$$

which you can check if you like, so in fact the simpler estimate is pretty accurate as long as x is large! In other words, for large x the three line segments in Example 3.6.5 form a somewhat length-efficient approximation of the hyperbolic line segment.

Exercise 3.6.11. You and a friend walk upwards along the vertical half lines $x = a$ and $x = b$ at unit hyperbolic speed. Parametrically, your paths are α and β, where

$$\alpha(t) = (a, e^t), \quad \beta(t) = (b, e^t)$$

since the hyperbolic speeds at time t are

$$\frac{|\alpha'(t)|}{\alpha_2(t)} = \frac{|\beta'(t)|}{\beta_2(t)} = \frac{\sqrt{0^2 + (e^t)^2}}{e^t} = 1.$$

Show that the distance between you and your friend at time t satisfies

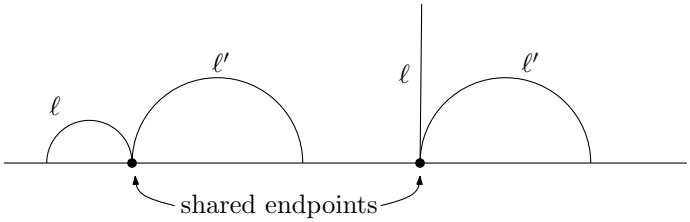
$$d_{\mathbb{H}^2}(\alpha(t), \beta(t)) \leqslant \frac{|b-a|}{e^t}.$$

Two paths α, β in $\mathbb{H}^2$ are *asymptotic* if they can be parameterized so that

$$\lim_{t\to\infty} d_{\mathbb{H}^2}(\alpha(t), \beta(t)) = 0.$$

The exercise above shows that any two vertical half lines in $\mathbb{H}^2$ are asymptotic, as $\frac{|b-a|}{e^t} \longrightarrow 0$ as $t \longrightarrow \infty$. As vertical half lines are those hyperbolic lines that have an 'endpoint at infinity', the following exercise is an extension of the previous one.

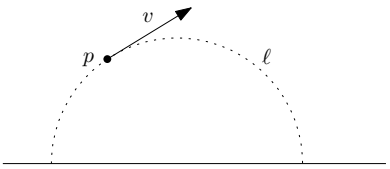
Exercise 3.6.12. Suppose that two hyperbolic lines ℓ, ℓ' share an endpoint on the x-axis, as pictured below. Show that there is some hyperbolic line m such that $R_m(\ell)$ and $R_m(\ell')$ are vertical half lines, and use this and the previous exercise to show that ℓ and ℓ' are asymptotic. *For the last part, you will need to parameterize* ℓ *and* ℓ' *as indicated above. However, don't worry about writing out an actual formula. Instead, compose parameterizations for the vertical half lines with* R_m.



So, even though hyperbolic distances near the x-axis are much larger than they appear, two hyperbolic lines that share an endpoint on the x-axis are getting close to each other quickly enough to overcome this distance distortion.

If $p \in \mathbb{R}^2$, there are lines through p in every direction. We'd like to say that the same is true in the hyperbolic plane.

Exercise 3.6.13. Show that if $p \in \mathbb{H}^2$ and v is a vector based at p, there is a unique hyperbolic line ℓ passing through p tangent to v.



3.7. Hyperbolic trigonometric functions

The *hyperbolic trigonometric functions* sinh, cosh, tanh, sech and csch, pronounced 'hyperbolic sin/cos/tan/sec/csc', are defined as follows.

$$\sinh(t) = \frac{e^t - e^{-t}}{2}, \quad \cosh(t) = \frac{e^t + e^{-t}}{2}, \quad \tanh(t) = \frac{\sinh(t)}{\cosh(t)},$$

$$\operatorname{csch}(t) = \frac{1}{\sinh(t)}, \quad \operatorname{sech}(t) = \frac{1}{\cosh(t)}, \quad \coth(t) = \frac{1}{\tanh(t)}.$$

Hyperbolic trigonometric functions have many features that are similar to their Euclidean counterparts. For instance, easy calculations show that

$$\sinh'(t) = \cosh(t), \quad \cosh'(t) = \sinh(t), \quad \tanh'(t) = \frac{1}{\cosh^2(t)},$$

There are also addition formulas similar to those in the Euclidean case.

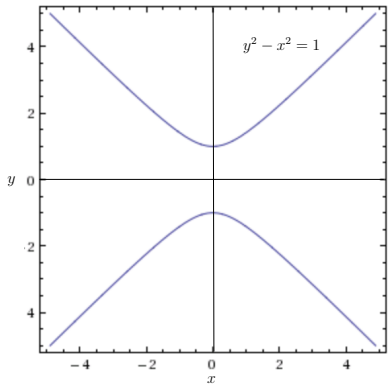
$$\sinh(t+s) = \sinh(t)\cosh(s) + \cosh(t)\sinh(s),$$
$$\cosh(t+s) = \cosh(t)\cosh(s) + \sinh(t)\sinh(s).$$

You can find many more identities on the Wikipedia page 'Hyperbolic function'. However, there's one identity we should discuss more thoroughly:

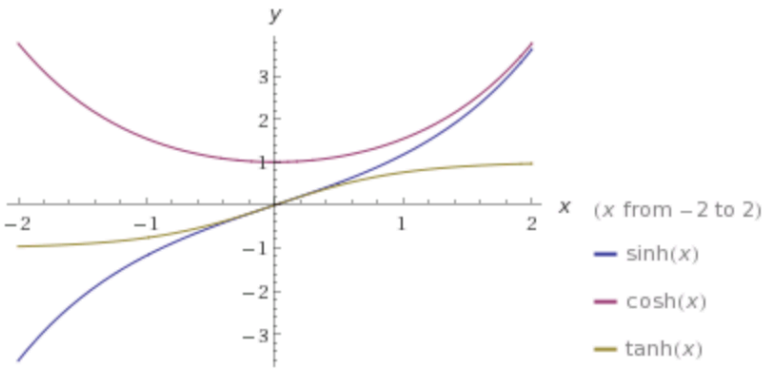
$$\begin{aligned}\cosh^2(t) - \sinh^2(t) &= \frac{(e^t + e^{-t})^2 - (e^t - e^{-t})^2}{4} \\ &= \frac{e^{2t} + 2 + e^{-2t} - e^{2t} + 2 - e^{-2t}}{4} \\ &= 1.\end{aligned} \tag{15}$$

Of course, this is similar to the familiar identity $\sin^2(t) + \cos^2(t) = 1$. The meaning of the latter is that for each t, the point $(\sin(t), \cos(t))$ lies on the unit circle $x^2 + y^2 = 1$; of course, we even know that the path $t \mapsto (\sin(t), \cos(t))$ parameterizes the unit circle.

The equation $y^2 - x^2 = 1$ defines a *hyperbola* instead of the circle, and (15) reflects that the path $t \mapsto (\sinh(t), \cosh(t))$ parameterizes (the top half of) this hyperbola.

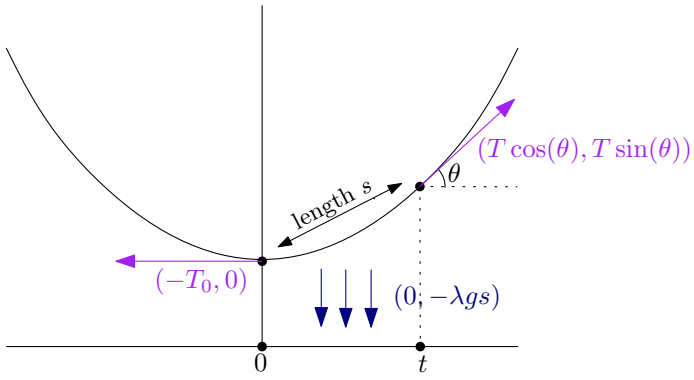


Let's analyze the graphs of sinh, cosh and tanh. When $t >> 0$, we have $e^{-t} \approx 0$, so for large positive t, it follows that $\sinh(t) \approx \cosh(t)$. Similarly, for large negative t we have $e^t \approx 0$, so $\sinh(t) \approx -\cosh(t)$. In the graphs, one sees that sinh and cosh are asymptotic as $t \to \infty$ and tanh has horizontal asymptotes at ± 1.



3.7.1. Catenaries. The graph of cosh is an example of a *catenary*, a curve that a rope traces out when it hangs under its own weight from two anchors.


To prove this, we must consider the forces acting on such a length of rope. In the figure below, fix attention on a part of the rope with length s that starts at the lowest point on the rope. There are three forces acting on this part of the rope. The downward force of *gravity* is represented by the vector $(0, -\lambda g s)$, where g is a gravitational constant and λ is the mass per unit length of the rope. There are also *tension* forces at each end that are tangent to the rope. The tension at the lowest point has magnitude T_0 and the tension at the other endpoint has magnitude T.



Since the rope is stationary, these three forces sum to zero. Thus, we have

$$T\cos(\theta) = -T_0, \quad T\sin(\theta) = -\lambda g s, \implies \tan(\theta) = \frac{\lambda g}{T_0} s.$$

Imagine now that the rope is the graph of a function f. Then $\tan(\theta)$ is just the rise over run at that point, i.e. the derivative $f'(t)$.

Notice that T_0, λ and g are all just constants that depend on the environmental conditions and the type of rope used. So, if we combine them into one constant $a = \frac{T_0}{\lambda g}$, then our equation becomes $f'(t) = s/a$. That is,

($\star$) *The length of the graph of $f(x)$ from $x = 0$ to $x = t$ is $af'(t)$.*

EXERCISE 3.7.1. Show that the function $f(x) = a\cosh(x/a)$ satisfies ($\star$).

Therefore, the graphs of the functions $f(x) = a\cosh(x/a)$ model the shapes of hanging ropes, where a depends on the environmental conditions and type of rope. To help you with the problem, note that the graph of f can be parameterized as

$$\gamma : [0, t] \longrightarrow \mathbb{R}^2, \quad \gamma(x) = (x, f(x))$$

so the length of the graph between $x = 0$ and $x = t$ is

$$\text{length}(\gamma) = \int_0^t |\gamma'(x)|\, dx = \int_0^t \sqrt{1 + f'(x)^2}\, dx.$$

3.7.2. Inverse hyperbolic trig functions and a distance formula. We defined hyperbolic trig functions above using exponentials, so it may not be a surprise that their inverses can be conveniently described using logarithms. For instance,

$$\cosh(x) = \frac{1}{2}(e^x + e^{-x})$$

is 1-1 for nonnegative x and has range equal to $[1, \infty)$, so there is an inverse function

$$\cosh^{-1} : [1, \infty) \longrightarrow [0, \infty),$$

where $\cosh^{-1}(y)$ is the unique nonnegative number x such that $\cosh(x) = y$.

EXERCISE 3.7.2. Show that $\cosh^{-1}(y) = \ln\left(y + \sqrt{y^2 - 1}\right)$, where $y \geqslant 1$.

If you are so inclined, try to come up with formulas using logarithms for the inverses of the other hyperbolic trigonometric functions.

In Section 3.6, we saw that the hyperbolic distance between two points in $\mathbb{H}^2$ is the length of the hyperbolic line segment joining them. Here is an actual formula for hyperbolic distance using the inverse hyperbolic cosine.

THEOREM 3.7.3. *The distance between $p = (p_1, p_2)$ and $q = (q_1, q_2)$ in $\mathbb{H}^2$ is*

$$d_{\mathbb{H}^2}(p, q) = \cosh^{-1}\left(1 + \frac{|p - q|^2}{2p_2q_2}\right).$$

We will split the proof of Theorem 3.7.3 into three exercises (3.7.4 - 3.7.6). The idea is to first prove it for pairs of points that lie on a vertical line, then to use hyperbolic reflections to reduce the general case to this first case.

EXERCISE 3.7.4. Prove the theorem when $p_1 = q_1$. That is, if p, q lie on a vertical line, show that their hyperbolic distance is given by the formula above. *Hint: in this case, the shortest path is just the segment of the vertical line between them.*

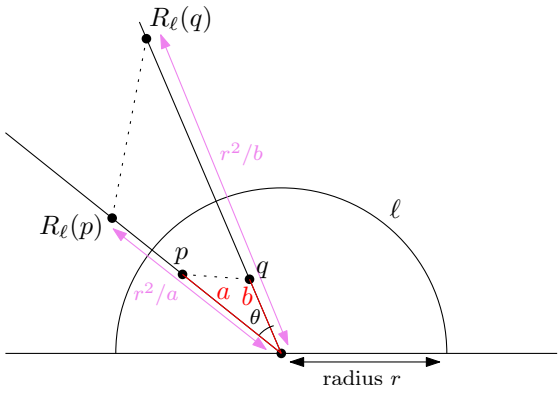
To prove the theorem in general, let's define D to be the right-hand side

$$D(p,q) = \cosh^{-1}\left(1 + \frac{|p-q|^2}{2p_2q_2}\right).$$

EXERCISE 3.7.5. If ℓ is a hyperbolic line, show that $D(R_\ell(p), R_\ell(q)) = D(p,q)$ for all $p, q \in \mathbb{H}^2$. *Hint: the figure below depicts the reflection of p, q through ℓ. The points p, q are assumed to be at distances a, b from the center of ℓ. Show that*

$$|R_\ell(p) - R_\ell(q)|^2 = \frac{r^4}{(ab)^2}|p-q|^2, \quad (R_\ell(p))_2 = \frac{r^2}{a^2}p_2, \quad \textit{and} \quad (R_\ell(q))_2 = \frac{r^2}{b^2}q_2.$$

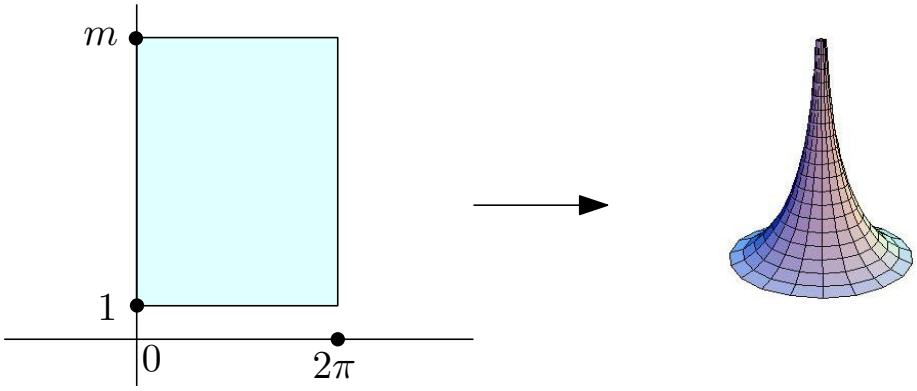
For the first equation, you might find the law of cosines useful. The second two equations should not take more than one sentence to prove.



EXERCISE 3.7.6. Using the previous two problems, prove the theorem in general. *Hint: if $p, q \in \mathbb{H}^2$, let ℓ be the line through p, q. There is a hyperbolic reflection R such that $R(\ell)$ is a vertical half-line in $\mathbb{H}^2$. Now apply the previous problems to $R(p)$ and $R(q)$.*

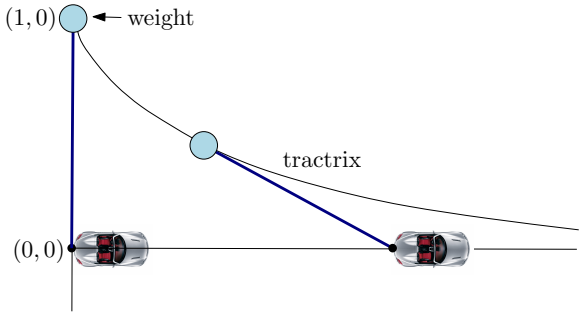
3.8. The pseudosphere and the tractrix

Consider the rectangle R in the hyperbolic plane below. The surface created by identifying the vertical sides of R is called a *pseudosphere*. As the hyperbolic length of the cross-section of R at height y decreases as y increases, the 'circumference' of the pseudosphere decreases with height.



With respect to the specific dimensions of the rectangle in the picture, the circumference of the pseudosphere is 2π at the bottom and decreases to $\frac{2\pi}{m}$ at the top.

The pseudosphere above is the surface of revolution of a *tractrix*. Imagine that the xy-plane represents the Earth and the x-axis is a road. If a car located at $(0,0)$ is anchored by a taught chain to a weight at $(0,1)$, the tractrix is the path traced out by the weight as the car moves to the right along the x-axis.

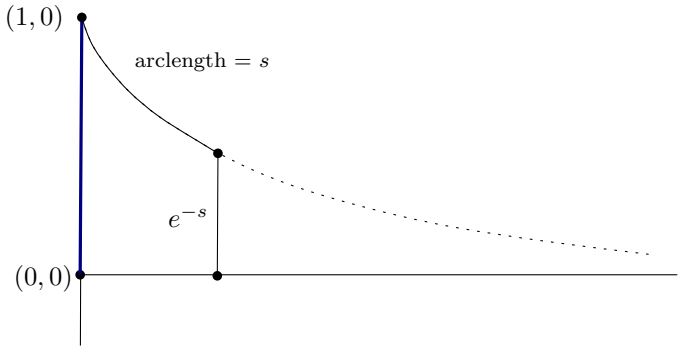


The name comes from the Latin word *trahere*, meaning to pull or drag. In the picture above, the force the car applies to the weight is always in the direction towards the car. Therefore, if you move to the right along the x-axis at unit speed, then at time t the box should be at distance 1 from $(t,0)$ and moving in the direction of $(t,0)$.

To explain why the pseudosphere is obtained by revolving the tractrix, we should analyze how fast the cross-sectional circumferences of the pseudosphere are decreasing. Initially, you might be tempted to say that the pseudosphere is obtained by revolving the curve $y = 1/x$ around the x-axis, since the cross-sectional lengths of R decay inversely with height. However, this is not accurate because height in $\mathbb{H}^2$ does not quite correspond with height in the pseudosphere.

Instead, looking at the rectangle R we see that the height y cross-section has length $2\pi/y$ and is at hyperbolic distance $\ln(y)$ from the bottom of the rectangle; in other words, the cross-section at hyperbolic distance s from the bottom has length $2\pi e^{-s}$. The same is then true for the pseudosphere, and the distance on a surface of revolution

between two cross-sections is just the arc length of the revolved curve. So, we want to show that after the tractrix has used arc length s, its height is e^{-s}.



To show this, we will need to find a parameterization of the tractrix.

PROPOSITION 3.8.1. *The tractrix can be prescribed parametrically as*

$$\gamma : [0, \infty) \longrightarrow \mathbb{R}^2, \quad \gamma(t) = (t - \tanh(t), \operatorname{sech}(t)).$$

EXERCISE 3.8.2. We leave the proof of the above proposition as a guided exercise.

(a) If $\gamma(t)$ is the weight's position at time t, explain why

$$\gamma(t) + \frac{1}{|\gamma'(t)|}\gamma'(t) = (t, 0).$$

(b) Show that if $\gamma(t) = (t - \tanh(t), \operatorname{sech}(t))$, then

$$\gamma'(t) = (\tanh^2(t), -\tanh(t)\operatorname{sech}(t)), \quad |\gamma'(t)| = \tanh(t).$$

(c) Show that $\gamma(t) = (t - \tanh(t), \operatorname{sech}(t))$ satisfies the differential equation from a) and the initial conditions $\gamma(0) = (0, 1)$ and $\gamma'(0) = (0, 0)$. As the tractrix is completely determined by the initial conditions and the differential equation of a), this proves the proposition.

We can now verify that the pseudosphere is obtained by revolving the tractrix around the x-axis. First, the arc length of the tractrix from $t = 0$ to $t = a$ is given by

$$\begin{aligned}\int_0^a |\gamma'(t)|dt &= \int_0^a \tanh(t)dt \\ &= \ln\cosh(t)\big|_0^a \\ &= \ln\cosh(a).\end{aligned}$$

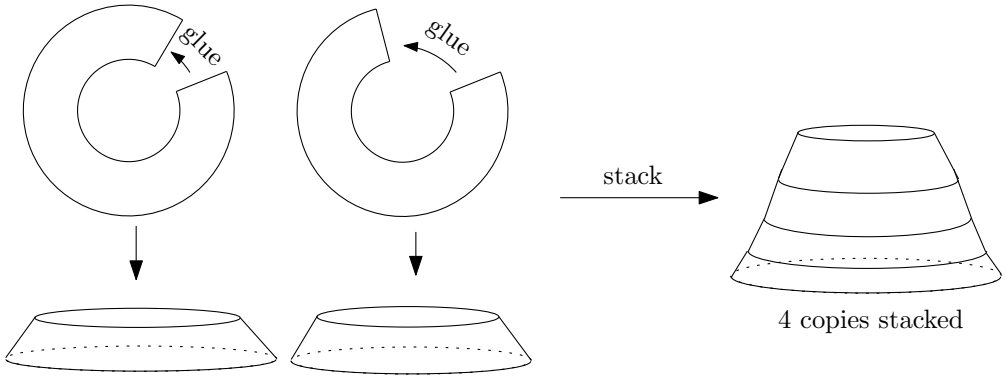
This arc length is s when $a = \cosh^{-1}(e^s)$, at which point the height of the tractrix is

$$\operatorname{sech}(\cosh^{-1}(e^a)) = e^{-s},$$

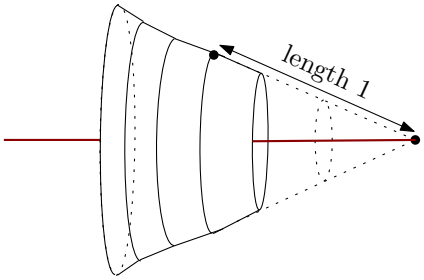
as desired. This shows that the surface of revolution is the pseudosphere.

The physical description of the tractrix can be used to construct paper models of the pseudo-sphere. Cut out many identical copies of an *annulus*, the region between

two concentric circles. Cut larger and larger sectors out of these annuli and attach the exposed cuts with tape. We now have a number of paper bracelets, that we stack in order of size to create an approximation to a pseudosphere.



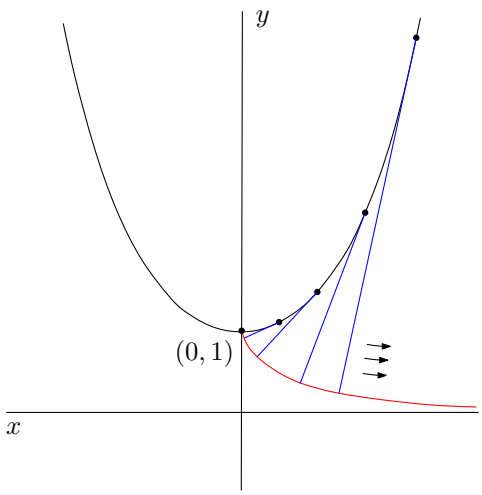
This approximation is also a surface of revolution, so our claim is that the profile curve approximates a tractrix. The tractrix is determined by the condition that at each point, the tangent line intersects the x-axis after one unit of length. As long as the outer circle used in creating the annuli has radius 1, the same condition holds at every point on the approximate that lies on one of these outer circles.



When the increment between adjacent sector sizes is small, every point on our paper model is very close to these outer circles. So, in the limit the profile curve becomes the tractrix and our paper model becomes the pseudosphere.

Cutting our paper model along a profile curve gives a geometric approximation to the region R in $\mathbb{H}^2$, which is our first physical model for the hyperbolic plane. Try to use this model to understand some of the properties of hyperbolic geometry we have discussed – for instance, can you see the homogeneity within the model?

There is an interesting relationship between the tractrix and the catenary. Imagine laying a strip of tape on the entire length of the graph of $y = \cosh(x)$, for $x \geqslant 0$. Then grab the end of the tape at $(0, 1)$, and slowly pull downwards. As the tape unwraps, the end you are holding will sweep out the tractrix. Below, the tractrix is in red and a few snapshots of the tape are drawn in blue.



One summarizes the above by saying that the tractrix is the *involute* of the catenary. More precisely, the involute of a parameterized curve $\gamma : [0, b] \longrightarrow \mathbb{R}^2$ is the curve $\alpha : [0, b] \longrightarrow \mathbb{R}^2$ such that

$$\alpha(t) = \gamma(t) - \left(\int_0^t |\gamma'(s)|\, ds \right) \frac{\gamma'(t)}{|\gamma'(t)|}.$$

This formula may look complicated, but it is just saying that at time t, the point $\alpha(t)$ is obtained by traveling from $\gamma(t)$ in the direction opposite to the velocity $\gamma'(t)$ for a distance that is equal to the length of γ from 0 to t. In other words, the subtracted term on the right represents the blue lines in the picture above.

EXERCISE 3.8.3. Verify that if $\gamma : [0, \infty) \longrightarrow \mathbb{R}^2$, $\gamma(t) = (t, \cosh(t))$ parameterizes the right half of the catenary, its involute is the tractrix $\alpha(t) = (t - \tanh(t), \operatorname{sech}(t))$.

EXERCISE 3.8.4. Find, and draw, the involute of $\gamma : [0, 2\pi] \longrightarrow \mathbb{R}^2$ when

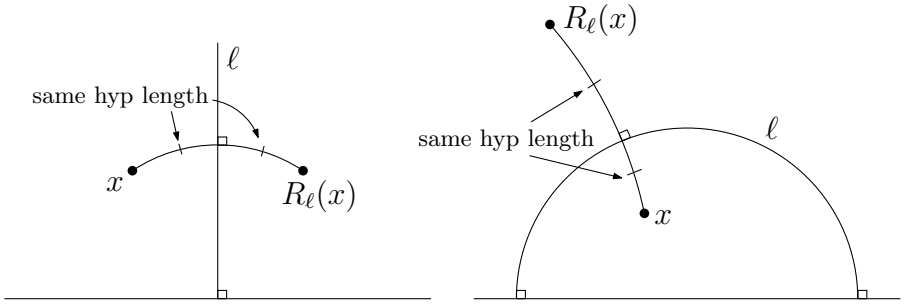
(a) $\gamma(t) = (\cos(t), \sin(t))$.
(b) $\gamma(t) = ((1 + \cos t)\cos t, (1 + \cos t)\sin t)$. *This γ is called a cardioid, since when drawn it looks like a heart. Its involute will also be a cardioid, but scaled, reflected and translated. Feel free to use a computer to plot γ and its involute, as it'll be a bit difficult to do by hand.*

3.9. Hyperbolic isometries

In Theorem 3.6.8, we saw that hyperbolic reflections are isometries of $\mathbb{H}^2$. We defined the hyperbolic reflection through a hyperbolic line ℓ as either the reflection or inversion through ℓ, depending on whether ℓ is a half line or a semicircle.

Using hyperbolic distance, however, hyperbolic reflections admit a characterization analogous to that of Euclidean reflections.

PROPOSITION 3.9.1. *If ℓ is a hyperbolic line, then $R_\ell(x) = x$ for all $x \in \ell$. If $x \notin \ell$, then ℓ is the perpendicular bisector of the hyperbolic line segment $xR_\ell(x)$.*



PROOF. Since R_ℓ exchanges x and $R_\ell(x)$ and takes hyperbolic lines to hyperbolic lines, it must take the hyperbolic line segment $xR_\ell(x)$ to itself. This implies that $xR_\ell(x)$ is perpendicular to ℓ. If p is the point of intersection, then R_ℓ takes the segment xp to the segment $R_\ell(x)p$. So, as R_ℓ preserves hyperbolic path lengths these two segments must have the same length. □

PROPOSITION 3.9.2. *If $x, y \in \mathbb{H}^2$, there is a hyperbolic line ℓ such that $R_\ell(x) = y$.*

PROOF. Let p be the hyperbolic midpoint of the hyperbolic line segment xy, and let ℓ be the unique hyperbolic line through p perpendicular to xy. (The existence of this line follows from Exercise 3.6.13.) Then ℓ is the hyperbolic perpendicular bisector of xy, and it follows from the above proposition that $R_\ell(x) = y$. □

This shows that for any pair of points $x, y \in \mathbb{H}^2$, there is an an isometry of $\mathbb{H}^2$ taking x to y. In other words, we have proven the following.

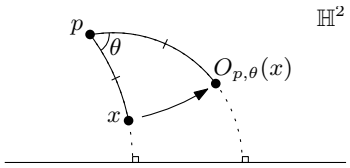
COROLLARY 3.9.3. $\mathbb{H}^2$ *is homogenous.*

What other isometries of $\mathbb{H}^2$ are there?

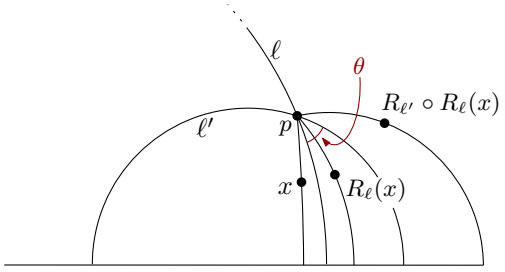
DEFINITION 3.9.4. If $p \in \mathbb{H}^2$ and $\theta \in [0, 2\pi)$, the *rotation* around p by θ is the map

$$O_{p,\theta} : \mathbb{H}^2 \longrightarrow \mathbb{H}^2,$$

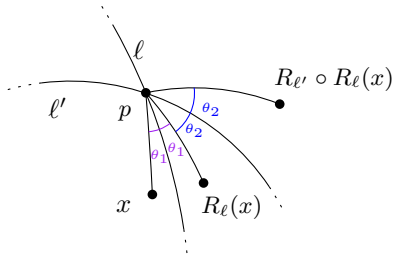
where $O_{p,\theta}(p) = p$ and for $x \neq p$ the hyperbolic line segments px and $pO_{p,\theta}(x)$ have the same length and meet with angle θ, measured counterclockwise from px to $pO_{p,\theta}(x)$.



PROPOSITION 3.9.5. *If ℓ, ℓ' are hyperbolic lines that intersect at $p \in \mathbb{H}^2$ with an angle of θ, measured counterclockwise from ℓ to ℓ', then $R_{\ell'} \circ R_\ell = O_{p,2\theta}$.*



PROOF. The composition $R_{\ell'} \circ R_\ell$ fixes p and is an isometry, so if $x \neq p$ then the hyperbolic line segments px and $pR_{\ell'} \circ R_\ell(x)$ have the same length.



Since hyperbolic reflections preserve angles, the two θ_1 angles and the two θ_2 angles are equal in the picture below. As $\theta = \theta_1 + \theta_2$, the angle between the segments px and $pR_{\ell'} \circ R_\ell(x)$ is $2\theta_1 + 2\theta_2 = 2\theta$. □

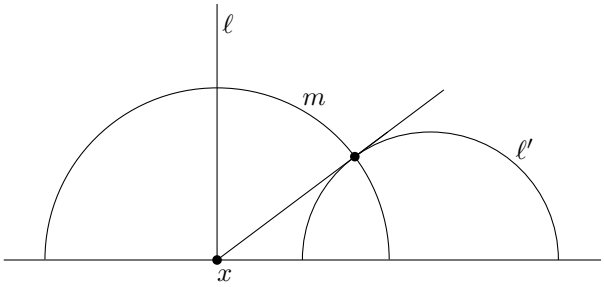
It is easy to see using Proposition 3.9.5 that any hyperbolic rotation can be written as the product of two reflections, so as reflections are isometries we have:

COROLLARY 3.9.6. *Hyperbolic rotations are isometries of* $\mathbb{H}^2$.

What happens if we compose two reflections through hyperbolic lines ℓ, ℓ' that do not intersect in $\mathbb{H}^2$? The answer is complicated, since ℓ and ℓ' can intersect on the x-axis, or not, and can also 'intersect at ∞' when they are both vertical lines. Define *the boundary of* $\mathbb{H}^2$, written $\partial\mathbb{H}^2$, as the x-axis together with the 'point' ∞. The type of the isometry $f = R_{\ell'} \circ R_\ell$ depends on whether ℓ and m intersect on $\partial\mathbb{H}^2$ or not.

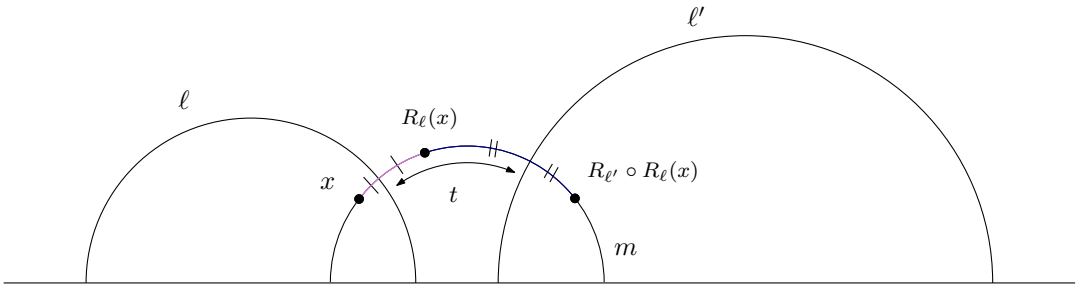
LEMMA 3.9.7. *Suppose* ℓ *and* ℓ' *are hyperbolic lines that do not intersect, even in* $\partial\mathbb{H}^2$. *Then there is a unique hyperbolic line* m *that intersects both* ℓ *and* ℓ' *orthogonally.*

PROOF. As a warm-up, let's assume ℓ is a vertical half-line. Then ℓ' is a semicircle, since ℓ, ℓ' do not intersect at $\infty \in \partial\mathbb{H}^2$. A hyperbolic line m intersects ℓ orthogonally if and only if it is a semicircle centered at the endpoint x of ℓ on the x-axis. There is a unique such semicircle that also intersects ℓ' orthogonally, the semicircle that goes through the point at which a line through x intersects ℓ' tangentially.

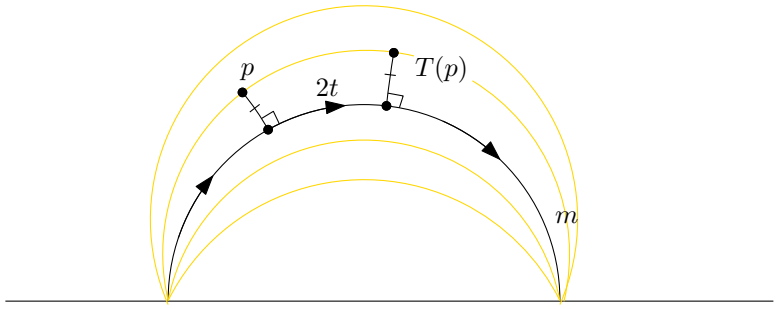


In general, ℓ may be a semicircle. If R is a reflection through a hyperbolic line centered at one of the endpoints of ℓ on the x-axis, $R(\ell)$ is a vertical half-line. From above, there is a unique hyperbolic line m that is perpendicular to $R(\ell)$ and $R(\ell')$, and then $R(m)$ is the hyperbolic line perpendicular to ℓ and ℓ'. □

Under the hypotheses of the lemma, Corollary 3.5.2 implies that $R_{\ell'}$ and R_ℓ both preserve m, so $f(m) = m$. Also, one can see from the following picture that the distance between a point $p \in m$ and $f(p)$ is twice the hyperbolic distance t between the intersection points $m \cap \ell$ and $m \cap \ell'$. So, f acts as a translation on m by $2t$. We call f a *hyperbolic translation* and m the *axis* of f.



The image of a general point $p \in \mathbb{H}^2$ under f is pictured above: drop a hyperbolic perpendicular to m, travel a distance of $2t$ along m, then proceed perpendicularly the same distance as in the first step to $T(p)$. The reason the picture is accurate is that f translates by $2t$ along m, preserves angles and takes hyperbolic line segments to hyperbolic line segments with the same length.



EXAMPLE 3.9.8. Our $f = R_{\ell'} \circ R_\ell$ is particularly easy to describe when ℓ and ℓ' are concentric semicircles. For instance, suppose ℓ, ℓ' are centered at the origin and have radii r, r'. By Exercise 3.4.9, the composition is the dilation around 0 by r'/r:

$$f(p) = \frac{r'}{r}\, p.$$

When ℓ, ℓ' are semicircles with radii 1 and λ centered at the origin, $R_{\ell'} \circ R_\ell$ is a translation along the vertical half-line defined by $x = 0$, and $t = 2\ln(\lambda)$.

3.10. The disc model

Our description of $\mathbb{H}^2$ as the upper half plane in $\mathbb{R}^2$ is convenient in its simplicity, but just as different maps of the earth are useful for different purposes, there is another model of $\mathbb{H}^2$ that is better suited to understanding hyperbolic circles.

Let $D = \{(x, y) \in \mathbb{R}^2 \,|\, x^2 + y^2 < 1\}$ be the open unit disc in $\mathbb{R}^2$. The *hyperbolic length* of a path $\gamma : [a, b] \longrightarrow D$ is defined as

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

and the *distance* between points $p, q \in D$ is the infimum

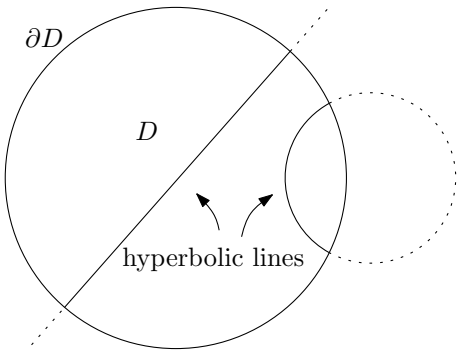
$$d_D(p, q) = \inf\big\{ \text{length}_D(\gamma) \mid \gamma : [a, b] \longrightarrow \mathbb{H}^2,\ \gamma(a) = p,\ \gamma(b) = q\big\}.$$

So, hyperbolic distances near $(x, y) \in D$ are distorted by the factor $\frac{2}{1-(x^2+y^2)}$, just as distances near $(x, y) \in \mathbb{H}^2$ were distorted by $1/y$.

The distortion factor $\frac{2}{1-(x^2+y^2)}$ goes to infinity when (x, y) approaches

$$\partial D = \{(x, y) \in \mathbb{R}^2 \,|\, x^2 + y^2 = 1\},$$

and is smallest at the center of D. So, the shortest path between two points $p, q \in D$ should bend toward the center to take advantage of the fact that distances are smaller there. In fact, we'll see that shortest paths in D lie along lircles orthogonal to ∂D, which will play the role of hyperbolic lines in the disc model.



EXERCISE 3.10.1. Show that the center of a circle orthogonal to ∂D lies outside of D.

The following theorem says that D and $\mathbb{H}^2$ are different representations of the same geometric object. So, results about the hyperbolic plane can be proved using either the D model or using $\mathbb{H}^2$, whichever is most convenient.

THEOREM 3.10.2. *Let C be the circle in $\mathbb{R}^2$ centered at $(0,-1)$ with radius $\sqrt{2}$. The inversion through C restricts to a bijection $i_C : D \longrightarrow \mathbb{H}^2$ with*

$$\text{length}_{\mathbb{H}^2}(i_C \circ \gamma) = \text{length}_D(\gamma)$$

for every path γ in D. Consequently, $d_{\mathbb{H}^2}(i_C(p), i_C(q)) = d_D(p,q)$ for all $p, q \in D$.

PROOF OF THEOREM 3.10.2. By definition, the inversion i_C can be expressed as

$$\begin{aligned} i_C(x,y) &= (0,-1) + \frac{(\sqrt{2})^2}{|(x,y)-(0,-1)|^2}\big((x,y)-(0,-1)\big) \\ &= \left(\frac{2x}{x^2+(y+1)^2}, \frac{2(y+1)}{x^2+(y+1)^2} - 1\right) \\ &= \left(\frac{2x}{x^2+(y+1)^2}, \frac{2(y+1)-x^2-(y+1)^2}{x^2+(y+1)^2}\right) \\ &= \left(\frac{2x}{x^2+(y+1)^2}, \frac{1-x^2-y^2}{2} \cdot \frac{2}{x^2+(y+1)^2}\right) \\ &= \left(\frac{2x}{x^2+(y+1)^2}, \frac{1-|(x,y)|^2}{2} \cdot \frac{2}{|(x,y)-(0,-1)|^2}\right). \end{aligned} \tag{16}$$

If $\gamma : [a,b] \longrightarrow D$, we have by Theorem 3.4.4 and (16) that

$$\begin{aligned} \text{length}_{\mathbb{H}^2}(i_C \circ \gamma) &= \int_a^b |(i_C \circ \gamma)'(t)| \frac{1}{(i_C \circ \gamma)_2(t)}\, dt \\ &= \int_a^b \left(\frac{2}{|\gamma(t)-(0,-1)|^2}\,|\gamma'(t)|\right)\left(\frac{2}{1-|\gamma(t)|^2} \cdot \frac{|\gamma(t)-(0,-1)|^2}{2}\right) dt \\ &= \int_a^b |\gamma'(t)| \frac{2}{1-|\gamma(t)|^2}\, dt, \\ &= \text{length}_D(\gamma). \end{aligned}$$

□

Using the theorem, we can transfer all our terminology and results about the upper half plane to D. First, the inversion i_C takes ∂D to the x-axis, and since inversions are conformal and send lircles to lircles, i_C must take lircles orthogonal to ∂D to lircles orthogonal to the x-axis, and vice versa. So, informed by the upper half plane, we call the intersections with D of lircles orthogonal to ∂D *hyperbolic lines* in D. This terminology makes sense, since the theorem, along with our characterization of

hyperbolic lines in $\mathbb{H}^2$ as shortest paths, implies that shortest paths in D do indeed lie along hyperbolic lines in D.

If ℓ is a hyperbolic line in D, the *hyperbolic reflection* through ℓ is the map

$$R_\ell : D \longrightarrow D$$

that is either a Euclidean reflection or an inversion through ℓ, depending on whether ℓ is the intersection with D of a line or a circle. By Exercise 3.4.11, we have that $R_\ell = i_C \circ R_{i_C(\ell)} \circ i_C$, where C is as in Theorem 3.10.2. Since $i_C : D \longrightarrow \mathbb{H}^2$ and $R_{i_C(\ell)} : \mathbb{H}^2 \longrightarrow \mathbb{H}^2$ are both isometries, the latter by Theorem 3.6.8, so is R_ℓ. In other words, *the reflection through a hyperbolic line in D is an isometry.*

Exercise 3.10.3. Prove directly that inversions through lircles orthogonal to ∂D are isometries of D, using Theorem 3.4.4 and mimicking the proof for the upper half plane.

Example 3.10.4. What is the hyperbolic distance in D from 0 to a point $p \in D$? We know that the shortest path γ is along the hyperbolic line through 0 and p, which in this case is a Euclidean line. Parametrically, we have $\gamma : [0,1] \longrightarrow D, \gamma(t) = tp$, so

$$\begin{aligned}\text{length}_D(\gamma) &= \int_0^1 |\gamma'(t)| \, \frac{2}{1-|\gamma(t)|^2} \, dt \\ &= \int_0^1 \frac{2|p|}{1-t^2|p|^2} \, dt \\ &= 2\tanh^{-1}(|p|).\end{aligned}$$

Above, the last line follows from a substitution and the fact that the integral of $1/(1-t^2)$ is $\tanh^{-1}$, which you can verify as an exercise.

Exercise 3.10.5. Show that $(\tanh^{-1})'(t) = \frac{1}{1-t^2}$. *Hint: first compute* $\tanh'(t)$*, then differentiate* $x = \tanh \circ \tanh^{-1}(x)$ *using the chain rule.*

So, we have shown that for $p \in D$, we have

$$d_D(0,p) = 2\tanh^{-1}(|p|). \tag{17}$$

Note that hyperbolic tangent goes to 1 as the input goes to infinity, so inverting, $\tanh^{-1}(|p|) \longrightarrow \infty$ as $|p| \longrightarrow 1$, i.e. as p approaches ∂D.

3.11. Hyperbolic circles

Circles in hyperbolic geometry are defined in the same way as in $\mathbb{R}^2$ and on the sphere, as a locus of points equidistant from a center. Namely, if $p \in \mathbb{H}^2$ and $r > 0$, the *hyperbolic circle* centered at p with radius r is

$$C(p,r) = \{q \in \mathbb{H}^2 \mid d_{\mathbb{H}^2}(p,q) = r\}.$$

Because of the curious form of the hyperbolic metric, it's not immediately clear what form a hyperbolic circle should take. However, Equation (17) implies:

FACT 3.11.1. *In the disc model, the Euclidean circle of radius r centered at the origin is a hyperbolic circle of radius $2\tanh^{-1}(r)$ centered at the origin. Conversely, a hyperbolic circle of radius r around 0 is a Euclidean circle of radius $\tanh\left(\frac{r}{2}\right)$.*

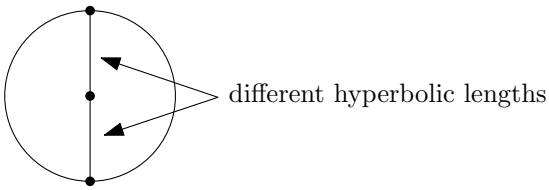
This is actually a general phenomenon!

COROLLARY 3.11.2. *In both the disc model and the upper half plane model, hyperbolic circles in $\mathbb{H}^2$ are Euclidean circles, but the Euclidean centers and radii are not always the same as their hyperbolic analogues.*

PROOF. When $p \in D$ is an arbitrary point, let ℓ be the hyperbolic line that perpendicularly bisects the hyperbolic line segment from 0 to p. As in Exercise **??**, $R_\ell(0) = p$. Since R_ℓ is an isometry, $C(p,r) = R_\ell\big(C(0,r)\big)$. And since R_ℓ is either an inversion through a circle (with center outside D, by Exercise 3.10.1) or a Euclidean reflection, it takes circles in D to circles, so $C(p,r)$ is a Euclidean circle!

To deduce the result for the upper half plane, let C be the circle in the statement of Theorem 3.10.2. When p is a point in the upper half plane, $i_C(p) \in D$, and as i_C is an isometry from D to the upper half plane, it takes hyperbolic circle of radius r around $i_C(p)$ (which is a Euclidean circle in D) to the hyperbolic circle $C(p,r)$ in the upper half plane. But i_C take circles to circles, so $C(p,r)$ is a Euclidean circle. □

The hyperbolic and Euclidean centers of circles in the upper half plane never agree. Because of the metric distortion, the Euclidean center of a circle C is always hyperbolically closer to the point directly above it on C than to the point below it on C.

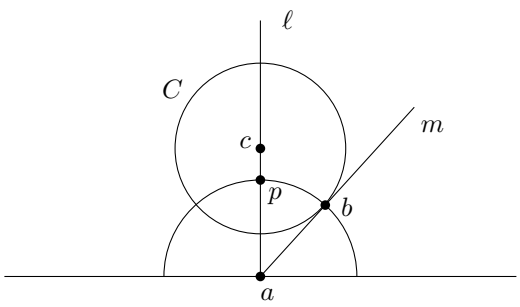


So, where is the hyperbolic center of a circle in the upper half plane?

LEMMA 3.11.3. *If C is a hyperbolic circle, a hyperbolic line ℓ passes through the (hyperbolic) center of C if and only if it is orthogonal to C.*

PROOF. A hyperbolic line ℓ passes through the center of C if and only if R_ℓ fixes C, which by Corollary 3.5.2 happens if and only if ℓ is orthogonal to C. □

This lemma can be used to give a ruler and compass construction of the hyperbolic center of a circle C in the upper half plane with Euclidean center c. Let a be the point at which the vertical line ℓ through c hits the x-axis, and let b be a point at which a line m through a is tangent to C. By the lemma, both ℓ and m pass through the hyperbolic center p of C, which must be their point of intersection.



EXERCISE 3.11.4. Let C be a Euclidean circle in $\mathbb{H}^2$ whose highest and lowest points are at heights a and b. Then the Euclidean center is at height $\frac{a+b}{2}$, the arithmetic mean. Show that the hyperbolic center is at height $\sqrt{ab}$, the geometric mean. *Hint: the ruler and compass construction is not the most efficient way to do this.*

So, the inequality of means (Exercise 1.9.9) is expressing the fact that hyperbolic distance becomes smaller relative to Euclidean distance when height is increased!

QUESTION. What is the circumference of a hyperbolic circle of radius r?

Implicit in this question is the statement is that the circumference of a hyperbolic circle should not depend on the center of the circle, just the radius. However, if $C(p,r)$ and $C(q,r)$ are hyperbolic circles with the same radius r, we've seen that there is a hyperbolic reflection R_ℓ such that $R_\ell(p) = q$, and then $R_\ell(C(p,r)) = C(q,r)$. As R_ℓ preserves hyperbolic path lengths, $\text{length}_{\mathbb{H}^2} C(q,r) = \text{length}_{\mathbb{H}^2} C(p,r)$.

So, it suffices to do the circumference computation for our favorite hyperbolic circle of radius r. My favorite is the one centered at the origin in the disc model, which Fact 3.11.1 says is a Euclidean circle with radius $\tanh\frac{r}{2}$. Parametrically, this is

$$\gamma : [0, 2\pi] \longrightarrow D, \quad \gamma(t) = \tanh\frac{r}{2}\,(\cos t, \sin t),$$

and we compute length as follows:

$$\begin{aligned}
\text{length}_{\mathbb{H}^2}(\gamma) &= \int_0^{2\pi} \frac{2|\gamma'(t)|}{1-|\gamma(t)|^2}\,dt \\
&= \int_0^{2\pi} \frac{2\tanh\frac{r}{2}}{1-\left(\tanh\frac{r}{2}\right)^2}\,dt \\
&= \frac{4\pi\tanh\frac{r}{2}}{1-\left(\tanh\frac{r}{2}\right)^2} \\
&= 4\pi\tanh\frac{r}{2}\left(\cosh\frac{r}{2}\right)^2, \qquad \text{since } 1-\tanh^2 = \text{sech}^2, \\
&= 4\pi\sinh\frac{r}{2}\cosh\frac{r}{2} \\
&= 2\pi\sinh r.
\end{aligned}$$

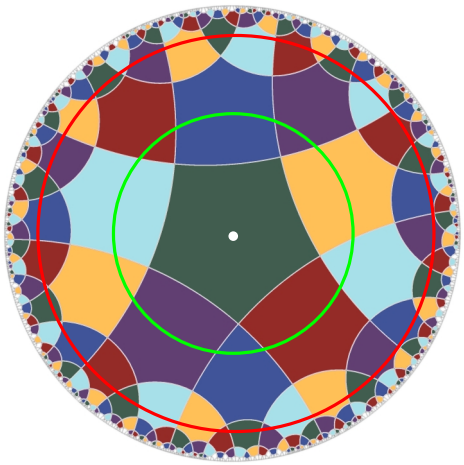
The last line is the hyperbolic double angle formula. You might try proving it and the identity $1 - \tanh^2 = \operatorname{sech}^2$ as a refresher on hyperbolic trigonometric functions.

In conclusion, we have now shown:

THEOREM 3.11.5. *The circumference of a hyperbolic circle of radius r is $2\pi \sinh r$.*

Note the difference between this and $2\pi r$, the Euclidean formula for the circumference of a radius r circle. As $\sinh(r) = \frac{1}{2}(e^r - e^{-r})$, the circumference of a hyperbolic circle grows *exponentially* as the radius increases. For comparison, the circumference of a Euclidean circle of radius 10 is $2\pi 10 \approx 62.8$, while the circumference of a hyperbolic circle of radius 10 is $2\pi \sinh(10) \approx 69,198$.

There's a nice way to see the exponential growth of circumference visually using monohedral tilings. The following is a tiling of the hyperbolic plane, in the disc model, by right angled hyperbolic pentagons. We will discuss hyperbolic polygons in more detail later, but note that it is plausible that all these pentagons are congruent, since the small polygons near ∂D are hyperbolically much larger than they appear.



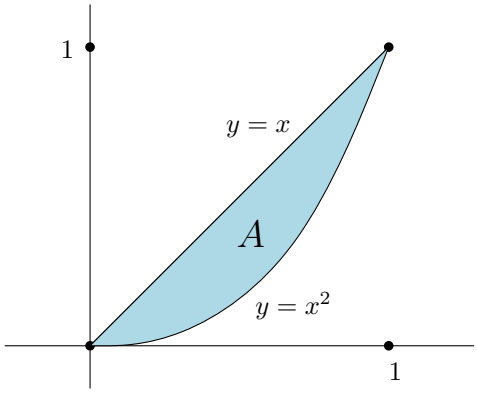
Superimposed on the image are two hyperbolic circle centered at the origin, in green and in red. The hyperbolic radius of the red circle is roughly 'one pentagon', while that of the green circle is around two pentagons. One can estimate the circumferences these circles by counting the number of polygons traversed. The green circle goes through 9 pentagons, while the red circle goes through 42. Each time the hyperbolic radius is increased by another pentagon, the circumference is multiplied by a factor vaguely around 4, which gives an exponential growth rate.

The large circumference of circles also explains the buckling you see after assembling a pseudo-sphere. If a loop of radius 10 in $\mathbb{R}^3$ is to have circumference $69,198$, the only option is to wave back and forth enough times to pick up the extra length.

3.12. Hyperbolic area

The goal of this section is to develop a theory of *hyperbolic area*. First, however, let's review some facts about Euclidean area in $\mathbb{R}^2$.

EXAMPLE 3.12.1. Let A be the region in the plane enclosed by the graphs of $y = x$ and $y = x^2$, as pictured below. What is the area of A?



Recall from calculus that the integral of a function is the area under its graph. So, to calculate the area of A we subtract two integrals:

$$\begin{aligned}\text{Area}(A) &= \int_0^1 x\,dx - \int_0^1 x^2\,dx \\ &= \frac{x^2}{2} - \frac{x^3}{3}\Big|_0^1 \\ &= \frac{1}{6}.\end{aligned}$$

The integral above can be rewritten as follows.

$$\text{Area}(A) = \int_0^1 x\,dx - \int_0^1 x^2\,dx = \int_0^1 x - x^2\,dx = \int_0^1 \int_{x^2}^x 1\,dy\,dx = \int\int_A 1$$

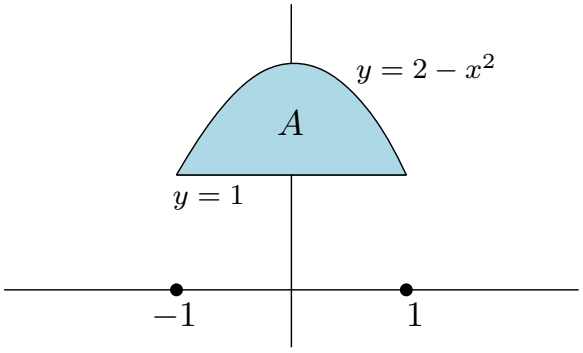
Recall: If $A \subset \mathbb{R}^2$ and $f : A \longrightarrow \mathbb{R}$ is a function, the *integral of f over A*, written

$$\int\int_A f,$$

is the double integral of f where the limits of integration are set to describe A.

So, the point is that Euclidean area may be calculated by integrating the constant 1 over the region in question, using a double integral whose limits describe A. It is best to see what we mean by 'describe A' in an example.

EXAMPLE 3.12.2. Let A be the region of $\mathbb{R}^2$ bounded by the line $y = 1$ and the graph of $y = 2 - x^2$, and let $f(x, y) = 2x$.



$$\begin{aligned}\iint_A f = \int_{-1}^{1}\int_{1}^{2-x^2} 2x\,dy\,dx &= \int_{-1}^{1} 2xy\Big|_{y=1}^{y=2-x^2}\,dx \\ &= \int_{-1}^{1} 2x(2-x^2) - 2x\,dx \\ &= \int_{-1}^{1} 2x - 2x^3\,dx \\ &= x^2 - \frac{1}{2}x^4\Big|_{-1}^{1} \\ &= 0.\end{aligned}$$

Here, the limits of integration describe A since a point $(x, y) \in A$ if and only if $-1 \leqslant x \leqslant 1$ and $1 \leqslant y \leqslant 2 - x^2$. So, in setting up the double integral we want the limits of integration to impose constraints on x and y equivalent to '$(x, y) \in A$'.

DEFINITION 3.12.3. If $A \subset \mathbb{H}^2$, the *hyperbolic area* of A is the integral

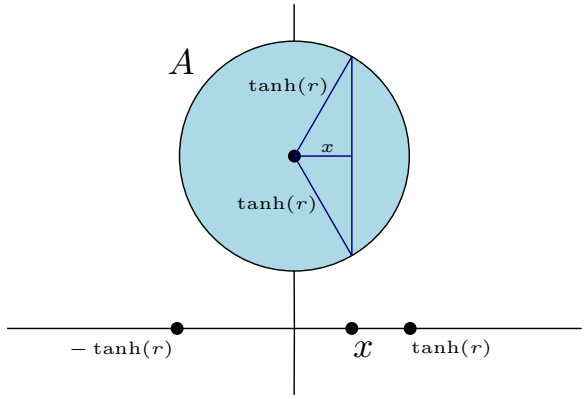
$$\text{Area}_{\mathbb{H}^2}(A) = \iint_A \frac{1}{y^2}.$$

The motivation behind this definition is that if the discrepancy between hyperbolic and Euclidean distance at a point (x, y) is a factor of $1/y$, then hyperbolic and Euclidean area should differ by a factor of $1/y^2$. For example, the area of a square with the side lengths t is t^2, but if the square is scaled by $1/y$, then the side lengths are t/y and the area is t^2/y^2.

EXAMPLE 3.12.4. Let A be the disc enclosed by the circle in $\mathbb{H}^2$ with Euclidean center at $(0, 1)$ and Euclidean radius $\tanh(r)$. We saw in Section 3.11 that the hyperbolic radius of this circle is r.

Using the Pythagorean theorem to calculate the limits of integration, we have

$$\iint_A 1 = \int_{-\tanh(r)}^{\tanh(r)} \int_{1-\sqrt{\tanh(r)^2-x^2}}^{1+\sqrt{\tanh(r)^2-x^2}} 1/y^2\,dy\,dx$$



$$= \int_{-\tanh(r)}^{\tanh(r)} \frac{-1}{1+\sqrt{\tanh(r)^2 - x^2}} + \frac{1}{1-\sqrt{\tanh(r)^2 - x^2}}\, dx,$$

which after substituting $c = \tanh(r)$ and integrating

$$= \frac{2}{\sqrt{1-c^2}} \tan^{-1}\left(\frac{x}{(\sqrt{1-c^2}\sqrt{c^2-x^2}}\right) - 2\tan^{-1}\left(\frac{x}{\sqrt{c^2-x^2}}\right)\Bigg|_{x\to -c}^{x\to c}. \quad (18)$$

As $x \to \pm c$, the two ratios inside the inverse tangents tend to $\pm\infty$. However, $\lim_{t\to\pm\infty} \tan^{-1}(t) = \pm\frac{\pi}{2}$, so we have

$$\begin{aligned}(18) &= \frac{2}{\sqrt{1-c^2}}\pi - 2\pi \\ &= \frac{2}{\sqrt{\operatorname{sech}^2(r)}}\pi - 2\pi \\ &= 2\pi(\cosh(r) - 1).\end{aligned}$$

In Section 3.11, we saw that all hyperbolic circles with radius r are congruent, so they all have the same circumference. We would like to say the same about area.

THEOREM 3.12.5. *If $f : \mathbb{H}^2 \longrightarrow \mathbb{H}^2$ is an isometry and $A \subset \mathbb{H}^2$, then* $\text{Area}_{\mathbb{H}^2}(f(A)) = \text{Area}_{\mathbb{H}^2}(A)$. *In other words, congruent subsets of $\mathbb{H}^2$ have the same area.*

As all hyperbolic circles of radius r are congruent, this implies:

COROLLARY 3.12.6. *The area of any disc of radius r in $\mathbb{H}^2$ is $2\pi(\cosh(r) - 1)$.*

Imagine a disc as being composed of all the concentric circles around its center. As the radius of the disc increases, we add in more concentric circles to the boundary of the disc. So, you might imagine that the derivative of the area of the disc of radius r is the circumference of its boundary circle.

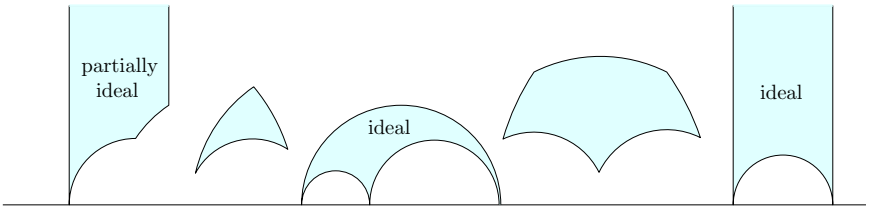
We can verify this directly in both hyperbolic and Euclidean geometry:

$$\frac{d}{dr}2\pi(\cosh(r)-1) = 2\pi\sinh(r), \qquad \frac{d}{dr}\pi r^2 = 2\pi r.$$

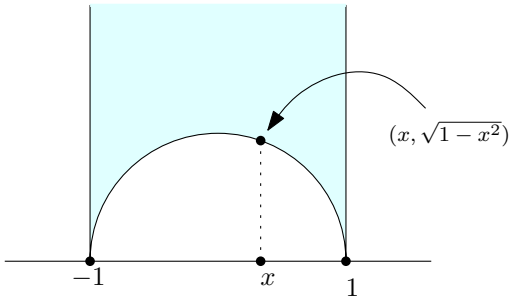
3.13. Hyperbolic polygons and their areas

A *hyperbolic n-gon* P is a subset of $\mathbb{H}^2$ bounded by n hyperbolic line segments (or full lines, or rays) $\ell_1, \ldots, \ell_n$, called the *sides of* T, such that each ℓ_i shares an endpoint with ℓ_{i+1} and its other endpoint with ℓ_{i-1}. (When $i = 1$ or $i = n$, we interpret $i-1$ and $i+1$ as n and 1, respectively.) These endpoints are called the *vertices* of T.

We will allow vertices to lie in $\partial\mathbb{H}^2 = x$–axis $\cup\, \infty$. These vertices are called *ideal*, and are not included in P since they do not lie in $\mathbb{H}^2$. Note that the interior angle at any ideal vertex is 0. If all the vertices of P are ideal, then P is an *ideal polygon*.



Example 3.13.1. Let's calculate the area of the ideal hyperbolic triangle T that has endpoints $0, 1, \infty$ in $\partial\mathbb{H}^2$.



$$\begin{aligned}
\int\int_T 1 &= \int_{-1}^{1}\int_{\sqrt{1-x^2}}^{\infty} \frac{1}{y^2}\,dy\,dx \\
&= \int_{-1}^{1} \frac{-1}{\sqrt{1-x^2}}\,dx \\
&= \int_{0}^{\pi} \frac{1}{\sqrt{1-\sin^2\theta}}\cos(\theta)\,d\theta \\
&= \int_{0}^{\pi} 1\,d\theta \\
&= \pi.
\end{aligned}$$

EXERCISE 3.13.2. All ideal hyperbolic triangles are congruent.

Therefore, since congruent regions have the same area, we have:

THEOREM 3.13.3. *Any ideal hyperbolic triangle in $\mathbb{H}^2$ has area π.*

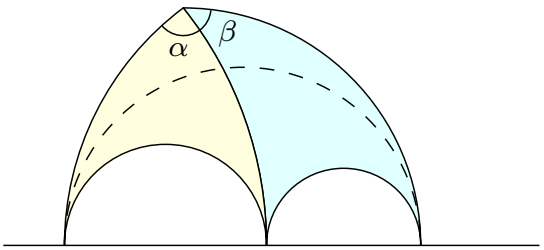
What about the areas of nonideal hyperbolic triangles? We will tackle this problem first for triangles with two vertices on $\partial\mathbb{H}^2$ and then for arbitrary triangles.

LEMMA 3.13.4. *Any two hyperbolic triangles that both have two ideal vertices and one vertex with angle α are congruent.*

PROOF. Suppose that T_1 and T_2 are two such triangles and that their nonideal vertices are p_1 and p_2. Pick an isometry R of $\mathbb{H}^2$ such that $R(p_1) = p_2$. Then $R(T_1)$ is a triangle with one nonideal vertex at p_2 with angle α. There is then a hyperbolic rotation O around p_2 such that $O \circ R(T_1) = T_2$. □

With the lemma in mind, we now define $A(\alpha)$ to be the area of any triangle that has two ideal vertices and one vertex with angle α. You can, of course, calculate $A(\alpha)$ directly by setting up an appropriate integral. However, here is another method.

From the picture below, we see that the union of the yellow and blue triangles can be expressed as the union of a triangle with angle $\alpha + \beta$ and an ideal triangle.



Therefore, $A(\alpha) + A(\beta) = A(\alpha + \beta) + \pi$. Differentiating, we obtain that

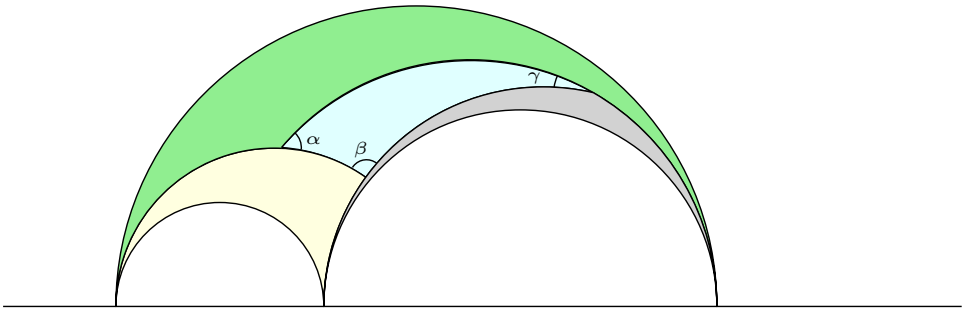
$$\frac{d}{d\alpha}A(\alpha) = \frac{d}{d\alpha}A(\alpha + \beta),$$

since $A(\beta)$ and π are constants when we differentiate in α. But this means that $A'(\alpha)$ is constant, so A is a linear function! We also know that $A(0) = \pi$ and $A(\pi) = 0$, since setting the final vertex to have angle 0 or π gives an ideal triangle or a hyperbolic line, respectively. So, if $A(\alpha) = c\alpha + d$ we must have

$$c \cdot 0 + d = \pi \quad \text{and} \quad c \cdot \pi + d = 0, \quad \Longrightarrow \quad c = -1,\ d = \pi.$$

THEOREM 3.13.5. *The area of a hyperbolic triangle with two ideal vertices and one vertex with angle α is $\pi - \alpha$.*

Now let T be an arbitrary triangle in $\mathbb{H}^2$ and denote the angles of T by α, β, γ. Then there is an ideal triangle that is the union of T with three triangles that have angles $\pi - \alpha, \pi - \beta, \pi - \gamma$, respectively, in addition to two ideal vertices.



Therefore, $\text{Area}(T) = \pi - (\pi - (\pi - \alpha)) - (\pi - (\pi - \beta)) - (\pi - (\pi - \gamma)) = \pi - \alpha - \beta - \gamma$.

We have now proven the following theorem.

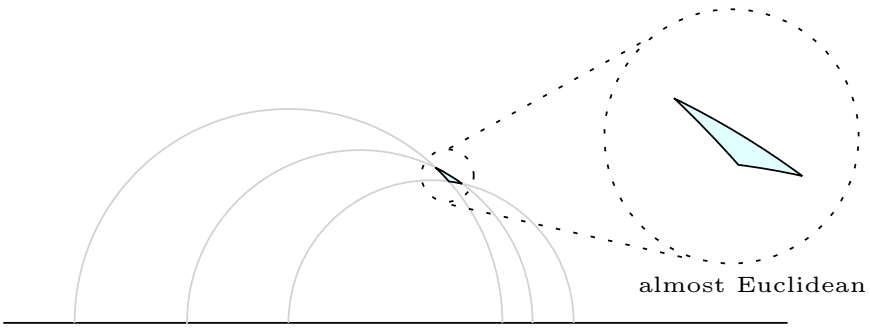
THEOREM 3.13.6. *The area of a hyperbolic triangle with angles α, β, γ is $\pi - \alpha - \beta - \gamma$.*

As in the Euclidean and spherical cases, by dividing an arbitrary polygon into triangles and applying the previous theorem, one can prove the following corollary.

COROLLARY 3.13.7. *If P is a hyperbolic n-gon with angles $\alpha_1, \ldots, \alpha_n$, then*

$$\text{Area}(P) = \pi(n-2) - \sum_{i=1}^{n} \alpha_i.$$

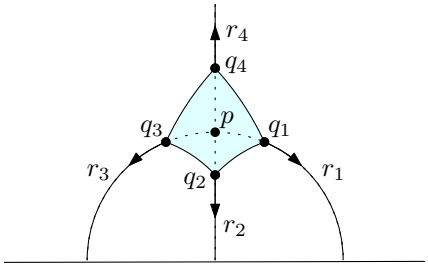
The term $\pi(n-2)$ is the interior angle sum of a Euclidean n-gon. As area is positive, the corollary implies that the interior angle sum of a hyperbolic n-gon is always strictly between 0 and the Euclidean angle sum, and that the smaller the polygon is, the closer the angle sum is to being Euclidean. This should conform to your intuition: the sides of an extremely small triangle in $\mathbb{H}^2$ look nearly straight, so in some sense the entire polygon should be nearly Euclidean.



A hyperbolic n-gon is *regular* if all its sides have the same length and all its angles are equal. The corollary above implies that the interior angles of a regular hyperbolic n-gon lie between 0 and $\frac{\pi(n-2)}{n}$; note that $\frac{\pi(n-2)}{n}$ is the interior angle of a regular Euclidean n-gon. In fact, there are regular hyperbolic n-gons with any such angle.

THEOREM 3.13.8. *If $\alpha \in \left(0, \frac{\pi(n-2)}{n}\right)$, there is a regular hyperbolic n-gon with all angles equal to α. Consequently, there are regular hyperbolic n-gons with interior angle sum equal to any number in $(0, \pi(n-2))$.*

PROOF. Pick a point p in $\mathbb{H}^2$ and let $r_1, \ldots, r_n$ be hyperbolic rays starting at p such that the angle between r_i and r_{i+1} is $2\pi/n$ for all i.



a regular hyperbolic quadrilateral

Given $t > 0$, let $q_1, \ldots, q_n$ be the unique points on the rays $r_1, \ldots, r_n$ such that $d(p, q_i) = t$ for all i. Then $q_1, \ldots, q_n$ form the vertices of a regular hyperbolic n-gon P_t, for the rotation around p by angle $2\pi/n$ is a symmetry of this polygon and by rotating an appropriate number of times we see that any two sides lengths or angles of P_t are the same. As $t \to 0$, the area of P_t goes to 0, so the angles of P_t converge to $\frac{\pi(n-2)}{n}$. As $t \to \infty$, the polygon P_t converges to an ideal polygon, so the angles of P_t converge to 0. By the intermediate value theorem, when we vary t from 0 to ∞, the angles of P_t take on every value between 0 and $\frac{\pi(n-2)}{n}$. □

Bibliography

[1] Robin Hartshorne. *Geometry: Euclid and beyond*. Springer Science & Business Media, 2013.
[2] Thomas Little Heath, Dana Densmore, et al. *Euclid's Elements*. 2003.
[3] John E Reeve. On the volume of lattice polyhedra. *Proceedings of the London Mathematical Society*, 3(1):378–395, 1957.