## 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 $x=\left(x_{1}, \ldots, x_{n}\right)$, 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

$$
d(p, q)=\sqrt{\left(p_{1}-q_{1}\right)^{2}+\left(p_{2}-q_{2}\right)^{2}}
$$

applying the Pythagorean theorem to the right triangle with vertices $p, q$ and $\left(q_{1}, p_{2}\right)$. 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

$$
d(p, q)=\sqrt{\left(p_{1}-q_{1}\right)^{2}+\cdots+\left(p_{n}-q_{n}\right)^{2}} .
$$

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

$$
d\left(p,\left(q_{1}, \ldots, q_{n-1}, p_{n}\right)\right)=\sqrt{\left(p_{1}-q_{1}\right)^{2}+\cdots+\left(p_{n-1}-q_{n-1}\right)^{2}}
$$

There is a right triangle with vertices $p, q$ and $\left(q_{1}, \ldots, q_{n-1}, p_{n}\right)$, 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) $\quad v \cdot(w+u)=v \cdot w+v \cdot u$,
- (Scalars come out) $v \cdot(r w)=r(v \cdot w)$, 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 $\theta$, 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

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

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

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

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

$$
\begin{aligned}
v \star w & =\left(\sum_{i} v_{i} e_{i}\right) \star\left(\sum_{j} w_{j} e_{j}\right) \\
& =\sum_{i, j}\left(v_{i} e_{i}\right) \star\left(w_{j} e_{j}\right), \quad \text { by distributivity, } \\
& =\sum_{i, j} v_{i} w_{j} e_{i} \star e_{j}, \quad \text { by pulling out scalars, } \\
& =\sum_{i} v_{i} w_{i}, \quad \text { by Equation }(1), \\
& =v \cdot w .
\end{aligned}
$$

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

$$
\operatorname{proj}_{v}(w)=w \cdot v \frac{v}{|v|^{2}}
$$

Show that $w-\operatorname{proj}_{v}(w)$ is perpendicular to $v$. Then show that $\operatorname{proj}_{v}(w)$ is the closest point to $w$ on the line $\{t v \mid t \in \mathbb{R}\}$. 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 $\theta$. Then $c^{2}=a^{2}+b^{2}-2 a b \cos \theta$.

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

$$
|c|^{2}=c \cdot c=(a-b) \cdot(a-b)=a \cdot a-2 a \cdot b+b \cdot b=|a|^{2}+|b|^{2}-2|a||b| \cos \theta .
$$



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 $\theta$ be the angle at $q$ in the triangle $p q z$. By the law of cosines,

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

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. $d(p, q)=\sqrt{\left(p_{1}-q_{1}\right)^{2}+\cdots+\left(p_{n}-q_{n}\right)^{2}}$ defines a metric on $\mathbb{R}^{n}$.

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

$$
d(p, q)=\left(p_{1}-q_{1}\right)^{2}+\ldots+\left(p_{n}-q_{n}\right)^{2}
$$

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

$$
\begin{aligned}
d_{\text {max }}(p, q) & =\max \left\{\left|p_{1}-q_{1}\right|, \cdots,\left|p_{n}-q_{n}\right|\right\}, \\
d_{\text {sum }}(p, q) & =\left|p_{1}-q_{1}\right|+\cdots+\left|p_{n}-q_{n}\right|
\end{aligned}
$$

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 $d_{\text {sum }}(p, q)=\left|p_{1}-q_{1}\right|+\left|p_{2}-q_{2}\right|$ 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 $2 \times 2$ matrix $(v w)$ is

$$
\operatorname{det}\left(\begin{array}{ll}
v_{1} & w_{1} \\
v_{2} & w_{2}
\end{array}\right)=v_{1} w_{2}-w_{1} v_{2}
$$

Note that $v_{1} w_{2}-w_{1} v_{2}=v^{\perp} \cdot w$, where $v^{\perp}=\left(-v_{2}, v_{1}\right)$. 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 $\pi$ if and only if $\operatorname{det}(v w) \geqslant 0$.
(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 $|\operatorname{det}(v w)|$.


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 $\theta$, 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 $i d: X \longrightarrow$ $X$ 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

$$
d\left(T_{v}(x), T_{v}(y)\right)=|x+v-(y+v)|=|x-y|=d(x, y)
$$

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 $\theta$ 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

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

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

$$
O_{0, \theta}\binom{x_{1}}{x_{2}}=\left(\begin{array}{cc}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta
\end{array}\right)\binom{x_{1}}{x_{2}}=\binom{x_{1} \cos \theta-x_{2} \sin \theta}{x_{1} \sin \theta+x_{2} \cos \theta} .
$$

To see why this is true, first note that

$$
\begin{aligned}
d\left(0, O_{0, \theta}(x)\right) & =\left|\binom{x_{1} \cos \theta-x_{2} \sin \theta}{x_{1} \sin \theta+x_{2} \cos \theta}\right| \\
& =\sqrt{\left(x_{1} \cos \theta-x_{2} \sin \theta\right)^{2}+\left(x_{1} \sin \theta+x_{2} \cos \theta\right)^{2}} \\
& =\sqrt{\left(x_{1}^{2}+x_{2}^{2}\right) \cos ^{2} \theta+\left(x_{1}^{2}+x_{2}^{2}\right) \sin ^{2} \theta} \\
& =\sqrt{x_{1}^{2}+x_{2}^{2}} \\
& =d(0, x) .
\end{aligned}
$$

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

$$
\begin{aligned}
\cos \psi & =\frac{x \cdot O_{p, \theta}(x)}{|x|\left|O_{p, \theta}(x)\right|} \\
& =\frac{x_{1}\left(x_{1} \cos \theta-x_{2} \sin \theta\right)+x_{2}\left(x_{1} \sin \theta+x_{2} \cos \theta\right)}{\sqrt{x_{1}^{2}+x_{2}^{2}} \sqrt{x_{1}^{2}+x_{2}^{2}}} \\
& =\cos \theta .
\end{aligned}
$$

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 $\theta$, then by angle $\psi$ 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:

$$
\left(\begin{array}{cc}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta
\end{array}\right)\left(\begin{array}{cc}
\cos \psi & -\sin \psi \\
\sin \psi & \cos \psi
\end{array}\right)=\left(\begin{array}{cc}
\cos \theta \cos \psi-\sin \theta \sin \psi & -\cos \theta \sin \psi-\sin \theta \cos \psi \\
\sin \theta \cos \psi+\cos \theta \sin \psi & \cos \theta \cos \psi-\sin \theta \sin \psi
\end{array}\right)
$$

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

$$
\begin{equation*}
O_{p, \theta}=T_{p} \circ O_{0, \theta} \circ T_{-p}, \tag{2}
\end{equation*}
$$

which implies that

$$
O_{p, \theta}(x)=\left(\begin{array}{cc}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta
\end{array}\right)(x-p)+p
$$

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 $\operatorname{Fix}(f)=\{x \in X \mid f(x)=x\}$. Show that

$$
\operatorname{Fix}\left(f \circ g \circ f^{-1}\right)=f(\operatorname{Fix}(g)),
$$

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 $\ell$ is a line in $\mathbb{R}^{2}$, the reflection through $\ell$ 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 $\ell$.

Exercise 1.2.9. Show that reflections are isometries.
When $\ell$ goes through the origin, we can write it as $\ell=\{t v \mid t \in \mathbb{R}\}$ for some $v \in \mathbb{R}^{2}$. In this case, the reflection $R_{\ell}$ has the following description in coordinates:

$$
R_{\ell}(x)=x-2 \operatorname{proj}_{v^{\perp}}(x), \text { where } v^{\perp}=\binom{-v_{2}}{v_{1}} .
$$

Recall from Exercise 1.1.4 that $\operatorname{proj}_{v^{\perp}}(x)=x \cdot v^{\perp} \frac{v^{\perp}}{\left|v^{\perp}\right|^{2}}$ 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 $\ell=\{p+t v \mid t \in \mathbb{R}\}$, in which case

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



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 $\theta$ be the angle from the segment $x y$ to the segment $x z$, and let $\psi$ 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 $\ell, \ell^{\prime}$ are parallel lines in $\mathbb{R}^{2}$. What's $R_{\ell^{\prime}} \circ R_{\ell}$ ? To get some evidence, let's pick some $x \in \mathbb{R}^{2}$ that lies on the far side of $\ell$ from $\ell^{\prime}$, and very close to $\ell$, and then compute $R_{\ell^{\prime}} \circ R_{\ell}(x)$.


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

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

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

To prove this, however, we have to show that $R_{\ell^{\prime}} \circ R_{\ell}(x)=T_{2 v}(x)$ for any $x \in \mathbb{R}^{2}$, not just the $x$ in the picture above. For instance, what happens if $x$ is on $\ell$ or $\ell^{\prime}$, or on the far side of $\ell^{\prime}$ ? 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 $R_{\ell^{\prime}} \circ R_{\ell}=T_{2 v}$, 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 $p q$. In particular, given $p, q$ all such $x$ are colinear.

Proof. Let $m$ be the midpoint of the segment $p q$. 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 $\pi$, 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 $f\left(x_{i}\right)=g\left(x_{i}\right)$ 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 $y_{i}=f\left(x_{i}\right)=g\left(x_{i}\right)$ for $i=1,2,3$. As both $f, g$ are isometries,

$$
d\left(f(x), y_{i}\right)=d\left(x, x_{i}\right)=d\left(g(x), y_{i}\right), \quad \text { for } i=1,2,3
$$

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 $\ell, \ell^{\prime}$ are parallel, we have $R_{\ell^{\prime}} \circ R_{\ell}=T_{2 v}$, where $v$ is a vector with its tail on $\ell$ and its head on $\ell^{\prime}$ that is perpendicular to both lines. Above, we showed that $R_{\ell^{\prime}} \circ R_{\ell}(x)=T_{2 v}(x)$ whenever $x$ is close to $\ell$, and on the opposite side of $\ell$ from $\ell^{\prime}$. But we can certainly find three such $x$ that are non-colinear, so $R_{\ell^{\prime}} \circ R_{\ell}=T_{2 v}$ 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 $\ell^{\prime}$ and $\ell$ intersect at a point $p$, the composition $R_{\ell^{\prime}} \circ R_{\ell}$ is the rotation $O_{p, \theta}$, where $\theta$ is twice the angle from $\ell$ to $\ell^{\prime}$. 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 $\ell$ is a line parallel to $v$. The composition $T_{v} \circ R_{\ell}$ is called the glide reflection along $\ell$ 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 $\ell$ is not parallel to $v$ ?

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

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

|  | $T_{v}$ | $O_{p, \theta}$ | $R_{\ell}$ | $T_{v} \circ R_{\ell}$ <br> $v \\| \ell$ |
| :---: | :---: | :---: | :---: | :---: |
| $T_{w}$ | id if $v=-w$ <br> otherwise translation | rotation | reflection if $\ell \perp w$ <br> otherwise glide | reflection if $v+w \perp \ell$ <br> otherwise glide |
| $O_{q, \psi}$ | rotation | id if $p=q$ and $\theta=2 \pi-\psi \psi$ <br> translation if $p \neq q$ and <br> $\theta=2 \pi-\psi$, o.w. rotation | reflection if $q \in \ell$ <br> otherwise glide | reflection or glide |
| $R_{\ell^{\prime}}$ | reflection if $v \perp \ell^{\prime}$ <br> otherwise glide | reflection if $p \in \ell$ <br> otherwise glide | id if $\ell=\ell^{\prime}$ <br> translation if $\ell \\| \ell^{\prime}, \ell^{\prime} \neq \ell$ <br> otherwise rotation | translation if $\ell^{\prime} \\| \ell$ <br> otherwise rotation |
| $T_{w} \circ R_{\ell^{\prime}}$ |  |  |  |  |
| $w \\| \ell^{\prime}$ |  |  |  |  | | reflection if $v+w \perp \ell^{\prime}$ |
| :---: |
| otherwise glide |$\quad$| reflection or glide |
| :---: |

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 $\ell$ 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 $f=i d$ 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

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

$f$ 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 $R_{a}(x)=2 a-x$.

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 $d_{\text {max }}(x, y)=\max _{i}\left\{\left|x_{i}-y_{i}\right|\right\}$. Show that all translations are $d_{\max }$-isometries, and that a rotation $O_{p, \theta}$ is a $d_{\max }$-isometry if and only if $\theta$ 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 ??.
(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 $\gamma(t)=\left(\gamma_{1}(t), \gamma_{2}(t)\right)$ is such a path, compare its length to that of the path $\alpha(t)=\left(\gamma_{1}(t), 0\right)$.
(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 $\lambda$. Dilations are similarities, since

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

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 $\S 1.2$, 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 $\ell$ and a translation parallel to $\ell$, 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.


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

$$
d\left(p, x_{i}\right)=d\left(q, x_{i}\right) \forall i=1, \ldots, 4 \Longrightarrow p=q .
$$

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

Exercise 1.3.3. Suppose $P$ and $P^{\prime}$ are planes in $\mathbb{R}^{3}$. Show that the composition of the reflections through $P$ and $P^{\prime}$ is a rotation around the line $\ell=P \cap P^{\prime}$ 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 $\ell$ and a plane $P$ pass through a point $p \in \mathbb{R}^{3}$. Show that the composition of a rotation around $\ell$ 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 $\ell$, show that $f$ is either the identity, a reflection through plane containing $\ell$, or a rotation around $\ell$. 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. The Chord Theorem

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

Theorem 1.4.1 (The Chord Theorem). If $C$ is a chord for $\gamma$ with length $a$, then for every $n=1,2, \ldots$, there is another chord for $\gamma$ with length $a / n$ that is parallel to $C$.


The proof will use the following lemma.
Lemma 1.4.2. Suppose $C$ is a chord for $\gamma$ with length $c$. Then for every $\alpha \in(0,1)$, there is a parallel chord either with length $\alpha c$ 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 $\gamma$, and $X_{\alpha}, X_{1}$, which are obtained by shifting the image of $\gamma$ to the right by $\alpha$ and 1 , respectively. It suffices to show that either $X \cap X_{\alpha} \neq \varnothing$ or $X_{\alpha} \cap X_{1} \neq \varnothing$, for $\gamma$ has a horizontal chord of length $\alpha$ exactly when $X_{0}$ intersects its translate $X_{\alpha}$, 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 $\beta$ by taking the part of $X_{\alpha}$ 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 $\beta$ is disjoint from $X_{0}$. First, $X_{0}$ cannot intersect the part of $\beta$ that lies along $X_{\alpha}$, since we assumed that $X_{0} \cap X_{\alpha}=\varnothing$. But $X_{\alpha}$ is a horizontal shift of $X_{0}$, so no point of $X_{0}$ can be higher than the highest point of $X_{\alpha}$, or lower than the lowest point of $X_{\alpha}$, so $X_{0}$ cannot intersect the other two parts of $\beta$. Similarly, $X_{1}$ is disjoint from $\beta$.

The path $\beta$ splits the plane into two pieces. We saw above that $\beta$ is disjoint from both $X_{0}$ and $X_{1}$. In fact, $X_{0}, X_{1}$ lie on different sides of $\beta$, 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 1.4.3. 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 $\gamma$ of length $a$, there is a parallel chord for $\gamma$ 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 $\alpha$, there is a path $\gamma$ with a chord of length $a$ that has no parallel chord with length $\alpha a$ ! Here is an explicit example.

ExERCISE 1.4.4. 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 $\alpha$.

Exercise 1.4.5. 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.

### 1.5. 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 $\gamma$ 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 1.5.1. 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 $(a) \Longrightarrow(b) \Longrightarrow(c) \Longrightarrow(d) \Longrightarrow(a)$. You might find the $S S S$, 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 $2 \pi$, 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 1.5.2. 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 $\pi$, 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 $\pi$.) The point of requiring that the interior angle is less than $\pi$ is that then, if you start moving into the triangle $p q r$ from $p$, you move into $P$ rather than out of it.


Figure 2. The two cases in the inductive step.
If the segment $q r$ is a diagonal, we are done. So, assume $q r$ 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 $p q r$ that lies farthest from the segment $q r$.

We claim that $p z$ is a diagonal. Since the segment $p z$ 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 $q r$ 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 $p q r$ and be farther from $q r$ than $z$, contrary to assumption. So, $p z$ is a diagonal.

ThEOREM 1.5.3. 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 $\operatorname{big} O$ notation, and is especially common in computer science. For example, you can check that

$$
\left(1+n^{5}\right)\left(\sin (n)+n^{2}\right)=O\left(n^{7}\right)
$$

ExERCISE 1.5.4. Analyzing the proof above, explain why the number of steps required to triangulate an $n$-gon is $O\left(n^{4}\right)$. 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 $100 n$ vs $2 n$, it's still $O(n)$.

There are faster ways to triangulate polygons, though. Here's one approach.
Definition 1.5.5. 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 $r p q$ and $p q z$.

Theorem 1.5.6. 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 $x y$, 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 $x y$. These are two nonoverlapping ears of $P$.

EXERCISE 1.5.7. By looking for ears instead of arbitrary diagonals, give an algorithm to triangulate an $n$-gon in $O\left(n^{3}\right)$ 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 $O\left(n^{2}\right)$-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 1.5.8. 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 $360^{\circ}$ field of vision.

THEOREM 1.5.9. $\lfloor n / 3\rfloor$ 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 1.5.10. 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 $x y$ be a diagonal of $P$. Then $x y$ 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 1.5.9. 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 $n=3 k$ 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 $n=3 k+1$ or $n=3 k+2$, just insert either one or two additional vertices into the bottom edge of the polygon drawn above.

### 1.5.1. Exercises.

ExERCISE 1.5.11. 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 $c$.


Exercise 1.5.12. 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 1.5.13. 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 1.5.14. The interior angles of a triangle sum to $\pi$. 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 1.5.15. 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 $x y \subset P$. Try to draw some examples of convex, and non-convex polygons.

Exercise 1.5.16. Let's say a vertex of a polygon $P$ is convex if its interior angle is at most $\pi$, and concave otherwise. Show that $P$ is convex if and only if all its vertices are convex.

EXERCISE 1.5.17. How many triangulations can a polygon have? We saw in Exercise 1.5.8 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 1.5.18 (A continuation of Exercise 1.5.17). The $n^{\text {th }}$ Catalan number is

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

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 1.5.17.

### 1.6. 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 1.6.1. 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 1.6.2. 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 1.6.3. 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 1.6.4 (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 1.6.5. 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 $\alpha$ 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 $[a / 2,2 a]$. So, repeated P-slides can be used to show that any two rectangles with the same area are scissors congruent.

Proof of Theorem 1.6.4. 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 1.5.3. 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 $\sqrt{\operatorname{Area}(A)}$. 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 $\pi$ '. The point is that now we are limiting the movement of the polygons to certain isometries.

Exercise 1.6.6. (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 1.6.7. Show that any two polygons with the same area are scissors congruent via translations and rotations by $\pi$. Hint: repeat the proof of Theorem 1.6.4 using Exercise 1.6.6.

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.


$$
v \text {-Hadwiger }=\operatorname{length}\left(e_{1}\right)-\operatorname{length}\left(e_{2}\right) .
$$



ExERCISE 1.6.8. 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.

ExERCISE 1.6.9. Suppose $P, Q$ are two polygons in $\mathbb{R}^{2}$ that have the same area. Can you write $P, Q$ as unions of polygons $P=P_{1} \cup P_{2}$ and $Q=Q_{1} \cup Q_{2}$ such that $P_{i}$ is congruent to $Q_{i}$ for $i=1,2$ ?

ExERCISE 1.6.10. Suppose that $P, Q$ are two polygons in $\mathbb{R}^{2}$ that have the same area and the same perimeter. Show that there is a scissors congruence from $P$ to $Q$ that takes points on the boundary of $P$ to points on the boundary of $Q$.

### 1.7. 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.

[^0]

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?

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

[^1]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 1.7.2. 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

$$
v \times w=\operatorname{det}\left(\begin{array}{ccc}
i & j & k \\
v_{1} & v_{2} & v_{3} \\
w_{1} & w_{2} & w_{3}
\end{array}\right)=\left(v_{2} w_{3}-w_{2} v_{3}\right) i-\left(v_{1} w_{3}-w_{1} v_{3}\right) j+\left(v_{1} w_{2}-w_{1} v_{2}\right) k
$$

Here, we are using the multivariable calculus notation $(a, b, c)=a i+b j+c k$, so in our usual notation the cross product is $v \times w=\left(v_{2} w_{3}-w_{2} v_{3}, w_{1} v_{3}-v_{1} w_{3}, v_{1} w_{2}-w_{1} v_{2}\right)$.

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

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

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

Proposition 1.7.3. There is an additive function $f: \mathbb{R} \longrightarrow \mathbb{R}$ such that

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

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 1.7.3 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 1.7.4. Show that if $f$ is additive, $f(q x)=q f(x)$ for all $x \in \mathbb{R}$ and $q \in \mathbb{Q}$.

Assuming Proposition 1.7.3, 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 1.7.3. The Dehn invariant $D_{f}$ associated to $f$ is defined as follows. If $P$ is a union of polyhedra, let

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

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

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

Theorem 1.7.5. 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 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 $e$ :

$$
\begin{aligned}
\operatorname{length}(e) f(\angle(e)) & =\left(\operatorname{length}\left(e_{1}\right)+\operatorname{length}\left(e_{2}\right)\right) f(\angle(e)) \\
& =\operatorname{length}\left(e_{1}\right) f(\angle(e))+\operatorname{length}\left(e_{2}\right) f(\angle(e)) \\
& =\operatorname{length}\left(e_{1}\right) f\left(\angle\left(e_{1}\right)\right)+\operatorname{length}\left(e_{2}\right) f\left(\angle\left(e_{2}\right)\right) .
\end{aligned}
$$

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 $\angle\left(e_{1}\right)+\angle\left(e_{2}\right)=\angle(e)$. So,

$$
\begin{aligned}
\operatorname{length}\left(e_{1}\right) f\left(\angle\left(e_{1}\right)\right)+\text { length }\left(e_{2}\right) f\left(\angle\left(e_{2}\right)\right) & =\text { length }(e)\left(f\left(\angle\left(e_{1}\right)\right)+f\left(\angle\left(e_{2}\right)\right)\right) \\
& =\operatorname{length}(e) f(\angle(e)) .
\end{aligned}
$$

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 $\pi$. So,

$$
\text { length }\left(d_{1}\right) f\left(\angle\left(d_{1}\right)\right)+\operatorname{length}\left(d_{2}\right) f\left(\angle\left(d_{2}\right)\right)=0
$$

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 1.7.1. Now suppose that both $f(\pi)=0$ and $f\left(\cos ^{-1}\left(\frac{1}{3}\right)\right)=$ 1, as in Proposition 1.7.3, 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, $\ell_{C}$ and $\ell_{T}$ are the lengths of the edges in $C$ and $T$, respectively. Since $D_{f}(C) \neq D_{f}(T)$, Theorem 1.7 .5 says that $C$ and $T$ are not scissors congruent.
1.7.1. Linear algebra over $\mathbb{Q}$ : a proof of Proposition 1.7.3. We now prove that there is an additive function $f$ with $f\left(\cos ^{-1}(1 / 3)\right)=1$ and $f(\pi)=0$, as required by Proposition 1.7.3. 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 $\pi$ 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 1.7.6 (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 $p \mid a_{0}$ and $q \mid a_{n}$.

Proof. As $a_{n}\left(\frac{p}{q}\right)^{n}+\cdots+a_{0}=0$, 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 $q \mid a_{n}$ 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 1.7.7. If $x \in \mathbb{Q}$ is a root of a monic, integer polynomial, then $x \in \mathbb{Z}$.

Proof. Suppose $\left(\frac{p}{q}\right)^{n}+a_{n-1}\left(\frac{p}{q}\right)^{n-1}+\cdots+a_{0}=0$, where $p / q$ is in lowest terms. By the Rational Root Theorem, $q \mid 1$, 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 1.7.8. Within the interval $[-\pi, \pi]$, the only rational multiples of $\pi$ 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 1.7.8 implies that $\cos ^{-1}(q)$ is not a rational multiple of $\pi$. In particular,

Corollary 1.7.9. $\cos ^{-1}\left(\frac{1}{3}\right)$ is not a rational multiple of $\pi$.
The proof of the proposition is an application of Corollary 1.7.7.
Proof of Proposition 1.7.8. By the angle sum formula, for $\alpha \in \mathbb{R}$ and $n \in \mathbb{N}$, we have

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

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

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

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 $P_{n}(x)=2 \cos \left(n \frac{m}{n} \pi\right)=2(-1)^{m}$. 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 1.7.7 implies that $x$ is an integer, implying $\cos (\alpha)$ is one of $0, \pm \frac{1}{2}, \pm 1$. This implies $\alpha$ is one of $0, \pm \frac{\pi}{3}, \pm \frac{\pi}{2}, \pm \pi$.

So, now we know that $\cos ^{-1}(1 / 3)$ and $\pi$ are not rational multiples of each other, and we want to conclude that there is an additive $f$ with $f\left(\cos ^{-1}(1 / 3)\right)=1$ 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 1.7.10. 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 $q x=0$ 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 1.7.11. 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 $b=q a$, where $a \in \mathbb{Q}$. Then $b-q a=0$, 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 $q a+r b=0$. If $r=0$, then $q a=0$, 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 $\left\{\cos ^{-1}(1 / 3), \pi\right\}$ are both $\mathbb{Q}$-independent sets.
EXERCISE 1.7.12. 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 1.7.13. 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 $f: \mathbb{R} \rightarrow \mathbb{R}$ 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 $q a+r b=0$ for some $q, r \in \mathbb{Q}$, then using Exercise 1.7.4, if $f$ is additive

$$
0=f(q a+r b)=q f(a)+r f(b)
$$

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

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

so Proposition 1.7 .3 is a direct corollary.
We will now proceed toward a proof of Theorem 1.7.13. We define:
Definition 1.7.14. If $S \subset \mathbb{R}$, the $\mathbb{Q}$-span of $S$ is

$$
\operatorname{span}_{\mathbb{Q}}(S)=\left\{q_{1} s_{1}+\cdots q_{n} s_{n} \mid q_{i} \in \mathbb{Q}, s_{i} \in S\right\}
$$

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

Lemma 1.7.15. Suppose that $S \subset \mathbb{R}$ is $\mathbb{Q}$-independent, and $x \in \mathbb{R} \backslash \operatorname{span}_{\mathbb{Q}}(\mathbb{R})$. Then $S \cup\{x\}$ is $\mathbb{Q}$-independent.

Proof. Suppose that we have $q_{1} s_{1}+\cdots+q_{n} s_{n}+r x=0$, 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

$$
x=\left(-q_{1} / r\right) s_{1}+\cdots+\left(-q_{n} / r\right) s_{n} \in \operatorname{span}_{\mathbb{Q}}(\mathbb{R}),
$$

a contradiction.
Definition 1.7.16. A $\mathbb{Q}$-basis for $\mathbb{R}$, or alternatively a Hamel basis, is a $\mathbb{Q}$ independent set $S \subset \mathbb{R}$ such that $\operatorname{span}_{\mathbb{Q}}(S)=\mathbb{R}$.

Lemma 1.7.17. $S \subset \mathbb{R}$ 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 $x \in \operatorname{span}_{\mathbb{Q}}(S)$, 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 $\left(q_{1}-r_{1}\right) s_{1}+\cdots+\left(q_{n}-r_{n}\right) s_{n}=0$, 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, $\operatorname{span}_{\mathbb{Q}}(S)=\mathbb{R}$, 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 1.7.18. 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 $\operatorname{span}_{\mathbb{Q}}(S)=\mathbb{R}$, we're done. If not, take some $x_{1} \in \mathbb{R} \backslash \operatorname{span}_{\mathbb{Q}}(S)$. By Lemma 1.7.15, $S \cup\left\{x_{1}\right\}$ 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 $S_{i}:=S \cup\left\{x_{1}, \ldots, x_{i}\right\}$, and preserving independence in every step. One would like to say that this terminates with some $S$ where $\operatorname{span}_{\mathbb{Q}}(S)=$ $\mathbb{R}$. 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 $S_{\infty}:=S \cup\left\{x_{1}, x_{2}, \ldots\right\}$ is still not $\mathbb{R}$ ! However, in this case one just continues as before: pick some $y_{1}$ outside the span of $S_{\infty}$ and add it to $S_{\infty}$, 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 1.7.19 (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 $\left\langle x, s_{i}\right\rangle:=q_{i}$, so that

$$
x=\sum_{s_{i} \in S}\left\langle x, s_{i}\right\rangle s_{i} .
$$

That is, $\left\langle x, s_{i}\right\rangle$ 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

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

ExERCISE 1.7.20. 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 1.7.21. 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

$$
x+y=\sum_{s_{i} \in S}\left\langle x, s_{i}\right\rangle s_{i}+\sum_{s_{i} \in S}\left\langle y, s_{i}\right\rangle s_{i}=\sum_{s_{i} \in S}\left(\left\langle x, s_{i}\right\rangle+\left\langle y, s_{i}\right\rangle\right) s_{i},
$$

so by definition, $\left\langle x+y, s_{i}\right\rangle=\left\langle x, s_{i}\right\rangle+\left\langle y, s_{i}\right\rangle$.
We can now prove Theorem 1.7.13, 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 1.7.13. 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 1.7.18. 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 1.7.20, $f(s)=a_{s}$ for all $s$. Additivity of $f$ follows from Fact 1.7.21: we leave the details as an exercise.

ExERCISE 1.7.22 (For students with an analysis background). Show that if $f$ : $\mathbb{R} \longrightarrow \mathbb{R}$ is additive and continuous, then $f(x)=a x$ for some $a \in \mathbb{R}$.

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

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

Examples include functions $f(x)=x^{a}$, where $a \in \mathbb{R}$.
Exercise 1.7.23. 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 1.7.24. 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 1.7.25 (For students with an analysis background). Suppose that $x$ is irrational. Show that the set of all numbers of the form $m+n x$, 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 1.7.26. 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 $a x \notin S$. Hint: if $a x \in S$ for all $x \in S$, the sum $\sum_{i} q_{i}$ of the coefficients in (3) will be the same for $x$ as it is for ax. (Why?)

In Exercise 1.7.26, if $a$ is rational, it can never be the case that $a x \in S$. For if so, we'd have $a \cdot x-1 \cdot(a x)=0$, 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 1.7.27. 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 $S$ :

$$
\begin{equation*}
x=q_{1} s_{1}+\cdots+q_{k} s_{k}, \quad q_{1}, \ldots, q_{k} \in \mathbb{Z}, s_{1}, \ldots, s_{k} \in S \tag{3}
\end{equation*}
$$

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

Exercise 1.7.28. 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

$$
a+a^{\prime} \in A, \text { for all } a, a^{\prime} \in A, \text { and } b+b^{\prime} \in B \text { for all } b, b^{\prime} \in A
$$

Exercise 1.7.29. A function $f: \mathbb{R} \longrightarrow \mathbb{R}$ is periodic with period $\ell$ if $f(x+\ell)=f(x)$ for all $x \in \mathbb{R}$. For example, sin and $\cos$ are periodic with period $2 \pi$.
(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 $f+g=i d$, 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, $x=\sum_{s \in S}\langle x, s\rangle$. 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 1.7.30. 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

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

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 $p, q \in S_{r}$,

$$
d_{S_{r}}(p, q)=\inf \left\{\operatorname{length}(\gamma) \mid \gamma:[a, b] \longrightarrow S_{r}, \gamma(a)=p, \gamma(b)=q\right\}
$$

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 $p, q \in A$,

$$
d_{A}(p, q)=\inf \{\operatorname{length}(\gamma) \mid \gamma:[a, b] \longrightarrow A, \gamma(a)=p, \gamma(b)=q\}
$$

Show that the function $d_{A}$ satisfies all the usual properties of a metric, except that $d_{A}(p, q)$ may equal $\infty$ 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

$$
\text { length }\left.\gamma\right|_{[0, \theta]}=\int_{0}^{\theta}\left|\gamma^{\prime}(t)\right| d t=\int_{0}^{\theta} r d t=r \theta
$$

we see that when the angle between $p, q \in S_{r}$ is $\theta$, 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 $\theta$, any path $\gamma$ from $p$ to $q$ has length at least $r \theta$, with equality only if $\gamma$ 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 $\gamma$ is piecewise differentiable, so that we can calculate its length via an integral formula. In general, a continuous path $\gamma$ 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 $\gamma$ radially onto the sphere. So, any continuous path with length less than $\theta$ would yield a piecewise differentiable path with length less than $\theta$. 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 $\gamma$ is contained in an open hemisphere around $p$, and in particular $\theta<\pi / 2$. If not, cut $\gamma$ into small pieces, and observe that if the length of $\gamma$ 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, $\gamma$ can be written as

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

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 $\gamma$ at time $t$ is then:

$$
\begin{aligned}
\left|\gamma^{\prime}(t)\right|^{2}= & \left|\left(r u^{\prime} \cos u \cos v-r v^{\prime} \sin u \sin v, r u^{\prime} \cos u \sin v+r v^{\prime} \sin u \cos v,-r u^{\prime} \sin u\right)\right|^{2} \\
= & \left(r u^{\prime} \cos u \cos v-r v^{\prime} \sin u \sin v\right)^{2}+\left(r u^{\prime} \cos u \sin v+r v^{\prime} \sin u \cos v\right)^{2}+ \\
& \left(-r u^{\prime} \sin u\right)^{2} \\
= & \left(r u^{\prime} \cos u \cos v\right)^{2}+\left(r v^{\prime} \sin u \sin v\right)^{2}+\left(r u^{\prime} \cos u \sin v\right)^{2}+ \\
& \left(r v^{\prime} \sin u \cos v\right)^{2}+\left(-r u^{\prime} \sin u\right)^{2}, \quad \text { as the middle terms above cancel, } \\
= & \left(r u^{\prime} \cos u\right)^{2}+\left(r v^{\prime} \sin u\right)^{2}+\left(-r u^{\prime} \sin u\right)^{2}, \quad \text { using } \sin ^{2}+\cos ^{2}=1, \\
= & \left(r u^{\prime}\right)^{2}+\left(r v^{\prime} \sin u\right)^{2},
\end{aligned}
$$

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

$$
\begin{aligned}
\operatorname{length}(\gamma) & =\int_{a}^{b} \sqrt{\left(r u^{\prime}(t)\right)^{2}+\left(r v^{\prime}(t) \sin u(t)\right)^{2}} d t \\
& \geqslant \int_{a}^{b}\left|r u^{\prime}(t)\right| d t \\
& \geqslant \int_{a}^{b} r u^{\prime}(t) d t \\
& =r \frac{\pi}{2}-r\left(\frac{\pi}{2}-\theta\right) \\
& =r \theta
\end{aligned}
$$

Moreover, we have equality above if and only if $v^{\prime}(t)=0$ and $u^{\prime}(t)>0$ for all $t$. If this happens, $v(t)=\pi / 2$ for all $t$, so $\gamma$ lies along the great circle determined by the $x z$-plane, and the condition $u^{\prime}(t)>0$ 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 $\gamma$ from $p$ to $q$. Then $\gamma$ is a segment of a great circle.

Proof. Let $P$ be a plane in $\mathbb{R}^{3}$ passing through the origin and $p, q^{\prime}$. The reflection through $P$ sends $\gamma$ to a path on $S_{r}$ with the same endpoints that has the same length as $\gamma$, which must then be $\gamma$, by our assumption. This means that the reflection fixes $\gamma$, so $\gamma^{\prime}$ 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 $\theta$ 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

$$
C(p, s)=\left\{q \in S_{r} \mid d_{S_{r}}(p, q)=s\right\} .
$$

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 $\theta$ is the angle between $p, q$. As $d_{S_{r}}(p, q)=r \theta$,

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

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

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

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 $\left(r \cos \frac{s}{r}\right) p \in P$.


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 $a x+b y+c z=d$. 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

$$
\lim _{r \rightarrow \infty} 2 \pi r \sin \frac{s}{r}=2 \pi s
$$

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

$$
\lim _{s \rightarrow 0} \frac{2 \pi r \sin \frac{s}{r}}{2 \pi s}=1
$$

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 $\pi$, 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 $\theta$ is the angle between the segments $0 p$ and $0 q$. By Lemma 1.2.12, as $f$ is an isometry the angle between $0 f(p)$ and $0 f(q)$ is $\theta$ 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

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

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 $\ell$, 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 $\ell$. These properties determine $R_{\ell}$, and are completely analogous to those in the geometric definition of a reflection in $\mathbb{R}^{2}$. So, by analogy, we call $R_{\ell}$ the reflection of $S_{r}$ through the great circle $\ell$.


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 $\theta$ 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 $p x$ and $p O_{p, \theta}(x)$ have the same length and meet with an angle of $\theta$, measured counterclockwise from $p x$ to $p O_{p, \theta}(x)$.


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 $\mathrm{t}(\mathrm{W})$ 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 $\ell$ and reflections through $\ell$, 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 $\ell$ by angle $\theta$.

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 $\pi$.

Exercise 2.3.2. Show that a proper triangle on $S_{r}$ has side lengths less than $\pi r$. 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 $\theta$ 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 $2 n$ lunes with angle $\pi / n$ and disjoint interiors, a lune with angle $\theta=\frac{\pi}{n}$ must have area $4 \pi r^{2} / 2 n=2(\pi / n) r^{2}$. 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 $\theta$ is rational. In general, given any angle $\theta$, we can write $\theta / \pi$ 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 $\pi$, so a lune with angle $\theta$ is the union of lunes with angles that are rational multiples of $\pi$ 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 $\alpha$ 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 $\alpha, \beta, \gamma$, then

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

In particular, this implies that the angle sum of a spherical triangle is always greater than $\pi$, 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 $\alpha, \beta, \gamma$. 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 $\alpha$, $\beta$, and $\gamma$. 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 $T^{\prime}$ are contained in 3 lunes. So,

$$
\begin{aligned}
4 \pi r^{2} & =\operatorname{Area}(S) \\
& =2 \cdot\left(2 \alpha r^{2}\right)+2 \cdot\left(2 \beta r^{2}\right)+2 \cdot\left(2 \gamma r^{2}\right)-2 \operatorname{Area}(T)-2 \operatorname{Area}\left(T^{\prime}\right) \\
& =4 r^{2}(\alpha+\beta+\gamma)-4 \operatorname{Area}(T)
\end{aligned}
$$

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 $T^{\prime}$ to compensate for the overcounting. Solving for $\operatorname{Area}(T)$ 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 $\pi$. 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 $\theta$ 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. 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 $\S 1.5$ 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

$$
H=\{(x, y, z) \in S \mid z<0\}, \quad Z=\left\{(x, y, z) \in \mathbb{R}^{3} \mid z=-1\right\} .
$$

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 $\gamma(t)=(t x, t y, t z)$. This ray intersects the plane at height -1 when $t z=-1$, 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 2.4.1. 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 2.4.2. 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 2.4.3. 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.

### 2.5. 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 2.5.1. If $P$ is a proper $n$-gon on $S_{r}$ with interior angle sum $s$, then

$$
\operatorname{Area}(P)=r^{2}(s-(n-2) \pi)
$$

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

$$
\operatorname{Area}(P)=\sum_{i=1}^{n-2} \operatorname{Area}\left(T_{i}\right)=\sum_{i=1}^{n-2} r^{2}\left(s_{i}-\pi\right)=r^{2}(s-(n-2) \pi)
$$

ExERCISE 2.5.2. Show directly that the conclusion of Theorem 2.5.1 holds for lunes and monogons.

ExERCISE 2.5.3. Show that the conclusion of Theorem 2.5.1 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 2.5.4. 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 2.5.5. 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 2.5.6. The Euler characteristic of a proper tiling of $S_{r}$ is 2 .
Proof. The interior angles around a vertex of the tiling sum to $2 \pi$, 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

$$
\begin{aligned}
4 \pi r^{2} & =\operatorname{Area}\left(S_{r}\right) \\
& =\sum_{\text {polygons } P} \operatorname{Area}(P) \\
& =\sum_{\text {polygons } P} r^{2}(s(P)-\pi(n(P)-2)), \quad \text { by Corollary } 2.5 .1 \\
& =r^{2}\left(\sum_{\text {polygons } P} s(P)-\pi \sum_{\text {polygons } P} n(P)+\pi \sum_{\text {polygons } P} 2\right) \\
& =r^{2}(2 \pi V-2 \pi E+2 \pi F) \\
& =2 \pi r^{2} \chi
\end{aligned}
$$

This implies that $\chi=2$.
EXERCISE 2.5.7. 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 2.5.3 instead of Corollary 2.5.1.

By Exercise 2.5.4, 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 2.5.8. The Euler characteristic of a convex polyhedron is 2.

[^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 $2-2 g$ ! This is a little tricky to prove (and state precisely) in general, but you can verify it in special cases.

Exercise 2.5.9. For each $g$, describe the construction of a specific polyhedron with $g$ holes in which you can easily show that the Euler characteristic is $2-2 g$.

### 2.6. 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 2.6.1. If two triangles $T, T^{\prime}$ 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 2.6.2. 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 2.6.3 (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 2.6.4. 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 2.6.2 and 2.6.3. 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 $(2 s+1)$-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 $n=2 s+1 \geqslant 5$ is odd.

ExErcise 2.6.5. 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 2.6.6. 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 2.6.6. 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 2.6.7. 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 2.5.7 says that $\chi(P)=1$.

EXERCISE 2.6.8. 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 2.6.6. 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 $T_{r}^{\prime}$ be the union of all polygons in the tiling that intersect the $\operatorname{disc} D_{r}=\left\{x \in \mathbb{R}^{2}| | x \mid \leqslant r\right\}$, as pictured below.


We'd like to say that $T_{r}^{\prime}$ is a polygon, and indeed, it looks like the polygon in the picture above. But what some of the polygons in $T_{r}^{\prime}$ enclose other polygons of the tiling that don't touch $D_{r}$, as in the following picture? Then $T_{r}^{\prime}$ 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 $T_{r}^{\prime}$ or are enclosed by a loop in $T_{r}^{\prime}$. Then any 'holes' in $T_{r}^{\prime}$ 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, $\chi\left(T_{r}\right)=1$.

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 $V_{i n t}$ and $V_{\text {ext }}$, so that we have $V=V_{\text {int }}+V_{\text {ext }}$. The total angle sum of all the polygons in $T_{r}$ is at most $2 \pi V$, since each interior vertex contributes $2 \pi$ while the other vertices contribute less than $2 \pi$. 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,

$$
(n-2) \pi F \leqslant 2 \pi V \Longrightarrow \frac{n-2}{2} F \leqslant V
$$

As the polygons in $T_{r}$ are convex, their angles are less than $\pi$. So, each interior vertex in $T_{r}$ is adjacent to at least three edges, implying

$$
3 V_{i n t} \leqslant 2 E
$$

We now use these inequalities in the definition of the Euler characteristic of $T_{r}$, which we said above is 1 , giving

$$
1=\chi\left(T_{r}\right)=V-E+F \leqslant V-\frac{3}{2} V_{i n t}+\frac{2}{5} V=\left(\frac{n}{n-2}-\frac{3}{2} \frac{V_{i n t}}{V}\right) V
$$

The key now is in the following claim:
CLAIM 2.6.9. As $r \longrightarrow \infty$, we have $\frac{V_{i n t}}{V} \longrightarrow 1$.
Assuming the claim for a minute, the inequality above says that when $r$ is large, we have $0 \leqslant\left(\frac{n}{n-2}-\frac{3}{2}\right)$, 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 2.6.9. 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 $F_{i n t}$, and exterior polygons by $F_{\text {ext }}$. The intuition here is that $V_{\text {int }}$ and $F_{\text {int }}$ should both be proportional to the area $\pi r^{2}$ of $D_{r}$, while $V_{e x}$ and $F_{e x t}$ should be proportional to the circumference $2 \pi r$ of $D_{r}$, and for large $r$, we have $\pi r^{2} \gg 2 \pi r$.

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,

$$
\operatorname{Area}(P) F_{\text {int }} \geqslant \pi(R-D)^{2}
$$

Similarly, every exterior polygon lies outside the disk of radius $R-D$, and within a disk of radius $R+D$, so we have

$$
\operatorname{Area}(P) F_{e x t} \leqslant \pi\left((R+D)^{2}-(R-D)^{2}\right)=4 \pi D R
$$

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 $2 \pi$, this number is at most $2 \pi$ divided by the minimum angle of $P$. Denoting this last quantity by $\alpha(P)$, we have

$$
n \cdot F_{i n t} \leqslant \frac{2 \pi}{\alpha(P)} V_{i n t}
$$

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

$$
n \cdot F_{e x t} \geqslant V_{e x t}
$$

Putting this info together, we have

$$
\begin{aligned}
& V_{\text {int }} \geqslant \frac{n \alpha(P)}{2 \pi} \cdot \frac{\pi(R-D)^{2}}{\operatorname{Area}(P)}=\frac{n \alpha(P)}{2 \operatorname{Area}(P)}(R-D)^{2} \\
& V_{\text {ext }} \leqslant n \cdot \frac{4 \pi D R}{\operatorname{Area}(P)}=\frac{4 \pi n D}{\operatorname{Area}(P)} R
\end{aligned}
$$

This means we have

$$
\frac{V_{e x t}}{V_{\text {int }}} \leqslant \frac{\frac{4 \pi n D}{\operatorname{Area}(P)} R}{\frac{n \alpha(P)}{2 \operatorname{Area}(P)}(R-D)^{2}}=\frac{8 \pi D}{\alpha(P)} \frac{R}{(R-D)^{2}}
$$

which goes to zero as $R \rightarrow \infty$. Thus we have

$$
\frac{V_{\text {int }}}{V_{i n t}+V_{e x t}}=\frac{1}{1+\frac{V_{e x t}}{V_{i n t}}} \rightarrow 1
$$

as $R \rightarrow \infty$, as desired.

### 2.7. 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 2.7.1 (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 $\operatorname{Area}(P)=I+B / 2-1$.


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 2.7.2. Suppose $\Delta \subset \mathbb{R}^{2}$ is a minimal integer triangle. Then $\operatorname{Area}(\Delta)=\frac{1}{2}$.
Proof of Lemma 2.7.2. For notational convenience, let's translate $\Delta$ 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 $p q$. Then the rotation by $\pi$ around $p$ has the formula

$$
O_{m, \pi}: \mathbb{R}^{2} \longrightarrow \mathbb{R}^{2}, \quad O_{m, \pi}(x)=2 m-x=p+q-x
$$

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 $\Delta$ 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 $\operatorname{Area}(P)=1$. Tile the plane with parallelograms congruent to $P$. The parallelograms in our tiling all have the form $T_{i p+j q}(P)$, where $i, j \in \mathbb{Z}$. Since $i p+j q$ is an integer point, each $T_{i p+j q}(P)$ 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 $T_{a p+b q}(P)$ 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 $a p+b q$ of the vertex 0 of $P$. So, $a p+b q=(1,0)$. Similarly, there are $c, d \in \mathbb{Z}$ such that $c p+d q=(0,1)$. Then

$$
\left(\begin{array}{ll}
p & q
\end{array}\right)\left(\begin{array}{ll}
a & c \\
b & d
\end{array}\right)=\left(\begin{array}{ll}
1 & 0 \\
0 & 1
\end{array}\right)
$$

where $(p q)$ is the matrix with columns $p$ and $q$. So,

$$
1=\operatorname{det}\left(\begin{array}{ll}
1 & 0 \\
0 & 1
\end{array}\right)=\operatorname{det}\left(\begin{array}{ll}
p & q
\end{array}\right) \operatorname{det}\left(\begin{array}{ll}
a & c \\
b & d
\end{array}\right)
$$

and both determinants on the right hand side are integers, so they are both $\pm 1$. In particular, Exercise 1.1.10 says that $\operatorname{Area}(P)=\left|\operatorname{det}\left(\begin{array}{ll}p & q\end{array}\right)\right|=1$.

To deduce Pick's Theorem from the Lemma 2.7.2, we need to show:
Lemma 2.7.3. 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 $\Delta$. Here, the argument is by strong induction on badness of $\Delta$, which we define to be the number of integer points that lie in $\Delta$ but are not vertices of $\Delta$. If the badness of $\Delta$ is zero, then $\Delta$ 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 $\Delta$ with badness $n$. Pick some integer point $q$ in $\Delta$ that is not a vertex. If $q$ lies on an edge of $\Delta$, split $\Delta$ along the segment connecting $q$ to the opposite vertex of $\Delta$, into two integer triangles with smaller badness than $\Delta$. By induction, these triangles can be tiled by minimal integer triangles. If $q$ lies in the interior of $\Delta$, split $\Delta$ along the line segments connected $q$ to the vertices of $\Delta$, giving three integer triangles with smaller badness than $\Delta$. 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 2.7.3, 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 2.6.7. 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

$$
3 F=2(E-B)+B, \quad \Longrightarrow E=\frac{1}{2}(3 F+B)
$$

To see this, note that each side counts the number of pairs $(\Delta, e)$, where $\Delta$ is a triangle in our tiling, and $e$ is one of its edges. On the left side, we pick $\Delta$ 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

$$
1=V-E+F=I+B-\frac{1}{2}(3 F+B)+F=I+\frac{B}{2}-\frac{1}{2} F .
$$

But by Lemma 2.7.2, we have $A=F / 2$, implying that $F=2 A$, and hence

$$
1=I+B / 2-A, \quad \Longrightarrow A=I+B / 2-1 .
$$

### 2.7.1. Exercises.

ExERCISE 2.7.4 (Reeve tetrahedra, see [?]). 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 2.7.5. Figure out how to replace the last paragraph of the proof of Lemma 2.7.2 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 $I_{r}=\left|\mathbb{Z}^{2} \cap D_{r}\right|$. Arguing as in proof of Claim 2.6.9, show that we have:
(a) $\lim _{r \rightarrow \infty} \operatorname{Area}\left(D_{r}\right) / I_{r}=1$,
(b) $\lim _{r \rightarrow \infty} \operatorname{Area}\left(D_{r}\right) / N_{r}=\operatorname{Area}(P)$,
(c) $\lim _{r \rightarrow \infty} I_{r} / N_{r}=1$.

Then conclude that Area $(P)=1$.
ExERCISE 2.7.6. Suppose that $P$ is a (possibly non-integer) polygon in $\mathbb{R}^{2}$ with vertices $\left(x_{i}, y_{i}\right)$, where $i=1, \ldots, n$, and define $\left(x_{n+1}, y_{n+1}\right):=\left(x_{0}, y_{0}\right)$.
(a) Show that $\operatorname{Area}(A)=\sum_{i=1}^{n}\left(y_{i}+y_{i+1}\right)\left(x_{i}-x_{i+1}\right)$, Hint: compute the area of the trapezoid with vertices $\left(x_{i}, 0\right),\left(x_{i+1}, 0\right),\left(x_{i+1}, y_{i+1}\right),\left(x_{i}, y_{i}\right)$.
(b) Using (a), show that $\operatorname{Area}(A)=\sum_{i=1}^{n} \operatorname{det}\left(\begin{array}{ll}x_{i} & x_{i+1} \\ y_{i} & y_{i+1}\end{array}\right)$.

EXERCISE 2.7.7. Show that there is no equilateral integer triangle in $\mathbb{R}^{2}$.
Exercise 2.7.8 (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 $[a, c] \times[c, d]:=\left\{(x, y) \in \mathbb{R}^{2} \mid x \in[a, b], y \in[c, d]\right\}$, which is a rectangle in $\mathbb{R}^{2}$. For each pair of integers $i, j$, consider

$$
X_{i j}=X \cap([i, i+1] \times[j, j+1]),
$$

so in other words, $X_{i j}$ 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 $T_{-(i, j)}\left(X_{i, j}\right)$ for at least $\lceil A\rceil$ different choices of $(i, j)$.

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 $x y \subset C$.

THEOREM 2.7.9 (Minkowski). Any centrally symmetric, convex subset $C \subset \mathbb{R}^{2}$ with area bigger than 4 contains a nonzero integer point.

To get some intuition for the theorem, note that all three conditions on $C$ are necessary if you want to ensure that there's a nonzero integer point in $C$. Below, the first picture shows a subset $C$ with area 4 that is centrally symmetric and convex, but only contains one integer point, as long as you do not include the boundary of the square in the subset $C$. The second picture is convex, has area bigger than 4 , but isn't centrally symmetric. The third picture is centrally symmetry and has area bigger than 4, but isn't convex.


Exercise 2.7.10. Prove Minkowski's Theorem. 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 $v \in C$. Use Blichfeldt's Theorem.

ExERCISE 2.7.11. In class, we proved that every minimal integer triangle in $\mathbb{R}^{2}$ has area $1 / 2$. Here's another way to prove that lemma, using Minkowski's Theorem.
(a) If $\Delta$ is an integer triangle, show that $\operatorname{Area}(\Delta) \geqslant \frac{1}{2}$, using that determinant exercise from the first HW assignment.
(b) Suppose $\Delta$ is a minimal integer triangle. Construct an integer parallelogram $Q$ tiled by 8 copies of $\Delta$ such that $Q$ contains a single integer point in its interior. Using Minkowski's Theorem, show that 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 [?] for an English translation, and Hartshorne [?] 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 $\alpha, \beta$ below sum to less than $\pi$ then the lines $m$ and $n$ intersect on the indicated side of $\ell$.


[^3]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.
$(\mathrm{P})$ "If a point $p$ does not lie on a line $\ell$, there is a unique line passing through $p$ that is parallel to $\ell$."
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 $p q$ 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 $\S 2.1$. 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 $\pi r$, 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 $\pi$. 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 (12011274), 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

$$
\mathbb{H}^{2}=\left\{(x, y) \in \mathbb{R}^{2} \mid y>0\right\} \subset \mathbb{R}^{2}
$$

A hyperbolic line is the intersection $\ell \cap \mathbb{H}^{2}$, where $\ell$ 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 $\ell$ is a hyperbolic line and $x \in \mathbb{H}^{2}$ does not lie on $\ell$, there are infinitely many hyperbolic lines through $x$ that are parallel to $\ell$. For instance, in the following picture many hyperbolic lines are drawn through $p$ that do not intersect the given hyperbolic line $\ell$.


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 $\S 3.1$, 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 $x y$ is the center of a circle passing through both $x$ and $y$. As $z \rightarrow \infty$ 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

$$
\begin{equation*}
a x^{2}+a y^{2}+b x+c y+d=0 \tag{4}
\end{equation*}
$$

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

$$
\begin{equation*}
a x^{2}+a y^{2}+b x+c y+d=0 \tag{5}
\end{equation*}
$$

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

$$
b x+c y+d=0
$$

while a circle with center $(a, b)$ and radius $r$ is the solution set of

$$
(x-a)^{2}+(y-b)^{2}=r^{2} \Longleftrightarrow x^{2}+y^{2}-2 x a-2 b y+\left(a^{2}+b^{2}-r^{2}\right)=0
$$

which has the desired form.
Conversely, consider an equation $a x^{2}+a y^{2}+b x+c y+d=0$ 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

$$
x^{2}+y^{2}+\frac{b}{a} x+\frac{c}{a} y+\frac{d}{a}=\left(x+\frac{b}{2 a}\right)^{2}-\left(\frac{b}{2 a}\right)^{2}+\left(y+\frac{c}{2 a}\right)^{2}-\left(\frac{c}{2 a}\right)^{2}+\frac{d}{a}
$$

so the equation $a x^{2}+a y^{2}+b x+c y+d=0$ is equivalent to

$$
\left(x+\frac{b}{2 a}\right)^{2}+\left(y+\frac{c}{2 a}\right)^{2}=\left(\frac{b}{2 a}\right)^{2}+\left(\frac{c}{2 a}\right)^{2}-\frac{d}{a} .
$$

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 $q=\left(-\frac{b}{2 a},-\frac{c}{2 a}\right)$ 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 $\ell$ and $m$ be the perpendicular bisectors of $x y$ and $y z$. Since $x y$ and $y z$ are not collinear, $\ell$ 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 $x y$ and the perpendicular bisector of $x z$, 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 $x z$, 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 $\ell, \ell^{\prime}$ intersect at two points $x, y \in \mathbb{R}^{2}$. Show that the angles of intersection at $x, y$ are equal. In particular, $\ell, \ell^{\prime}$ are perpendicular at $x$ if and only if they are perpendicular at $y$.

### 3.3. 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 3.3.1. 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 $\left|i_{C}(p)-q \| p-q\right|=r^{2}$. 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 $\left|i_{C}(p)-q \| p-q\right|=r^{2}$ 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 3.3.2. 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 $\ell$ be the line through $q, p$ and let $\ell^{\prime}$ be the line perpendicular to $\ell$ through $p$. Suppose that $\ell^{\prime}$ intersects $C$ at $a$ and $b$. Show that the point $d$ on $\ell$ where the lines through $a$ and $b$ tangent to $C$ intersect is equal to $i_{C}(p)$.


In $\S 3.2$, 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 $\ell$ be a line and let $m$ be a line that intersects $\ell$ 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 $r \rightarrow \infty$, the circle $C_{r}$ limits onto the line $\ell$. We claim that for any $p \in \mathbb{R}^{2}$,

$$
\lim _{r \rightarrow \infty} i_{C_{r}}(p)=R_{\ell}(p),
$$

where $R_{\ell}$ is the reflection through $\ell$. So in other words, as $C_{r}$ approaches $\ell$, the inversion through $C_{r}$ approaches the reflection through $\ell$. To see this, note first that as $r \rightarrow \infty$, 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 $r \rightarrow \infty$, the point $x_{r}$ limits to the closest point $x \in \ell$ to $p$. Moreover,

$$
d\left(i_{C_{r}}(p), q_{r}\right)=r^{2} / d\left(p, q_{r}\right) \Longrightarrow d\left(i_{C}(p), x_{r}\right)=\frac{r^{2}}{r-d\left(x_{r}, p\right)}-r,
$$

so as $r \rightarrow \infty$, we have

$$
\lim _{r \rightarrow \infty} \frac{r^{2}}{r-d\left(x_{r}, p\right)}-r=\lim _{r \rightarrow \infty} \frac{r \cdot d\left(x_{r}, p\right)}{r-d\left(x_{r}, p\right)}=d(x, p)
$$

where the last equality is L'Hôpital's rule, using that $x_{r} \rightarrow x$. So, as $r \rightarrow \infty$ we have that $\left.d\left(i_{C_{r}}(p), x_{r}\right)\right) \rightarrow d(x, p)$. This implies that $i_{C_{r}}(p) \rightarrow R_{\ell}(p)$ 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 3.3.3 (Inversions send lircles to lircles). If $C$ is a circle and $C^{\prime}$ is a lircles, then $i_{C}\left(C^{\prime}\right)$ 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 $\S 3.1$ that lircles are solution sets of nondegenerate equations

$$
\begin{equation*}
a x^{2}+a y^{2}+b x+c y+d=0 . \tag{6}
\end{equation*}
$$

So, to show that inversions send lircles to lircles, we just have to check that an equation of the form in (6) 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

$$
i_{C}(x, y)=\frac{r^{2}}{|(x, y)|^{2}}(x, y)=\left(\frac{r^{2} x}{x^{2}+y^{2}}, \frac{r^{2} y}{x^{2}+y^{2}}\right) .
$$

So, plugging in the output into Equation (6) gives

$$
a\left(\frac{r^{2} x}{x^{2}+y^{2}}\right)^{2}+a\left(\frac{r^{2} y}{x^{2}+y^{2}}\right)^{2}+b\left(\frac{r^{2} x}{x^{2}+y^{2}}\right)+c\left(\frac{r^{2} y}{x^{2}+y^{2}}\right)+d=0 .
$$

The numerators of the first two terms combined to give a $x^{2}+y^{2}$, which cancels part of the $\left(x^{2}+y^{2}\right)^{2}$ denominator. So, after simplification this becomes

$$
\begin{equation*}
a r^{4}+b r^{2} x+c r^{2} y+d\left(x^{2}+y^{2}\right)=0 \tag{7}
\end{equation*}
$$

which is a quadratic equation exactly of the form in Equation (6), except that the coefficients have been rearranged and modified. Since the original equation (6) is nondegenerate, its solution set $C^{\prime}$ is a proper subset of $\mathbb{R}^{2}$ with at least 2 points, so the same is true of $i_{C}\left(C^{\prime}\right)$, which is the solution set of (7), 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 (6) describes a line that does not pass through 0
when $a=0$ and $d \neq 0$. In this case, Equation (7) 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 $C^{\prime}$ be a circle with the same radius centered at the origin. Then $i_{C}=T_{q} \circ i_{C^{\prime}} \circ T_{-q}$, so $i_{C}$ takes lircles to lircles as both $i_{C^{\prime}}$ 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 $i_{C^{\prime}}$.

If $\alpha:[a, b] \longrightarrow \mathbb{R}^{2}, \alpha(t)=\left(\alpha_{1}(t), \alpha_{2}(t)\right)$ is a differentiable path, recall that the velocity vector of $\alpha$ at time $t$ is the vector $\alpha^{\prime}(t):=\left(\alpha_{1}^{\prime}(t), \alpha_{2}^{\prime}(t)\right)$ of derivatives of the coordinates of $\alpha$. The speed of $\alpha$ at time $t$ is the length $\left|\alpha^{\prime}(t)\right|$ of the velocity vector. The following theorem compares the velocity of a path $\alpha$ to the velocity of its image $i_{C} \circ \alpha$ under an inversion.

Theorem 3.3.4. 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

$$
\left(i_{C} \circ \alpha\right)^{\prime}(t)=\frac{r^{2}}{|\alpha(t)-q|^{2}} R_{\ell}\left(\alpha^{\prime}(t)\right)
$$

where $\ell$ is the line through the origin perpendicular to the vector $\alpha(t)-q$.
Here, remember that $\alpha^{\prime}(t)$ and $\left(i_{C} \circ \alpha\right)^{\prime}(t)$ 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 $\ell$ 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, $\alpha^{\prime}(t)$ is reflected over the vertical axis, and then scaled up by $r^{2} /|\alpha(t)-q|^{2}$.


We'll prove Theorem 3.3.4 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 $\alpha$ is the path $\alpha(t)=(t, 0)$, then $i_{C}(\alpha(t))=\left(r^{2} / t, 0\right)$, so

$$
\left(i_{C} \circ \alpha\right)^{\prime}(t)=\left(-\frac{r^{2}}{t^{2}}, 0\right)=\frac{r^{2}}{t^{2}} \cdot(-1,0)=\frac{r^{2}}{|(t, 0)-(0,0)|^{2}} \cdot R_{\ell}\left(\alpha^{\prime}(t)\right)
$$

where $\ell$ 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

$$
\left(i_{C} \circ \beta\right)^{\prime}(t)=\frac{r^{2} / s}{s} \beta^{\prime}(t)
$$

i.e. the velocity is scaled by the ratio of the two radii. Here, $\beta$ is perpendicular to the vector $\beta(t)-(0,0)$, so reflection through the line $\ell$ perpendicular to $\beta(t)$ fixes $\beta^{\prime}(t)$.

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 $d(q, \alpha(t))=|\alpha(t)|$.

$$
\left(i_{C} \circ \alpha\right)^{\prime}(t)=\frac{d}{d t} r^{2} \frac{\alpha(t)}{|\alpha(t)|^{2}}=r^{2} \frac{\alpha^{\prime}(t)|\alpha(t)|^{2}-\alpha(t) \frac{d}{d t}|\alpha(t)|^{2}}{|\alpha(t)|^{4}}
$$

The mysterious term in the quotient on the right is

$$
\begin{aligned}
\frac{d}{d t}|\alpha(t)|^{2} & =\frac{d}{d t} \alpha_{1}(t)^{2}+\alpha_{2}(t)^{2} \\
& =2 \alpha_{1}(t) \alpha_{1}^{\prime}(t)+2 \alpha_{2}(t) \alpha_{2}^{\prime}(t) \\
& =2 \alpha(t) \cdot \alpha^{\prime}(t)
\end{aligned}
$$

which is no longer so mysterious. Plugging it in,

$$
\left(i_{C} \circ \alpha\right)^{\prime}(t)=\frac{r^{2}}{|\alpha(t)|^{2}}\left(\alpha^{\prime}(t)-2 \alpha(t) \frac{\alpha(t) \cdot \alpha^{\prime}(t)}{|\alpha(t)|^{2}}\right)
$$

Interpreting everything as vectors based at the origin, the vector $b=\alpha(t) \frac{\alpha(t) \cdot \alpha^{\prime}(t)}{|\alpha(t)|^{2}}$ is the projection of $\alpha^{\prime}(t)$ onto $\alpha(t)$, as pictured below. So, $\alpha^{\prime}(t)-2 b$ is the reflection of $\alpha^{\prime}(t)$ over the line perpendicular to the vector $\alpha(t)$.


This means that $\left(i_{C} \circ \alpha\right)^{\prime}(t)$ is obtained from $\alpha^{\prime}(t)$ 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 3.3.4, we have:
Corollary 3.3.5. Suppose $C$ is a circle in $\mathbb{R}^{2}$ with center $q$. Then $i_{C}$ is conformal, or angle preserving: if two paths $\alpha$ and $\beta$ intersect at an angle $\theta$ at $p \neq q$, then the paths $i_{C} \circ \alpha$ and $i_{C} \circ \beta$ meet with angle $\theta$ at $i_{C}(p)$.

Proof. Theorem 3.3.4 says that $\left(i_{C} \circ \alpha\right)^{\prime}(t)$ and $\left(i_{C} \circ \beta\right)^{\prime}(t)$ are obtained from $\alpha^{\prime}(t)$ and $\beta^{\prime}(t)$ 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.

Here's an application of Corollary 1.2.12 that is an example of the following philosophy: if you can prove something about lircles and angles when one of the lircles is a line, you can also prove it when the lircle is a circle.

Lemma 3.3.6. If $\ell$ is a lircle and $x, y \in \ell$, there is a unique lircle $\ell^{\prime}$ in $\mathbb{R}^{2}$ such that $\ell^{\prime}$ intersects $\ell$ orthogonally, exactly at the points $x, y$.

Proof. Let's first prove the lemma when $\ell$ is a line. If $x=y$, then the perpendicular line to $\ell$ at $x=y$ is the unique such $\ell^{\prime}$. If $x \neq y$, the unique such $\ell^{\prime}$ is the circle centered at the midpoint of the segment $x y \subset \ell$.

Now suppose $\ell$ is a circle. Take a circle $C$ centered at some point $q$ that lies on $\ell$. Then Theorem 3.3.3 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 3.3.5, $i_{C}(m)$ is the unique lircle intersects $\ell$ orthogonally in the points $x, y$.

We can then show:
Theorem 3.3.7. Suppose $C$ is a circle in $\mathbb{R}^{2}$, and $\ell$ is a lircle.
(a) If $\ell$ is orthogonal to $C$, then $i_{C}(\ell)=\ell$.
(b) Conversely, if $\ell$ contains some point $p \notin C$ and its inversion $i_{C}(p)$, then $\ell$ is orthogonal to $C$.

In particular, $\ell$ is orthogonal to $C$ if and only if $i_{C}(\ell)=\ell$, it's just that for the backwards direction all you need is that $\ell$ contains the inversion of one of its points, instead of all of them.

Proof of Theorem 3.3.7. Suppose $\ell$ intersects $C$ orthogonally at $x, y$. As noted above, $i_{C}(\ell)$ also intersects $C$ orthogonally at $x, y$, and Theorem 3.3.3 says that $i_{C}(\ell)$ is a lircle, so Lemma 3.3.6 says that $i_{C}(\ell)=\ell$.

For the reverse direction, suppose $\ell$ contains $p \notin C$ and $i_{C}(p)$. If we pick some $x \in C$, then the lircles $\ell$ 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 3.3.5 then says that the two angles $\theta$ below are equal, and therefore are $\pi / 2$. So, $\ell$ is orthogonal to $C$.


Corollary 3.3.8. 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 $\ell$ 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 $\ell$ 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 $\ell, m$ be two lircles that intersect $C$ orthogonally at $p$. Then $\ell, m$ 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 $\ell$ 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 3.3.7, both $\ell$ and $m$ are orthogonal to $C$, so we have $f(\ell)=\ell$ and $f(m)=m$. So, $f(p) \in \ell \cap m=\left\{p, i_{C}(p)\right\}$.

Finally, by continuity, either $f(p)=p$ for all $p$ or $f(p)=i_{C}(p)$ for all $p$.

### 3.3.1. Exercises.

Exercise 3.3.9. Suppose that $C$ and $C^{\prime}$ are two circles with the same center $q$ and radii $r, r^{\prime}$. Show that the composition $i_{C^{\prime}} \circ i_{C}$ is the dilation $D_{q, \lambda}$ around $q$ by a factor of $\lambda$, where $\lambda=\left(\frac{r^{\prime}}{r}\right)^{2}$. (See the end of Section 1.2.)

ExErcise 3.3.10. If $\ell$ is a line through a point $p$ and $q \neq p$, show that there is a unique lircle $C$ that is tangent to $\ell$ 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 3.3.11. 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 3.3 .8 and its analogue for reflections useful.

A hot pursuit in the 18th and 19th centuries was to construct machines to convert between rotational motion and linear motion. This 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: $a p=a q$ and $p r=q r=p s=s q$, where for brevity these are the lengths of the rods with the indicated endpoints.


Exercise 3.3.12. Show that $s$ is the inversion of $r$ through a circle with center $a$ and radius $\sqrt{a p^{2}-p r^{2}}$.

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 3.3.13. Augment the linkage above with an additional bar and anchor to create a new linkage in which $r$ is constrained to move along an arc of a circle, and $s$ is moves along a line segment. 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

$$
C=\left\{p \in \mathbb{R}^{n}| | p-q \mid=r\right\}
$$

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 3.3.14. 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 3.3.15. 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 3.3.16. 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 $\left(E_{i}\right)$ is a Steiner chain.


Exercise 3.3.17. Assume the circles $D$ and $C$ are concentric with radii $r$ and $s$, respectively. Show that a chain $\left(E_{i}\right)$ 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 3.3.18 (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 3.3.19. 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. First, we prove:
ExERCISE 3.3.20. Suppose that $\ell$ is a line in $\mathbb{R}^{2}$, that $C$ is a circle, and that $\ell \cap C=\varnothing$. Show that there is a circle $D \subset \mathbb{R}^{2}$ that is perpendicular to $\ell$ and $C$.

ExERCISE 3.3.21. Using an inversion and the previous exercise, show that if $C, D$ are non-intersecting circles in $\mathbb{R}^{2}$, then there are two intersecting lircles that are both perpendicular to both $C, D$.

Exercise 3.3.22. Prove the lemma, using the previous exercise.

Exercise 3.3.23. The cross ratio of four points $a, b, c, d \in \mathbb{R}^{2}$ is the ratio

$$
C R(a, b, c, d)=\frac{\frac{a c}{b d}}{\frac{b c}{b d}}
$$

where for brevity we write $a c:=|a-c|$, and similarly for the other distances. If $C$ is a circle in $\mathbb{R}^{2}$ centered at some $q$ that is not any of the points above, show that

$$
C R\left(i_{C}(a), i_{C}(b), i_{C}(c), i_{C}(d)\right)=C R(a, b, c, d) .
$$

That is, inversions preserve cross ratio!

### 3.4. 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. In $\mathbb{R}^{2}$, lines are shortest paths, and in this section we show that hyperbolic lines are also shortest paths in $\mathbb{H}^{2}$, with respect to a different notion of path length.

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.


This is similar to a phenomenon you may have seen when looking at shortest aerial paths on maps of the Earth. For instance, the shortest path from New York to London curves strangely upward toward Greenland when viewed in a map using the Mercator projection, to take advantage of the fact that distances near the poles are much smaller than they appear in the map.

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 3.4.1 (Hyperbolic length). Suppose that $\gamma:[a, b] \longrightarrow \mathbb{H}^{2}$ is a path, where $\gamma(t)=\left(\gamma_{1}(t), \gamma_{2}(t)\right)$. The hyperbolic length of $\gamma$ is defined to be

$$
\text { length }_{\mathbb{H}^{2}}(\gamma)=\int_{a}^{b}\left|\gamma^{\prime}(t)\right| \frac{1}{\gamma_{2}(t)} d t
$$

Recall that $\left|\gamma^{\prime}(t)\right|$ is the Euclidean speed of $\gamma$, which integrates to the Euclidean length of $\gamma$. The integrand $\left|\gamma^{\prime}(t)\right| \frac{1}{\gamma_{2}(t)}$ is called the hyperbolic speed of $\gamma$. 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 $\gamma$ 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 3.4.2. 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 $\alpha_{2}(t)=y$,

$$
\operatorname{length}_{\mathbb{H}^{2}}(\alpha)=\int_{a}^{b}\left|\alpha^{\prime}(t)\right| \frac{1}{y} d t=\int_{a}^{b}\left|\alpha^{\prime}(t)\right| \frac{1}{y} d t=\frac{1}{y} \text { length }(\alpha),
$$

so hyperbolic length is just Euclidean length divided by height.
Example 3.4.3. Let's now compute the length of a vertical line segment from $\left(x, y_{1}\right)$ to $\left(x, y_{2}\right)$ in $\mathbb{H}^{2}$. Parametrically, $\gamma:\left[y_{1}, y_{2}\right] \longrightarrow \mathbb{H}^{2}$, where $\gamma(t)=(x, t)$. Well,

$$
\begin{aligned}
\text { length }_{\mathbb{H}^{2}}(\gamma) & =\int_{y_{1}}^{y_{2}}\left|\gamma^{\prime}(t)\right| \frac{1}{\gamma_{2}(t)} d t \\
& =\int_{y_{1}}^{y^{\prime}} 1 \cdot \frac{1}{t} d t \\
& =\ln \left(y_{2}\right)-\ln \left(y_{1}\right) \\
& =\ln \left(\frac{y_{2}}{y_{1}}\right)
\end{aligned}
$$

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 3.4.4. The hyperbolic distance between two points $p, q \in \mathbb{H}^{2}$ is

$$
d_{\mathbb{H}^{2}}(p, q)=\inf \left\{\operatorname{length}_{\mathbb{H}^{2}}(\gamma) \mid \gamma:[a, b] \longrightarrow \mathbb{H}^{2}, \gamma(a)=p, \gamma(b)=q\right\}
$$

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 3.4.5. 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 $\gamma$ 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 3.4.2 and 3.4.3, these three segments have length $\ln (b-a), 1$ and $\ln (b-a)$, respectively, so

$$
d_{\mathbb{H}^{2}}((a, 1),(b, 1)) \leqslant \operatorname{length}_{\mathbb{H}^{2}}(\gamma)=2 \ln (b-a)+1
$$

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 3.4.6. The path with the shortest hyperbolic length between points $p=$ $\left(x, y_{1}\right)$ and $q=\left(x, y_{2}\right)$ 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 $\ln \left(y_{2} / y_{1}\right)$. So if $\gamma:[a, b] \longrightarrow \mathbb{H}^{2}$ is a path from $p$ to $q$, we claim that length $\mathbb{H}^{2}(\gamma) \geqslant \ln \left(y_{2} / y_{1}\right)$, with equality if and only if $\gamma$ is a vertical line segment.

Writing $\gamma(t)=\left(\gamma_{1}(t), \gamma_{2}(t)\right)$, we compute:

$$
\begin{aligned}
\operatorname{length}_{\mathbb{H}^{2}}(\gamma) & =\int_{a}^{b}\left|\gamma^{\prime}(t)\right| \frac{1}{\gamma_{2}(t)} d t \\
& =\int_{a}^{b} \sqrt{\gamma_{1}^{\prime}(t)^{2}+\gamma_{2}^{\prime}(t)^{2}} \frac{1}{\gamma_{2}(t)} d t \\
& \geqslant \int_{a}^{b} \sqrt{0^{2}+\gamma_{2}^{\prime}(t)^{2}} \frac{1}{\gamma_{2}(t)} d t \\
& \geqslant \int_{a}^{b} \gamma_{2}^{\prime}(t) \frac{1}{\gamma_{2}(t)} d t \\
& =\ln \left(\gamma_{2}(b)\right)-\ln \left(\gamma_{2}(a)\right) \\
& =\ln \left(y_{2} / y_{1}\right)
\end{aligned}
$$

as desired. Equality occurs exactly when $\gamma_{1}^{\prime}(t)=0$ and $\gamma_{2}^{\prime}(t)>0$ for all $t$, which means that $\gamma$ 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 3.4.7. If $\ell$ 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 $\ell$ 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 3.4.8. If $\ell$ is a hyperbolic line and $\gamma$ is a path in $\mathbb{H}^{2}$, then

$$
\operatorname{length}_{\mathbb{H}^{2}}\left(R_{\ell} \circ \gamma\right)=\text { length }_{\mathbb{H}^{2}}(\gamma)
$$

So, $R_{\ell}$ is a hyperbolic isometry: $d_{\mathbb{H}^{2}}\left(R_{\ell}(p), R_{\ell}(q)\right)=d_{\mathbb{H}^{2}}(p, q)$ 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 $t \in[a, b]$,

$$
\begin{equation*}
\left|\gamma^{\prime}(t)\right| \frac{1}{\gamma_{2}(t)}=\left|\left(R_{\ell} \circ \gamma\right)^{\prime}(t)\right| \frac{1}{\left(R_{\ell} \circ \gamma\right)_{2}(t)} \tag{8}
\end{equation*}
$$

That is, the hyperbolic speeds of $\gamma$ and $R_{\ell} \circ \gamma$ agree at time $t$. Integrating, this implies that length $\mathbb{H}^{2}(\gamma)=\operatorname{length}_{\mathbb{H}^{2}}\left(R_{\ell} \circ \gamma\right)$, so $R_{\ell}$ preserves path lengths. As hyperbolic distance is defined in terms of path lengths, $R_{\ell}$ must be an isometry.
Case 1. If $\ell$ is a vertical half line, then $R_{\ell}$ is a Euclidean isometry. Therefore, the Euclidean speeds of $\gamma$ and $R_{\ell} \circ \gamma$ must agree for all $t$. Furthermore, since $\ell$ is vertical, the heights of $\gamma(t)$ and $R_{\ell} \circ \gamma(t)$ are the same. So, we have

$$
\left|\gamma^{\prime}(t)\right|=\left|\left(R_{\ell} \circ \gamma\right)^{\prime}(t)\right| \quad \text { and } \quad \gamma_{2}(t)=\left(R_{\ell} \circ \gamma\right)_{2}(t)
$$

from which Equation (8) follows.
Case 2. Suppose now that $\ell$ is a semicircle with center $c$ and radius $r$. In this case, $R_{\ell}$ preserves neither height nor Euclidean speed, and the goal is to show that the height distortion exactly matches the speed distortion, i.e.

$$
\frac{\left|\left(R_{\ell} \circ \gamma\right)^{\prime}(t)\right|}{\left|\gamma^{\prime}(t)\right|}=\frac{\left(R_{\ell} \circ \gamma\right)_{2}(t)}{\gamma_{2}(t)} .
$$

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

$$
\begin{equation*}
\frac{\left(R_{\ell} \circ \gamma\right)_{2}(t)}{\gamma_{2}(t)}=\frac{\left|R_{\ell} \circ \gamma(t)-c\right|}{|\gamma(t)-c|}=\frac{\frac{r^{2}}{|\gamma(t)-c|}}{|\gamma(t)-c|}=\frac{r^{2}}{|\gamma(t)-c|^{2}} \tag{9}
\end{equation*}
$$

This matches the speed distortion, by Theorem 3.3.4, so Equation (8) follows.
Corollary 3.4.9. If $p, q \in \mathbb{H}^{2}$, the (hyperbolically) shortest path from $p$ to $q$ is the hyperbolic line segment joining them.

Proof. Let $\ell$ 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 $\ell$ 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 3.4.6, 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 $\ell$.

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 ${ }^{3}$, 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.

[^4]

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 3.4.10. In Example 3.4.5, 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

$$
\begin{equation*}
d_{\mathbb{H}^{2}}((-x, 1),(x, 1))=2 \ln \cot \left(\frac{\cot ^{-1}(x)}{2}\right) . \tag{10}
\end{equation*}
$$

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 3.4.5, which in this case is $2 \ln (2 x)$. It turns out that

$$
\lim _{x \rightarrow \infty} \frac{2 \ln \cot \left(\frac{\cot ^{-1}(x)}{2}\right)}{2 \ln (2 x)}=1
$$

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 3.4.5 form a somewhat length-efficient approximation of the hyperbolic line segment.

ExERCISE 3.4.11. You and a friend walk upwards along the vertical half lines $x=a$ and $x=b$ at unit hyperbolic speed. Parametrically, your paths are $\alpha$ and $\beta$, where

$$
\alpha(t)=\left(a, e^{t}\right), \quad \beta(t)=\left(b, e^{t}\right)
$$

since the hyperbolic speeds at time $t$ are

$$
\frac{\left|\alpha^{\prime}(t)\right|}{\alpha_{2}(t)}=\frac{\left|\beta^{\prime}(t)\right|}{\beta_{2}(t)}=\frac{\sqrt{0^{2}+\left(e^{t}\right)^{2}}}{e^{t}}=1 .
$$

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 $\alpha, \beta$ in $\mathbb{H}^{2}$ are asymptotic if they can be parameterized so that

$$
\lim _{t \rightarrow \infty} d_{\mathbb{H}^{2}}(\alpha(t), \beta(t))=0 .
$$

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 3.4.12. Suppose that two hyperbolic lines $\ell, \ell^{\prime}$ share an endpoint on the $x$-axis, as pictured below. Show that there is some hyperbolic line $m$ such that $R_{m}(\ell)$ and $R_{m}\left(\ell^{\prime}\right)$ are vertical half lines, and use this and the previous exercise to show that $\ell$ and $\ell^{\prime}$ are asymptotic. For the last part, you will need to parameterize $\ell$ and $\ell^{\prime}$ 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 3.4.13. Show that if $p \in \mathbb{H}^{2}$ and $v$ is a vector based at $p$, there is a unique hyperbolic line $\ell$ passing through $p$ tangent to $v$.


### 3.5. Hyperbolic trigonometric functions

We now know that the hyperbolically shortest path between points $p, q \in \mathbb{H}^{2}$ is the hyperbolic line segment connecting them. What's the length of this path? In other words, can we find a formula for $d(p, q)$ in terms of the coordinates of $p, q$ ? To do this, it will be convenient to intoduce the hyperbolic trigonometric functions sinh, cosh, tanh, sech and csch. These are pronounced 'hyperbolic sine, hyperbolic cosine, hyperbolic tangent, etc...', or more informally as 'sinsh, cosh, tansh, etc...', and are defined as follows.

$$
\begin{array}{lll}
\sinh (t)=\frac{e^{t}-e^{-t}}{2}, & \cosh (t)=\frac{e^{t}+e^{-t}}{2}, & \tanh (t)=\frac{\sinh (t)}{\cosh (t)} \\
\operatorname{csch}(t)=\frac{1}{\sinh (t)}, & \operatorname{sech}(t)=\frac{1}{\cosh (t)}, & \operatorname{coth}(t)=\frac{1}{\tanh (t)}
\end{array}
$$

Hyperbolic trigonometric functions have many features that are similar to their Euclidean counterparts. For instance, easy calculations show that

$$
\sinh ^{\prime}(t)=\cosh (t), \quad \cosh ^{\prime}(t)=\sinh (t), \quad \tanh ^{\prime}(t)=\frac{1}{\cosh ^{2}(t)}
$$

There are also addition formulas similar to those in the Euclidean case.

$$
\begin{aligned}
\sinh (t+s) & =\sinh (t) \cosh (s)+\cosh (t) \sinh (s) \\
\cosh (t+s) & =\cosh (t) \cosh (s)+\sinh (t) \sinh (s)
\end{aligned}
$$

You can find many more identities on the Wikipedia page 'Hyperbolic function'. However, there's one identity we should discuss more thoroughly:

$$
\begin{align*}
\cosh ^{2}(t)-\sinh ^{2}(t) & =\frac{\left(e^{t}+e^{-t}\right)^{2}-\left(e^{t}-e^{-t}\right)^{2}}{4}  \tag{11}\\
& =\frac{e^{2 t}+2+e^{-2 t}-e^{2 t}+2-e^{-2 t}}{4} \\
& =1
\end{align*}
$$

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 (11) 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 $t \gg 0$, 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 $t \rightarrow \infty$ and tanh has horizontal asymptotes at $\pm 1$.

3.5.1. 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 $\lambda$ 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

$$
T \cos (\theta)=-T_{0}, \quad T \sin (\theta)=-\lambda g s, \Longrightarrow \tan (\theta)=\frac{\lambda g}{T_{0}} s
$$

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 $f^{\prime}(t)$. Combine $T_{0}, \lambda$ and $g$ into one constant $a=\frac{T_{0}}{\lambda g}$. Then our equation becomes $f^{\prime}(t)=s / a$. That is,
$(\star)$ The length of the graph of $f(x)$ from $x=0$ to $x=t$ is a $f^{\prime}(t)$.
Exercise 3.5.1. 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

$$
\text { length }(\gamma)=\int_{0}^{t}\left|\gamma^{\prime}(x)\right| d x=\int_{0}^{t} \sqrt{1+f^{\prime}(x)^{2}} d x
$$

3.5.2. 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,

$$
\sinh : \mathbb{R} \longrightarrow \mathbb{R}, \quad \sinh (x)=\frac{1}{2}\left(e^{x}-e^{-x}\right)
$$

is a bijection, so it has an inverse function $\sinh ^{-1}: \mathbb{R} \longrightarrow \mathbb{R}$. In fact,

$$
\sinh ^{-1}(y)=\ln \left(y+\sqrt{1+y^{2}}\right)
$$

which we can verify by plugging in $y=\sinh (x)$ as follows:

$$
\ln \left(\sinh (x)+\sqrt{1+\sinh (x)^{2}}\right)=\ln (\sinh (x)+\cosh (x))=\ln e^{x}=x
$$

Exercise 3.5.2. Hyperbolic cosine restricts to a bijection

$$
\cosh :[0, \infty) \longrightarrow[1, \infty)
$$

Show that for $y \in[1, \infty)$, we have

$$
\cosh ^{-1}(y)=\ln \left(y+\sqrt{y^{2}-1}\right)
$$

In Section 3.4, 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 3.5.3. The distance between $p=\left(p_{1}, p_{2}\right)$ and $q=\left(q_{1}, q_{2}\right)$ in $\mathbb{H}^{2}$ is

$$
d_{\mathbb{H}^{2}}(p, q)=2 \sinh ^{-1}\left(\frac{|p-q|}{2 \sqrt{p_{2} q_{2}}}\right) .
$$

For example, we saw in Example 3.4.3 that when $p, q$ lie on a vertical line, say with $p_{2} \geqslant q_{2}$, their hyperbolic distance is given by the formula $\ln \left(p_{2} / q_{2}\right)$. Since

$$
\begin{aligned}
\sinh \left(\frac{1}{2} \ln \frac{p_{2}}{q_{2}}\right) & =\frac{1}{2}\left(e^{\ln \sqrt{\frac{p_{2}}{q_{2}}}}-e^{-\ln \sqrt{\frac{p_{2}}{q_{2}}}}\right) \\
& =\frac{1}{2}\left(\sqrt{\frac{p_{2}}{q_{2}}}-\sqrt{\frac{q_{2}}{p_{2}}}\right) \\
& =\frac{p_{2}-q_{2}}{2 \sqrt{p_{2} q_{2}}}
\end{aligned}
$$

we have that Theorem 3.5.3 is true when $p, q$ lie on a vertical line. To prove the theorem in general, let's define $D$ to be the right-hand side

$$
D(p, q)=2 \sinh ^{-1}\left(\frac{|p-q|}{2 \sqrt{p_{2} q_{2}}}\right) .
$$

EXERCISE 3.5.4. If $\ell$ is a hyperbolic line, show that $D\left(R_{\ell}(p), R_{\ell}(q)\right)=D(p, q)$ for all $p, q \in \mathbb{H}^{2}$. Hint: the figure below depicts the reflection of $p, q$ through $\ell$. The points $p, q$ are assumed to be at distances $a, b$ from the center of $\ell$. Show that

$$
\left|R_{\ell}(p)-R_{\ell}(q)\right|^{2}=\frac{r^{4}}{(a b)^{2}}|p-q|^{2}, \quad\left(R_{\ell}(p)\right)_{2}=\frac{r^{2}}{a^{2}} p_{2}, \quad \text { and } \quad\left(R_{\ell}(q)\right)_{2}=\frac{r^{2}}{b^{2}} q_{2}
$$

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 3.5.5. Using the previous two problems, prove the theorem in general. Hint: if $p, q \in \mathbb{H}^{2}$, let $\ell$ 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 $R(q)$.

ExERCISE 3.5.6. Prove the following alternative distance formula:

$$
d_{\mathbb{H}^{2}}(p, q)=\cosh ^{-1}\left(1+\frac{|p-q|^{2}}{2 p_{2} q_{2}}\right),
$$

either by transforming the one above into this, or by repeating the proof structure above.
3.5.3. Hyperbolic circles. Here is a cool application of the distance formula above. If $p \in \mathbb{H}^{2}$ and $r>0$, the hyperbolic circle of radius $r$ centered at $p$ is

$$
C_{\mathbb{H}^{2}}(p, r):=\left\{q \in \mathbb{H}^{2} \mid d_{\mathbb{H}^{2}}(p, q)=r\right\} .
$$

We will call $p$ the hyperbolic center of $C$, and $r$ the hyperbolic radius.
Remark 3.5.7. A priori it may be possible that we have $C_{\mathbb{H}^{2}}(p, r)=C_{\mathbb{H}^{2}}(q, s)$, where $p \neq q$, in which case 'the' hyperbolic center of a hyperbolic circle may not be well defined. This actually happens in spherical geometry: on a unit sphere, the circle of radius $\pi / 2$ centered at the north pole is the equator, which is also the circle of radius $\pi / 2$ centered at the south pole, so the 'center' of the equator is not well defined. However, if $p, q \in \mathbb{H}^{2}$, hyperbolic circles $C_{\mathbb{H}^{2}}(p, r)=C_{\mathbb{H}^{2}}(q, s)$ only agree if $p=q$ and $r=s$. Indeed, let $\ell$ be the hyperbolic line through $p, q$, and suppose for instance that $s \geqslant r$. Then the point $z \in \ell$ with $d_{\mathbb{H}^{2}}(q, z)=s$ that lies on the opposite side of $q$ from $p$ must lie in $C_{\mathbb{H}^{2}}(q, s)$, but can't lie in $C_{\mathbb{H}^{2}}(p, s)$ since $d_{\mathbb{H}^{2}}(p, z)=d_{\mathbb{H}^{2}}(p, q)+s>r$.

The following is a bit surprising!
FACT 3.5.8. Every hyperbolic circle in $\mathbb{H}^{2}$ is also a Euclidean circle in $\mathbb{R}^{2}$.
Note that we are not saying that $C_{\mathbb{H}^{2}}(p, r)$ is the Euclidean circle of radius $r$ centered at $p$. Indeed, while $C_{\mathbb{H}^{2}}(p, r)$ is a Euclidean circle, its Euclidean center and radius will always be different than $p$ and $r$.

Proof. By the distance formula, given $p=\left(p_{1}, p_{2}\right)$ and $r>0$ the hyperbolic circle $C_{\mathbb{H}^{2}}(p, r)$ is the set of points $q=(x, y) \in \mathbb{H}^{2}$ such that

$$
\begin{gather*}
r=2 \sinh ^{-1}\left(\frac{|p-q|}{2 \sqrt{p_{2} y}}\right) \\
\Longleftrightarrow \sinh \left(\frac{r}{2}\right)=\frac{|p-q|}{2 \sqrt{p_{2} y}} \\
\Longleftrightarrow 4 \sinh \left(\frac{r}{2}\right)^{2} p_{2} y=\left(x-p_{1}\right)^{2}+\left(y-p_{2}\right)^{2} \tag{12}
\end{gather*}
$$

This is a quadratic equation in the variables $x, y$ with no $x y$ terms, and where the coefficients of $x^{2}$ and $y^{2}$ are both 1 . If $r>0$, there are at least two solutions to this equation, since the points $\left(p_{1}, e^{r} p_{2}\right)$ and $\left(p_{1}, e^{-r} p_{2}\right)$ both have distance $r$ to $p$. So, by Fact 3.2.2, the equation (12) describes a circle.

Exercise 3.5.9. By completing the square in (12), find the Euclidean center and radius of $C_{\mathbb{H}^{2}}(p, r)$ in terms of $p=\left(p_{1}, p_{2}\right)$ and $r$. Then show that every Euclidean circle contained in $\mathbb{H}^{2}$ is also a hyperbolic circle.

In Euclidean geometry, a line $\ell \subset \mathbb{R}^{2}$ passes through the center of a circle $C \subset \mathbb{R}^{2}$ if and only if $\ell \perp C$, which happens if and only if the reflection $R_{\ell}(C)=C$. Prove this as an exercise if you like! Here is the hyperbolic analogue .

Lemma 3.5.10. If $C:=C_{\mathbb{H}^{2}}(p, r)$ is a hyperbolic circle and $\ell$ is a hyperbolic line, then $\ell$ passes through the hyperbolic center of $C \Longleftrightarrow \ell \perp C \Longleftrightarrow R_{\ell}(C)=C$, where $R_{\ell}$ is the hyperbolic reflection through $\ell$.

Proof. Since $d_{\mathbb{H}^{2}}\left(R_{\ell}(q), R_{\ell}(p)\right)=d_{\mathbb{H}^{2}}(q, p)$, the former is $r$ if and only if the latter is, implying that $R_{\ell}\left(C_{\mathbb{H}^{2}}(p, r)\right)=C_{\mathbb{H}^{2}}\left(R_{\ell}(p), r\right)$. Therefore $p \in \ell \Longleftrightarrow R_{\ell}(p)=$ $p \Longleftrightarrow R_{\ell}(C)=C$. If $\ell$ is a semicircle, Theorem 3.3 .7 says that $R_{\ell}(C)=C \Longleftrightarrow$ $\ell \perp C$. If $\ell$ is a vertical half-line, then $R_{\ell}$ is a Euclidean reflection, and we have $R_{\ell}(C)=C \Longleftrightarrow \ell \perp C$ as mentioned above.

This lemma can be used to give a ruler and compass construction of the hyperbolic center of a hyperbolic circle $C$ with Euclidean center $c$. Let $a$ be the point at which the vertical line $\ell$ 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 $\ell$ and $m$ pass through the hyperbolic center $p$ of $C$, which must be their point of intersection.


ExErcise 3.5.11. 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 $\sqrt{a b}$, the geometric mean. Hint: the ruler and compass construction is not the most efficient way to do this.

The inequality of means states that $\sqrt{a b} \leqslant \frac{a+b}{2}$. This makes sense in terms of the exercise above: since hyperbolic distances becomes smaller relative to Euclidean distance when height is increased, the hyperbolic center of hyperbolic circle should be lower down than the Euclidean center.

### 3.6. 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 $2 \pi$ 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 $x y$-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 3.6.1. The tractrix can be prescribed parametrically as

$$
\gamma:[0, \infty) \longrightarrow \mathbb{R}^{2}, \quad \gamma(t)=(t-\tanh (t), \operatorname{sech}(t))
$$

ExErcise 3.6.2. 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

$$
\gamma(t)+\frac{1}{\left|\gamma^{\prime}(t)\right|} \gamma^{\prime}(t)=(t, 0) .
$$

(b) Show that if $\gamma(t)=(t-\tanh (t), \operatorname{sech}(t))$, then

$$
\gamma^{\prime}(t)=\left(\tanh ^{2}(t),-\tanh (t) \operatorname{sech}(t)\right), \quad\left|\gamma^{\prime}(t)\right|=\tanh (t)
$$

(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 $\gamma^{\prime}(0)=(0,0)$. 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

$$
\begin{aligned}
\int_{0}^{a}\left|\gamma^{\prime}(t)\right| d t & =\int_{0}^{a} \tanh (t) d t \\
& =\left.\ln \cosh (t)\right|_{0} ^{a} \\
& =\ln \cosh (a)
\end{aligned}
$$

This arc length is $s$ when $a=\cosh ^{-1}\left(e^{s}\right)$, at which point the height of the tractrix is

$$
\operatorname{sech}\left(\cosh ^{-1}\left(e^{a}\right)\right)=e^{-s},
$$

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

$$
\alpha(t)=\gamma(t)-\left(\int_{0}^{t}\left|\gamma^{\prime}(s)\right| d s\right) \frac{\gamma^{\prime}(t)}{\left|\gamma^{\prime}(t)\right|}
$$

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 $\gamma^{\prime}(t)$ for a distance that is equal to the length of $\gamma$ from 0 to $t$. In other words, the subtracted term on the right represents the blue lines in the picture above.

ExERCISE 3.6.3. 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 3.6.4. 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 $\gamma$ 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 $\gamma$ and its involute, as it'll be a bit difficult to do by hand.


[^0]:    ${ }^{1}$ 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

[^1]:    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.

[^2]:    ${ }^{1}$ 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.

[^3]:    ${ }^{1}$ 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.

[^4]:    ${ }^{3}$ This assumes pressure is the dominant factor in determining the speed of sound. Temperature also plays a role, so the picture is a bit more complicated than presented here.

