PMath 321 - Non-Euclidean geometry

Course Notes, University of Waterloo

Tyrone Ghaswala

Spring 2025

Contents

These notes are for the Spring 2025 offering of PMath 321 - Non-Euclidean geometry. They will be updated as I go, and are definitely not free of typos and mistakes. If you find any, please let me know about it (either by email or through Piazza)!

Lecture 1 - 05/05

1 Introduction

Geometry is fundamental to human civilisation. This is undeniable if one observes the organisation of cities, or the construction of amazing structures (think the Pyramids, for example). However, geometry was not formalised in a mathematical sense until around the 5th century BC, when Greek mathematician Thales of Miletus applied deductive reasoning to geometry (at least our best guess is that he was the first). A few hundred years later, Euclid came along and wrote The Elements, a collection of 13 books that spectacularly organised mathematics.

The first book deals with the foundations of geometry. Euclid states 23 definitions, followed by 5 postulates, and 5 common notions. He then goes on to carefully prove 48 propositions in geometry. The common notions and postulates are what we would call axioms today. You can read about all of these at this wonderful interactive web-version of Euclid's elements: http://aleph0.clarku.edu/~djoyce/elements/bookI/bookI.html.

For example, here are some of the definitions:

1. A **point** is that which has no part.

2. A **line** is a breadthless length.

3. The ends of a line are points.

10. When a straight line standing on a straight line makes the adjacent angles equal to one another, each of the equal angles is **right**, and the straight line standing on the other is called a **perpendicular** to that on which it stands.

23. **Parallel** straight lines are straight lines which, being in the same plane and being produced indefinitely in both directions, do not meet one another in either direction.

And it goes on like this. These definitions seem a little imprecise by today's standards, but remember, this is an English translation, and it's also the first time in human history (to our best guess) that definitions and axioms were laid out! Here are the common notions, followed in parentheses what they say in modern notation:

1. Things which equal the same thing also equal one another. [*If* $a = c$ *and* $b = c$*, then* $a = b$.]

2. If equals are added to equals, then the wholes are equal. [*If* $a = b$ *and* $c = d$*, then* $a+c = b+d$.]

3. If equals are subtracted from equals, then the remainders are equal. [*If* $a = b$ *and* $c = d$*, then* $a - c = b - d$.]

4. Things which coincide with one another equal one another. [$a = a$.]

5. The whole is greater than the part. [*If* $a, b \geq 0$*, then* $a + b \geq a$.]

The postulates are where things get interesting for us. These are the axioms of what we now call **Euclidean geometry**.

1. A straight line may be drawn from any point point to any other point.
2. A terminated line can be produced indefinitely.
3. A circle can be drawn with any centre and any radius.
4. All right angles are equal to one another.
5. If a straight line falling on two straight lines makes the interior angles on the same side of it taken together less than two right angles, then the two straight lines, if produced indefinitely, meet on that side on which the sum of angles is less than two right angles.

The fifth postulate is a mouthful! We'll come back to this. Using these postulates, Euclid then carefully deduces a bunch of geometry facts. For example,

Proposition 20: In any triangle the sum of any two sides is greater than the remaining one.

Proposition 32: In any triangle, if one of the sides is produced, then the exterior angle equals the sum of the two interior and opposite angles, and the sum of the three interior angles of the triangle equals two right angles.

In more modern terminology, Proposition 20 is the triangle inequality, and Proposition 32 states that the sum of the angles of a triangle is 180°.

The geometry most of us are used to, modelled by $\mathbb{R}^n$ with the dot product, statisfies Euclid's postulates, and therefore all Propositions proved in Book 1 of The Elements hold in $\mathbb{R}^n$ with the dot product.

Lecture 2 - 07/05

The fifth postulate seems out of place, and many mathematicians tried in vain to prove that the fifth postulate followed from the other four. Here are some statements that are equivalent to the fifth postulate. That is, we could replace the fifth postulate with any of the following and still deduce the same theorems.

1. **Proclus' axiom**: If a line intersects one of two parallel lines, it must also intersect the other.
2. **Playfair's axiom**: Through a point not on a given line there exists a unique line parallel to the given line.
3. The sum of the angles of a triangle is two right angles.
4. Similar triangles exsit which are not congruent.
5. Two lines parallel to the the same line are parallel to each other.
6. Any three distinct noncollinear points have a circle going through them.

Although each of these statements is completely acceptable as being true, none of them can be proved from just the first four postulates.

In the early 18th century (AD of course), Italian mathematician Girolamo Saccheri made a valiant attempt at proving Euclid's fifth postulate from the other four. Although he was unsuccessful, he proved the first results of what is now called elliptic and hyperbolic geometry. Geometry derived from the first four postulates (or a different set of axioms altogether) is called **non-Euclidean geometry**.

For example, consider the surface of a sphere. If we define points to be points on the surface, and lines to be great circles (the interection of a plane passing through the centre of the sphere and the surface of the sphere), then it is no longer true that the sum of the angles of a triangle is 180°! Indeed, consider a triangle with one vertex on the north pole, and two vertices on the equator. The angle sum is 270°! Wild.

In this course we will study non-Euclidean geometries.

2 Circle inversion

The first part of the course will be a study of hyperbolic geometry. In order to get there, we must first, perhaps surprisingly, do some regular, good ol' fashioned, Euclidean geometry.

Consider the plane $\mathbb{R}^2$ with our usual notion of distance and angle (given by the dot product). A reflection across any line preserves length and angle, and is therefore referred to as an **isometry** of $\mathbb{R}^2$. A reflection R has the additional property that if you do it twice, every point gets mapped back to itself! Such a transformation of the plane is called an **involution**. Symbolically, $R \circ R(x) = x$ for all $x \in \mathbb{R}^2$.

Our path towards hyperbolic geometry begins with a different type of involution, a circle inversion. Here is an example of a circle inversion.

Example. Consider the map $f : \mathbb{C} \setminus \{0\} \to \mathbb{C} \setminus \{0\}$ defined by $f(z) = \overline{z}^{-1}$. So, $f(1) = 1$, $f(i) = i$, and $f(2) = \frac{1}{2}$. In fact, if we write our complex number in exponential form, we have

$$f(re^{i\theta}) = (re^{-i\theta})^{-1} = \frac{1}{r}e^{i\theta}.$$

So, a point in the complex plane with modulus r gets mapped to a point with modulus $\frac{1}{r}$ pointing in the same direction (that is with the same argument as the original point)!

Furthermore, suppose $|z| = 1$. Then $f(z) = z$, so the circle of radius 1, centred at the origin, is fixed! Evenfurthermore, $f^2(z) = f(\frac{1}{\overline{z}}) = z$ for all $z \in \mathbb{C} \setminus \{0\}$.

In the previous argument, although we haven't defined the function at 0, it is helpful to imagine 0 being mapped to ∞, and ∞ being mapped to 0. We will come back to this idea and make it a bit more formal later on.

For the remainder of this section, we imagine everything occuring in the Euclidean plane, or if you prefer, $\mathbb{R}^2$ with length and angle being given by the dot product.

Lecture 3 - 09/05

Before we define general circle inversions, let's set some notation. For points A and B in the plane, $\overline{AB}$ denotes the line segment between A and B, $\overrightarrow{AB}$ denotes the half-infinite ray starting at A and passing through B, and AB denotes the distance beween A and B.

When convenient, we introduce a point at infinity, call it ∞, which is on every line. So, in particular, any two lines intersect at ∞. This allows us to define circle inversion in a more slick fashion, and to state some properties of circle inversions more concisely. The point ∞ is purely a formal mathematical object introduced to make our lives a little easier.

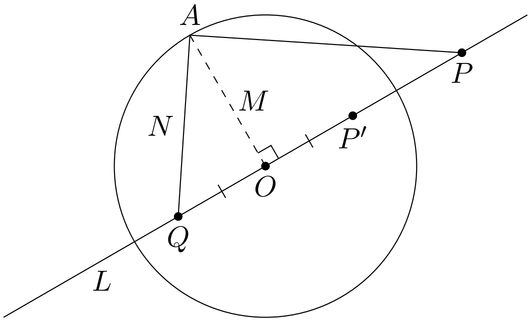
Definition. Let K be a circle with radius r and center O. Let $I_K : \mathbb{R}^2 \cup \{\infty\} \to \mathbb{R}^2 \cup \{\infty\}$ be the map that takes a point $P \notin \{O, \infty\}$ to a point P' on the ray $\overrightarrow{OP}$ so that $OP \cdot OP' = r^2$. Define $I_K(O) = \infty$ and $I_K(\infty) = O$. Then I_K is the circle inversion about the circle K.

Note that we may use $\mathbb{R}^2$ or $\mathbb{C}$ to represent the plane.

Exercise. Prove that $I_K^2(P) = P$ for all $P \in \mathbb{R}^2 \cup \{\infty\}$. This shows that the general circle inversion I_K is an **involution**.

Exercise. Verify that the map $f : \mathbb{C} \cup \{\infty\} \to \mathbb{C} \cup \{\infty\}$ given by $f(z) = \frac{1}{\bar{z}}$ is indeed the inversion in the unit circle in the complex plane.

Here is one way we can visualise what happens to the points under a circle inversion. First, an image, then the description.



Let P be a point in the plane other than the center O of the circle K. Draw the line L passing through OP, and draw a line M perpendicular to L and let A be one of the ponts of intersection of M with K. Draw a line N through A perpendicular to $\overline{PA}$ and let Q be the interesection of N with L. Now, on the ray $\overrightarrow{OP}$, draw a point P' distance OQ from O. Then $I_K(P) = P'$.

Exercise. Prove that $I_K(P) = P'$ in the previous construction.

Now, let's prove our first result about circle inversions.

Proposition 1. *Let L be a line passing through the center O of a circle K. Then $I_K(L) = L$.*

Proof. First note that $I_K(O) = \infty \in L$ and $I_K(\infty) = O \in L$. We now want to show that if $P \in L$ (other than O or ∞), then $I_K(P) \in L$. To this end, note that $I_K(P) \in \overrightarrow{OP} \subset L$. Therefore $I_K(L) \subset L$.

Conversely, suppose $Q \in L$. Then $Q = I_K(I_K(Q))$. However, we already proved that $I_K(Q) \in L$, and therefore $Q \in I_K(L)$. We can now conclude that $L \subset I_K(L)$. Alas, $L = I_K(L)$. ■

Before the next circle inversion fact, we need to recall the following Euclidean geometry fact, often called Thales' theorem.

Fact 2. *Let K be a circle and let $A, C \in K$ be such that $O \in \overline{AC}$ (that is, $\overline{AC}$ is a diameter of K). Let B be a point in the plane other than A or C. Then $\angle ABC = \frac{\pi}{2}$ if and only if $B \in K$.*

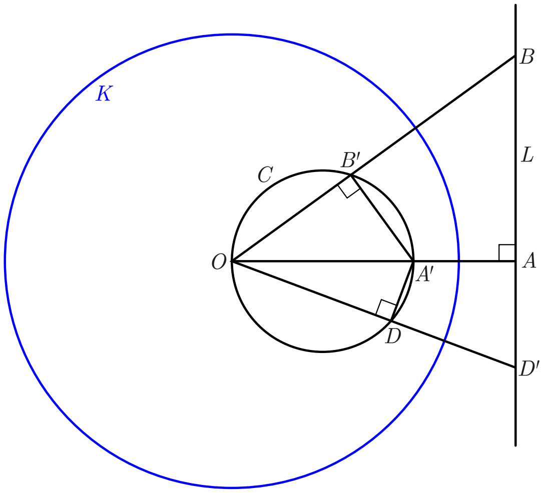
Lecture 4 - 12/05

Proposition 3. *Let K be a circle with center O and radius r.*

1. For a line L not containing O, $I_K(L)$ is a circle containing O.

2. If C is a circle containing O, $I_K(C)$ is a line not containing O.

Proof. 1. Since $\infty \in L$, $I_K(\infty) = O \in I_K(L)$. So, we just need to show that $I_K(L)$ is a circle. Let $A \in L$ be such that OA is perpendicular to L, and let $A' = I_K(A)$. I claim that $I_K(L) = C$ where C is the circle with diameter $\overline{OA'}$.



Let B be an arbitrary point on L (other than ∞), and let $B' = I_K(B)$. Since B and B' lie on the ray $\overrightarrow{OB}$ and A and A' lie on the ray $\overrightarrow{OA}$, we have $\angle A'OB' = \angle AOB$. Furthermore, $OB \cdot OB' = OA \cdot OA'$ and so $\frac{OA}{OB} = \frac{OB'}{OA'}$. Therefore the triangles $\triangle AOB$ and $\triangle B'OA'$ are similar. It follows that $\angle OB'A' = \angle OAB = \frac{\pi}{2}$. Alas, by Fact 2, $B' \in C$. Therefore, $I_K(L) \subset C$.

For the other inclusion, let $D \in C$. Let D' be the intersection of the ray $\overrightarrow{OD}$ with L. Since $D \in C$, $\angle ODA' = \frac{\pi}{2}$ by Fact 2 and $\triangle ODA'$ and $\triangle OAD'$ are similar. The similarity implies $\frac{OA'}{OD} = \frac{OD'}{OA}$ and therefore

$$OD \cdot OD' = OA \cdot OA' = r^2.$$

It follows from the definition of I_K that $I_K(D) = D'$. Applying I_K again gives $I_K(D') = I_K^2(D) = D$. So, D is in the image of L under I_K and we can conclude that $I_K(L) = C$.

2. This part follows from the first part and the fact that $I_K^2(P) = P$ for all points P in $\mathbb{R}^2 \cup \{\infty\}$.

Let $A \in C$ be such that OA is a diameter of C (that is, the line segment $\overline{OA}$ contains the center of C). Let $A' = I_K(A)$ and let L be a line perpendicular to $\overline{OA'}$ that contains A'. The claim is that $I_K(C) = L$. However, from the proof of Part 1 we know $I_K(L) = C$. Therefore $L = I_K^2(L) = I_K(C)$, completing the proof. ■

The proof of the previous proposition not only tells us that the statement is true, it also gives a construction of the circle and line! This is called a constructive proof in mathematics.

Lecture 5 - 14/05

Exercise. Let K be a circle in $\mathbb{R}^2$ with center $(0,0)$ and radius r. Let C be a circle containing $(0,0)$ of radius $s > \frac{r}{2}$ and center on the positive x-axis.

1. Prove that the circles K and C intersect.
2. Let A be an intersection point of K and C. Prove that the x-coordinate of A is equal to $\frac{r^2}{2s}$.

So, we know what happens to all lines, and circles that pass through the center of the inversion circle. Let's see what happens to circles that don't pass through the center of the inversion circle. This time we will take a more algebraic approach, and do things in the complex plane.

Proposition 4. *Let K be a circle with center O, and let C be a circle such that $O \notin C$. Then there is a circle C' not containing O such that $I_K(C) = C'$.*

Sketch of proof: By scaling and translating, we can prove the statement of the circle inversion $f : \mathbb{C} \cup \{\infty\} \to \mathbb{C} \cup \{\infty\}$ given by $f(z) = \overline{z}^{-1}$. That is, the circle inversion about the unit circle in $\mathbb{C}$.

Let C be an arbitrary circle in $\mathbb{C}$ that does not pass through $0 \in \mathbb{C}$ (the center of the inversion circle). So C has center $a \in \mathbb{C}$ and radius r where $|a|^2 = a\overline{a} \neq r^2$. The circle C is the set of complex numbers given by

$$C = \{z \in \mathbb{C} : |z - a| = r\} = \{z \in \mathbb{C} : (z - a)(\overline{z} - \overline{a}) = r^2\}.$$

Let $w = f(z) = \overline{z}^{-1}$. Then, as a consequence of a fun (for some definition of fun) exercise, we can show that $z \in C$ if and only if

$$\left(w - \frac{\overline{a}}{a\overline{a} - r^2}\right)\left(\overline{w} - \frac{a}{\overline{a}a - r^2}\right) = \left(\frac{r}{a\overline{a} - r^2}\right)^2.$$

We can then conclude that the image of C under f is the circle of radius $\frac{r}{|a\overline{a}-r^2|}$ centered at $\frac{\overline{a}}{a\overline{a}-r^2}$. ■

Exercise. Let r be a positive real number, and let a be a complex number such that $|a| \neq r$. Prove that $z \in \mathbb{C}$ satisfies

$$|z - a| = r$$

if and only if $w = \overline{z}^{-1}$ satisfies

$$\left|w - \frac{\overline{a}}{a\overline{a} - r^2}\right| = \frac{r}{|a\overline{a} - r^2|}.$$

Exercise. Let K be some circle in the plane. Which circles C, not containing the center of K, have the property that I_K maps the center of C to the center of $I_K(C)$?

The next thing we want to investigate is how circle inversion acts with respect to angles. It will turn out that circle inversions preserve angles.

Definition. The angle between two curves at a point P of intersection is the angle between the tangent lines to the curves at P. As a converntion, we take the angle to lie in the interval $[0, \frac{\pi}{2}]$.

This definition only makes sense if the curves are differentiable, which means they have well-defined tangent lines.

Exercise. Compute the angle between the curves $y = x^2$ and $(x - 1)^2 + y^2 = 1$ at the point $(1, 1)$ in $\mathbb{R}^2$.

Lecture 6 - 16/05

In order to prove that angles are preserved under circle inversions, we will need to compute angles between curves by approximating them by smaller and smaller triangles. Here is a fact that we will use without proof:

Fact 5. *Suppose C_1 and C_2 are curves intersecting at P with well-defined tangent lines at P. Let $A_1, A_2, A_3, \ldots$ be a sequence of points on C_1 that approach P. Let $B_1, B_2, B_3, \ldots$ be a sequence of points on C_2 that approach P. Then the angle between C_1 and C_2 at P is equal to $\lim_{n\to\infty} \angle A_n P B_n$.*

Also recall the following result which was proved in the proof of Proposition 3.

Lemma 6. *Let O be the center of a circle K, and let M, N be points other than O in the plane. Let $M' = I_K(M)$ and $N' = I_K(N)$. Then the traingles $\triangle OMN$ and $\triangle ON'M'$ are similar.*

Proof. This is an exercise, and is hidden in the proof of Poposition 3. ■

We are now ready to prove the following two results.

Theorem 7. *The angle between curves is preserved under circle inversion.*

Proof. Let α and β be curves intersecting at a point P ($P \neq O$, the center of the circle K of inversion). Let $P' = I_K(P)$. Then the curves $I_K(\alpha)$ and $I_K(\beta)$ intersect at P'. We want to show the angle between α and β at P is equal to the angle between $I_K(\alpha)$ and $I_K(\beta)$ at P'.

Consider a ray from O not passing through P that intersects α at N and β at M. Let $I_K(N) = N'$ and $I_K(M) = M'$. Note that N' is on $I_K(\alpha)$ and M' is on $I_K(\beta)$. Now, triangles $\triangle OMP$ and $\triangle OP'M'$ are similar, as are $\triangle ONP$ and $\triangle OP'N'$.

WIthout loss of generality, suppose $\angle OPM > \angle OPN$. Then $\angle NPM = \angle OPM - \angle OPN$. The angle $\angle OM'P'$ is an exterior angle for the triangle $\triangle M'P'N'$, so $\angle OM'P' = \angle ON'P' + \angle M'P'N'$. Rearranging and using similarity of the triangles above gives

$$\angle M'P'N' = \angle OM'P' - \angle ON'P' = \angle OPM - \angle OPN = \angle NPM.$$

Great! Now we can choose a sequence of rays starting from O that approach the ray $\overrightarrow{OP}$. The intersection points of these rays with the curves $\alpha, \beta, I_K(\alpha)$, and $I_K(\beta)$ form sequences $\{N_1, N_2, \ldots\}$ and $\{M_1, M_2, \ldots\}$ that approach P, and, $\{N'_1, N'_2, \ldots\}$ and $\{M'_1, M'_2, \ldots\}$ that approach P'. We showed above that $\angle M'_n P N'_n = \angle N_n P M_n$ for all n. Therefore, the angle between α and β at P is equal to the angle of $I_K(\alpha)$ and $I_K(\beta)$ at P'. ■

Note that although the angles are equal, the direction the angle is measured in is reversed. This is to be expected since a circle inversion is a reflection of sorts!

As a consequence of the fact that the angle between curves is preserved by circle inversion, we have a characterisation of which circles are preserved by circle inversion. We already know that lines passing through the center of inversion are preserved.

Before we do that, we need the following result from Euclidean geometry.

Lemma 8. *Let K be a circle and $P, Q \in K$ be such that $\overline{PQ}$ does not contain the center O of K. There is a unique circle C such that $P, Q \in C$ and C and K are orthogonal.*

Proof. Let T_P and T_Q be the tangent lines at P and Q to the circle C, and let N be the intersection point of T_P and T_Q (they must intersect since $\overline{PQ}$ is not a diameter of K, and thus T_P and T_Q are not parallel). Then $\overline{OP}$ is orthogonal to $\overline{PN}$ and $\overline{OQ}$ is orthogonal to $\overline{QN}$. By the Pythagorean theorem, $PN^2 = ON^2 - OP^2 = ON^2 - OQ^2 = QN^2$. Therefore N is the center of a circle C such that $P \in C$ and $Q \in C$. By the construction C is orthogonal to K.

Uniqueness is left as an exercise. ■

Theorem 9. *Let K and C be distinct circles. Then $I_K(C) = C$ if and only if C is orthogonal to K.*

Proof. Suppose $I_K(C) = C$. If C is contained in the intrior of K, then $I_K(C)$ is in the exterior of K, and vice versa. Therefore it must be the case that C and K intersect. Let P be a point of intersection, and since $P \in K$, we know that $I_K(P) = P$. Let L_K be the tangent to K at P, and L_C the tangent to C at P. The line L_K forms two complementary angles with L_C, call their measures θ and ϕ. Since angles are preserved by inversion, $\theta = \phi$. Since they are complementary, we must have $\theta = \phi = \frac{\pi}{2}$.

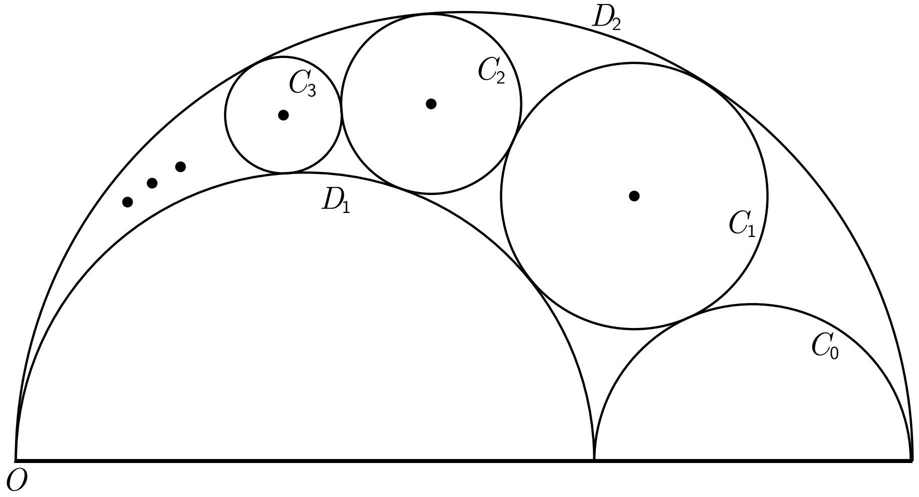
Conversely, suppose C is orthogonal to K. Then $I_K(C)$ is orthogonal to K since inversion preserves angles. Furthermore, the intersection points lie on K, so $I_K(C)$ is a circle orthogonal to K, intersecting K at the same intersection points as C. It follows that $I_K(C) = C$ by Lemma 8. ■

Lecture 7 - 21/05

Let's put all of the properties of circle inversion that we've seen so far to good use.

The arbelos

The word arbelos comes from a Greek word meaning "shoemaker's knife." Here's how one constructs an arbelos. Take three colinear points O, A and B, and let $\overline{OA}$, $\overline{OB}$, and $\overline{AB}$ be diameters of semi-circles D_1, D_2, and C_0 respectively.



This object is what's called an arbelos. However, since we're not making shoes, we're going to add some more circles, just for fun. For every positive integer n, let C_n be a circle tangent to C_{n-1}, D_1, and D_2. At this point it's not even clear that such a circle exists, but for the moment, take my word for it.

Let h_n be the height of the center of the circle C_n above the diameters. So, for example, $h_0 = 0$. Here's an amazing fact.

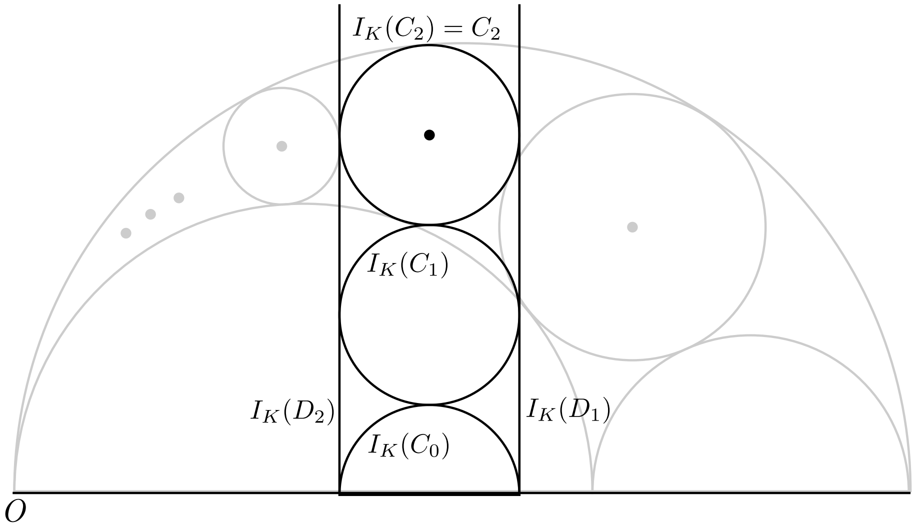
Proposition 10. *For all positive integers n, $h_n = nd_n$ where d_n is the diameter of the circle C_n.*

Let's prove this, using the full force of the theory we've built up so far. In order to do that, we first need the following fact from Euclidean geometry, which you should prove.

Fact 11. *Given a point O outside a circle C, there is a unique circle K with center O so that C and K are orthogonal.*

We can now prove the proposition by cleverly choosing a circle to invert about.

Proof. Let's prove the statement for $n = 2$, the rest of the cases follow similarly. Let K be a circle with center O that is orthogonal to C_2. Now invert the entire arbelos (and all the circles) about K. We get the following image, overlaid over the original image.



The line L that contains the diameter of D_1 passes through O, and so is preserved under I_K. The semi-circles D_1 and D_2 pass through O, and so map to lines. Since D_1 and D_2 are orthogonal to the diameters, the lines $I_K(D_1)$ and $I_K(D_2)$ are orthogonal to L, and are therefore parallel. The circle C_2 is orthogonal to K and is therefore sent to itself. All points of tangency correspond to curves intersecting with angle 0, and are therefore sent to points of tangency.

Putting all of this together we see that all the circles $I_K(C_n)$ have the same diameter, and therefore the height of the center of C_2 is twice the diameter of C_2. ■

Wild. Notice that this proof also provides us a way to prove that a cofiguration of circles $C_1, C_2, \ldots$ as we have described actually exists.

Exercise. Consider the arbelos with all the circles C_n as in the image above. Let P_n be the point of tangence between C_n and C_{n-1}. Prove that all the points P_n lie on a circle.

Lecture 8 - 23/05

3 Hyperbolic geometry

We are now ready to finally do some non-Euclidean geometry. We are going to create a "geometry" (whatever that means) which shows that you can have the first four postulates of Euclid's hold, but the fifth one be completely false. This geometry will be called hyperbolic geometry, and we begin by creating a model of it (not that different to how $\mathbb{R}^2$ with the dot product is a model of Euclidean geometry).

The model we will begin with is something called the **Poincaré disk model** of hyperbolic geometry. To set this up, first we establish some notation.

$$\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\} \quad \text{and} \quad \partial\mathbb{D} = \{z \in \mathbb{C} : |z| = 1\}.$$

So $\mathbb{D}$ is the interior of the unit circle in $\mathbb{C}$, and $\partial\mathbb{D}$ is the unit circle. Notice that $\partial\mathbb{D}$ is **not** a subset of $\mathbb{D}$. If we like, we can instead think of $\mathbb{D}$ and $\partial\mathbb{D}$ as living in $\mathbb{R}^2$, in which case

$$\mathbb{D} = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\} \quad \text{and} \quad \partial\mathbb{D} = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}.$$

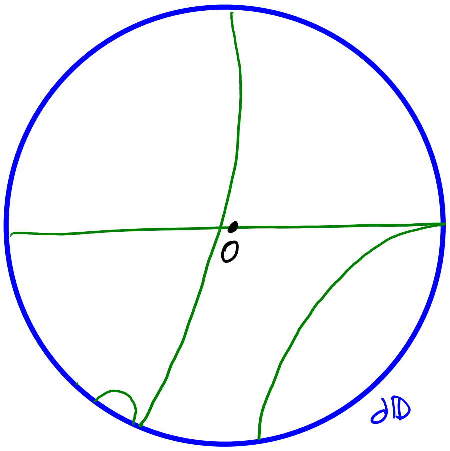
In some circumstances it's better to have $\mathbb{D}$ living in $\mathbb{C}$, and sometimes it's better to be in $\mathbb{R}^2$. We can choose whatever is most convenient for us at the time.

Now, $\mathbb{D}$ is going to consist of the points of our geometry. We first define which subsets of $\mathbb{D}$ are straight lines.

In what follows, a **Euclidean line** or a **Euclidean circle** is simply a regular line or circle in $\mathbb{R}^2$ or $\mathbb{C}$.

Definition. A *d*-**line** is part of a Euclidean circle or line that is orthogonal to $\partial\mathbb{D}$, and is in $\mathbb{D}$.

The *d* in *d*-line is for disk, and *d*-lines will play the role of straight lines in our geometry. Here are some *d*-lines in $\mathbb{D}$:



Exercise. Prove that a *d*-line that is part of a line must contain the origin.

Exercise. Prove that the center of a (Euclidean) circle othogonal to $\partial\mathbb{D}$ must lie outside $\partial\mathbb{D}$.

Just like in Euclidean geometry, we have a notion of parallel lines. In Euclidean geometry, lines are parallel if they do not intersect. In hyperbolic geometry, the situation is a little more subtle, and two lines can not intersect in two different ways.

Definition. Two *d*-lines that do not intersect are

- **parallel** if the two Euclidean circles (or lines) defining them intersect at $\partial\mathbb{D}$,
- **ultra-parallel** otherwise.

Exercise. Sketch out $\mathbb{D}$, $\partial\mathbb{D}$, and three lines l_1, l_2, l_3 such that

- l_1 and l_2 are parallel,
- l_2 and l_3 are ultra-parallel, and
- l_1 and l_3 intersect in $\mathbb{D}$.

3.1 Hyperbolic transformations

We are now ready to define hyperbolic reflections. In $\mathbb{R}^2$, a reflection across a straight line is an **isometry**, meaning it preserves the geometry (length, angle etc). Although we don't have a notion of length or angle on $\mathbb{D}$ yet, we will first define our set of reflections, and then create the geometry on $\mathbb{D}$ so that the reflections are isometries! Seems like cheating, I know.

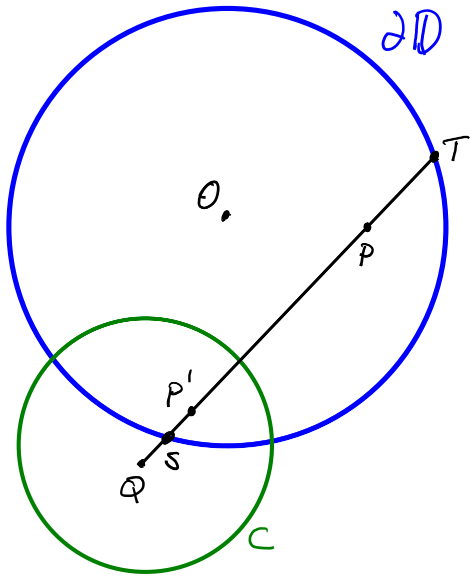
Lecture 9 - 26/05

The straight lines in $\mathbb{D}$ are d-lines, which are parts of Euclidean circles or lines. We know how to "reflect" across Euclidean circles and lines - we can take a circle inversion about a circle, or an honest-to-goodness reflection across a line.

First, we need to make sure that $\mathbb{D}$ is sent to itself under our proposed reflections.

Theorem 12. *Let l be a d-line that is part of a circle C. Then $I_C(\partial D) = \partial D$ and $I_C(\mathbb{D}) = \mathbb{D}$.*

Proof. Since C and $\partial\mathbb{D}$ are orthogonal, we have $I_C(\partial D) = \partial D$.



Let Q be the center of C, and O the origin (the center of $\partial\mathbb{D}$). Let $P \in \mathbb{D}$, and let S and T be the intersection of $\overrightarrow{QP}$ with $\partial\mathbb{D}$. We can choose S and T so that $QS < QP < QT$. Let $P' = I_C(P)$.

Since $I_C(\partial\mathbb{D}) = \partial\mathbb{D}$, $I_C(S) = T$. Therefore $QS \cdot QT = QP \cdot QP'$ by the definition of circle inversion. Therefore

$$\frac{QS}{QP} = \frac{QP'}{QT}$$

implying $QP' < QT$. Also,

$$\frac{QS}{QP'} = \frac{QP}{QT}$$

implying $QP' > QS$. Therefore $P' \in \mathbb{D}$ and $I_C(\mathbb{D}) \subset \mathbb{D}$.

Conversely, suppose $P \in \mathbb{D}$. We just showed that $I_C(P) \in \mathbb{D}$. Then $I_C(I_C(P)) = P$ and so $P \in I_C(\mathbb{D})$. Therefore, $I_C(\mathbb{D}) = \mathbb{D}$. ∎

Exercise. Let l be a d-line that is defined by a Euclidean line L. If R is a reflection across L, prove that $R(\partial\mathbb{D}) = \partial\mathbb{D}$ and $R(\mathbb{D}) = \mathbb{D}$.

We now know that reflection about d-lines is truly a reflection of $\mathbb{D}$ (and $\partial\mathbb{D}$ for that matter).

Definition. A **hyperbolic reflection** in a d-line l is the restriction to $\mathbb{D}$ of the circle inversion (or reflection) about the Euclidean circle (or line) containing l.

We want the hyperbolic reflections to generate all isometries of the hyperbolic plane. One thing we do know is that circle inversions and reflections across lines both preserve angles, so if we define angles in $\mathbb{D}$ to be our usual angles, hyperbolic reflections will preserve them!

Definition. The **hyperbolic angle** between two curves in $\mathbb{D}$ intersecting at P is the Euclidean angle of the two curves in the Euclidean plane intersecting at P.

So, we know that hyperbolic reflections presrve hyperbolic angle (although they reverse the direction of the angle). More importantly, we have the following theorem.

Theorem 13. *If l is a d-line, then the image of l under a hyperbolic reflection is also a d-line.*

Proof. This is an exercise. ∎

In usual Euclidean geometry in $\mathbb{R}^2$, it turns out that you can generate any isometry (ie, transformation of $\mathbb{R}^2$ that preserves angle and distance) by composing together a bunch of reflections across lines. If you're not convinced by this, try to write a rotation, and then a translation, as a composition of two reflections.

The same thing will be true for us in hyperbolic geometry.

Definition. A composition of finitely many hyperbolic reflections is called a **hyperbolic transformation**. The set of all hyperbolic transformations is called the **hyperbolic transformation group** and is denoted $\mathcal{G}_{\mathbb{D}}$.

Exercise. Show that a rotation by θ about the origin is in $\mathcal{G}_{\mathbb{D}}$.

Lecture 10 - 28/05

3.2 Hyperbolic lines

We now shift our attention to some basic results about d-lines. There are many facts we take for granted in Euclidean geometry. For example, between any two points there exists a unique line. A point and a direction define a line. Between two points there is a perpendicular bisector.

Statements of this nature, while true in hyperbolic geometry, need careful proving. To do this, we first prove an unreasonably useful lemma that tells us we can always find a hyperbolic reflection that sends a chosen point to the origin.

Lemma 14 (The Origin Lemma)**.** *Let $A \in \mathbb{D}$, $A \neq 0$. Then there exists a hyperbolic reflection $\sigma \in \mathcal{G}_{\mathbb{D}}$ satisfying $\sigma(A) = 0$.*

Proof. Let $P = I_{\partial\mathbb{D}}(A)$, so $0A \cdot 0P = 1$. Let C be the circle centered at P that is orthogonal to $\partial\mathbb{D}$. We will show that $I_C(A) = 0$.

By Pythagoras we have $0P^2 = 1 + r^2$, where r is the radius of C. Therefore $0P^2 = 0A \cdot 0P + r^2$. Rearranging gives $P0 \cdot PA = r^2$. Since $0, P$, and A lie on the ray $\overrightarrow{A0}$, it follows that $I_C(P) = 0$. ■

The usefulness lies in the fact that we understand *d*-lines that pass through 0. So if we want to solve some problem, it often helps to first reflect so some point of interest is at 0, solve the problem there, and then reflect back.

Corollary 15. *Let $A \in \mathbb{D}$. There are infinitely many d-lines passing through A.*

Proof. Let $\sigma \in \mathcal{G}_{\mathbb{D}}$ be the hyperbolic reflection satisfying $\sigma(A) = 0$. Every line through the origin is a *d*-line containing 0, so there are infinitely many *d*-lines containing 0. Let l be one of these lines. Then $\sigma(l)$ contains $\sigma(0) = A$. It remains to show that if l_1 and l_2 are two distinct *d*-lines, then so are $\sigma(l_1)$ and $\sigma(l_2)$, which is left as an exercise.

It follows that we can now reflect our infinitely many *d*-lines through 0 to infinitely many *d*-lines through A. ■

Exercise. Complete the proof of the previous corollary by proving that if l_1 and l_2 are distinct *d*-lines, then $\sigma(l_1)$ and $\sigma(l_2)$ are distinct *d*-lines.

Let's see how the Origin Lemma can help us again!

Theorem 16. *Let $A, B \in \mathbb{D}$ be disjoint points. Then there exists a unique d-line containing A and B.*

Proof. Let $\sigma \in \mathcal{G}_{\mathbb{D}}$ be a hyperbolic reflection such that $\sigma(A) = 0$ and $\sigma(B) = B'$. There is a unique Euclidean line passing through B' and 0. Since *d*-lines containing 0 are Euclidean lines, it follows that there is a unique *d*-line passing through B' and 0. Call this line l.

Applying σ gives us that $\sigma(l)$ is a *d*-line containing $\sigma(0) = A$ and $\sigma(B') = B$.

For uniqueness, suppose l' is a *d*-line containing A and B. Then $\sigma(l')$ is a *d*-line containing 0 and B'. Since there is only one such line, it must be the case that $\sigma(l') = l$. Therefore $l' = \sigma(l)$, completing the proof. ■

Here's another result.

Theorem 17. *Let $A_1, A_2 \in \mathbb{D}$ be points on d-lines l_1 and l_2 respectively. There exists a $\tau \in \mathcal{G}_{\mathbb{D}}$ such that $\tau(A_1) = A_2$ and $\tau(l_1) = l_2$.*

Proof. Here's the idea, and the details are left as an exercise. By the origin lemma, there are hyperbolic reflections $\sigma_1, \sigma_2 \in \mathcal{G}_{\mathbb{D}}$ satisfying $\sigma_1(A_1) = 0 = \sigma_2(A_2)$. Now, $\sigma_1(l_1)$ and $\sigma_2(l_2)$ are possibly different *d*-lines passing through 0. By a previous exercise, there is some rotation about 0, call it $r \in \mathcal{G}_{\mathbb{D}}$, such that $r(\sigma_1(l_1)) = \sigma_2(l_2)$ and $r(0) = 0$. Now let's compose all of these! Let $\tau = \sigma_2 \circ r \circ \sigma_1$. Then

$$\tau(A_1) = \sigma_2 \circ r(0) = \sigma_2(0) = A_2$$

and

$$\tau(l_1) = \sigma_2 \circ r(\sigma_1(l_1)) = \sigma_2 \circ \sigma_2(l_2) = l_2$$

completing the proof. ■

Lecture 11 - 30/05

3.3 Some helpful formulas

We know from earlier in the course, that given a Euclidean circle C, and a point P outside C, there exists a unique circle D centered at P orthogonal to C. With this in mind, we can make the following definition.

Definition. Let $\alpha \in \mathbb{C}$ be such that $|\alpha| = 1$. Define the ***d*-line defined by** α to be the d-line defined by the unique circle centered at α and orthogonal to $\partial\mathbb{D}$.

Recall that the function $f : \mathbb{C} \cup \{\infty\} \to \mathbb{C} \cup \{\infty\}$ given by $f(z) = \overline{z}^{-1}$ (along with $f(\infty) = 0$ and $f(0) = \infty$) is a circle inversion across the unit circle centered at 0.

Now, if we want to find the formula for a circle inversion about a circle centered at α with radius r, we can do the following: First translate the complex plane so that α is sent to 0, then scale by $\frac{1}{r}$. Now we can perform the circle inversion about the unit circle. Once that is done, we can scale back by r, and translate back so that 0 ends up at α. That will do it!

Formally, let $T : \mathbb{C} \cup \{\infty\} \to \mathbb{C} \cup \{\infty\}$ be given by $T(z) = z - \alpha$ and $T(\infty) = \infty$. Then T is a translation that sends α to 0. Let $S : \mathbb{C} \cup \{\infty\} \to \mathbb{C} \cup \{\infty\}$ be given by $S(z) = r^{-1}z$ and $S(\infty) = \infty$. Then S scales a circle centered at 0 with radius r to the unit circle. Putting all of this together, we have

$$\begin{aligned}
T^{-1}S^{-1}fST(z) &= T^{-1}S^{-1}fS(z-\alpha) \\
&= T^{-1}S^{-1}f\left(\frac{z-\alpha}{r}\right) \\
&= T^{-1}S^{-1}\left(\frac{r}{\overline{z}-\overline{\alpha}}\right) \\
&= T^{-1}\left(\frac{r^2}{\overline{z}-\overline{\alpha}}\right) \\
&= \frac{r^2}{\overline{z}-\overline{\alpha}} + \alpha.
\end{aligned}$$

This gives us the formula for a circle inversion across a circle with center α and radius r.

Exercise. Prove that the function $f(z) = \frac{r^2}{\overline{z}-\overline{\alpha}} + \alpha$, together with the declaration that $f(\alpha) = \infty$ and $f(\infty) = \alpha$ is indeed a circle inversion about the circle with center α and radius r.

Great, now let C be the unique circle centered at α and orthogonal to $\partial\mathbb{D}$. By Pythagoras, we have that $r^2 + 1 = |\alpha|^2$. Substituting this new information into the formula above gives

$$\frac{r^2}{\overline{z}-\overline{\alpha}} + \alpha = \frac{\alpha\overline{\alpha} - 1 + \alpha\overline{z} - \alpha\overline{\alpha}}{\overline{z}-\overline{\alpha}} = \frac{\alpha\overline{z} - 1}{\overline{z}-\overline{\alpha}}.$$

We have just proved the following:

Lemma 18. *Let $\alpha \in \mathbb{C}$ be such that $|\alpha| > 1$. Let σ be the hyperbolic reflection across the d-line defined by α. Then*

$$\sigma(z) = \frac{\alpha\overline{z} - 1}{\overline{z}-\overline{\alpha}}$$

for all $z \in \mathbb{D} \cup \partial\mathbb{D}$.

Great! It's useful to have formulas. For example, we now have an algebraic proof of the origin lemma within our reach.

Exercise. Prove the Origin Lemma (Lemma 14) by using Lemma 18.

What about if the d-line is a line through the origin? We can approach a derivation of the formula much the same way. We know $B(z) = \overline{z}$ is a reflection across the horizontal axis. We also know that $r(z) = e^{i\theta}z$ is a rotation by θ counterclockwise. Thus

$$rBr^{-1}(z) = rB(e^{-i\theta}z) = r(e^{i\theta}\overline{z}) = e^{2i\theta}\overline{z}.$$

We have just derived a formula for the reflection across a line going through the origin that forms an angle of θ with the positive real axis. Such a line has equation $y = x\tan(\theta)$ or can be written as the set of points $\{a + bi \in \mathbb{C} : b = a\tan(\theta)\}$.

Lemma 19. *Let l be the d-line defined by the Euclidean line $\{a + bi \in \mathbb{C} : b = a\tan(\theta)\}$. Then the hyperbolic reflection σ across l is given by $\sigma(z) = e^{2i\theta}\overline{z}$ for all $z \in \mathbb{D} \cup \partial\mathbb{D}$.*

These formulas give us a way of investigating what happens when you compose hyperbolic reflections.

Example. Let σ_1, σ_2 be the hyperbolic reflections $\sigma_1(z) = e^{2i\theta}\overline{z}$ and $\sigma_2(z) = e^{2i\phi}\overline{z}$. Then if we compose them we get

$$\sigma_2\sigma_1(z) = \sigma_2(e^{2i\theta}\overline{z}) = e^{2i(\phi-\theta)}z.$$

This is a rotation about the origin by $2(\phi - \theta)$ counterclockwise! By cleverly choosing my ϕ and θ, I can obtain any rotation about the origin. The take-home message here is that every rotation about the origin is an element of $\mathcal{G}_{\mathbb{D}}$.

Exercise. Using the formulas, prove that for any hyperbolic reflection σ, $\sigma^2(z) = z$ for all $z \in \mathbb{D} \cup \partial\mathbb{D}$.

Let's compose two reflections across d-lines defined by circles.

Example. Let $\alpha, \beta \in \mathbb{C}$ be such that $|\alpha| > 1$ and $|\beta| > 1$. Let

$$\sigma_1(z) = \frac{\alpha\overline{z} - 1}{\overline{z} - \overline{\alpha}} \quad \text{and} \quad \sigma_2(z) = \frac{\beta\overline{z} - 1}{\overline{z} - \overline{\beta}}.$$

That is, σ_1 and σ_2 are the hyperbolic reflections about the d-lines defined by α and β respectively. Then

$$\sigma_2\sigma_1(z) = \frac{(\beta\overline{\alpha} - 1)z + (\alpha - \beta)}{(\overline{\alpha} - \overline{\beta})z + (\alpha\overline{\beta} - 1)}.$$

Exercise. Verify the above formula.

Let's do the remaining case, which is a composition of two hyperbolic reflections, one of which is defined by a Euclidean line through the origin.

Example. Let

$$\rho(z) = \frac{\alpha\overline{z} - 1}{\overline{z} - \overline{\alpha}} \quad \text{and} \quad \sigma(z) = e^{2i\theta}\overline{z}.$$

Then

$$\sigma\rho(z) = \frac{e^{2i\theta}\bar{\alpha}z - e^{2i\theta}}{z - \alpha} \quad \text{and} \quad \rho\sigma(z) = \frac{\alpha e^{-2i\theta}z - 1}{e^{-2i\theta}z - \overline{\alpha}}.$$

Of course, you should verify both of these formulas.

In each of the examples we just went through, we composed two hyperbolic reflections. In all of the cases, the bar above the z disappeared. That is, there was no conjugation. Here's how to interpret that fact geometrically.

When a single hyperbolic reflection occurs, angles are preserved, but the orientation of the angle is reversed. Compose two hyperbolic reflections together, and the orientation of angles is preserved! You can detect whether or not the orientation is preserved by whether or not the z that appears in the formula is conjugated or not!

Lecture 12 - 02/06

3.4 Möbius transformations

The formula for a hyperbolic reflection has the form

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

where $a, b, c, d \in \mathbb{C}$. In the last few examples, we showed that when we compose two hyperbolic reflections we get a function of the form

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

where $a, b, c, d \in \mathbb{C}$. It turns out that functions of the latter form (so without the conjugation) play an important role in the study of the group of hyperbolic transformations, and they are called Möbius transformations.

Definition. A **Möbius transformation** is a function $M : \mathbb{C} \cup \{\infty\} \to \mathbb{C} \cup \{\infty\}$ of the form

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

where $a, b, c, d \in \mathbb{C}$ and $ad - bc \neq 0$. If $c = 0$, define $M(\infty) = \infty)$. Otherwise we define $M\left(\frac{-d}{c}\right) = \infty$ and $M(\infty) = \frac{a}{c}$.

Exercise. What kind of function is a Möbius transformation without the condition $ad - bd \neq 0$?

Here are some fun facts about Möbius transformations, which I won't prove here. A study of Möbius transformations with proofs of the following facts would likely occur in any course in complex analysis.

Fact 20.

- *Möbius transformations send circles and lines to circles and lines.*
- *Möbius transformations preserve angles, with orientation!*
- *Given two sets of three points $\{z_1, z_2, z_3\}, \{w_1, w_2, w_3\} \subset \mathbb{C} \cup \{\infty\}$ there is a unique Möbius transformation M such that $M(z_i) = w_i$ for all i.*

Example. As an example illustrating the last fact, suppose $z_1 = i$, $z_2 = \infty$, $z_3 = 3$, $w_1 = 0$, $w_2 = 1$, and $w_3 = \infty$. Then

$$M(z) = \frac{z - i}{z - 3}$$

is the unique Möbius transformation sending z_1, z_2, z_3 to w_1, w_2, w_3.

Let's return for a moment to our formula for a hyperbolic reflection. Recall that if the d-line l is defined by the point $\alpha \in \mathbb{C}$, then the hyperbolic reflection σ across l is given by

$$\sigma(z) = \frac{\alpha\overline{z} - 1}{\overline{z} - \overline{\alpha}}.$$

If we let $B(z) = \overline{z}$ (which is the hyperbolic reflection across the reall axis). Then $\sigma = M \circ B$ where M is the Möbius transformation $M(z) = \frac{\alpha z - 1}{z - \overline{\alpha}}$. In fact, you can check that the hyperbolic reflection across a d-line passing through the origin is also of the form $M \circ B$ for some Möbius transformation.

Exercise. Let $B(z) = \overline{z}$ and let M be a Möbius transformation. Show that $B \circ M \circ B$ is also a Möbius transformation. Use this to show that a composition of an even number of hyperbolic reflections is a Möbius transformation.

Now, let's see what happens when we compose two Möbius transformations.

Let $M_1(z) = \frac{az+b}{cz+d}$ and $M_2(z) = \frac{ez+f}{gz+h}$. Then

$$M_1 \circ M_2(z) = \frac{(ae + bg)z + (af + bh)}{(ce + dg)z + (cf + dh)}.$$

I'm being a little sloppy here, and we should certainly be making sure that anything involving ∞ makes sense. You can check that it does.

If we stare at those coefficients, we may be reminded of something.

Now for something completely different. Let's multiply some matrices.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{bmatrix}.$$

Gasp! Witchcraft! Somehow, composition of matrices exactly imitates composition of Möbius transformations. Even better, the condition $ad - bc \neq 0$ is exactly saying that the determinant of the corresponding matrix is non-zero, and therefore that it's invertible!

This is a super useful computational tool. Given a Möbius transformation, we can turn it into a matrix. Given a (invertible) matrix with complex entries, we can turn it back into a Möbius transformation. Be careful though, it's not a perfect dictionary. There are many complex matrices that correspond to the same Möbius transformation.

Exercise. Find two distinct matrices that correspond to the same Möbius transformation $M(z) = \frac{z+i}{z-i}$.

Matrices are wonderful computational tools. Not only can we use matrix multiplication to keep track of composition of Möbius functions, but we also know how to find inverses of 2 × 2 matrices, and this will give us inverses of Möbius transformations!

Example. Let's find the inverse of $M(z) = \frac{z-i}{iz+2}$. Here, when we mean inverse of a function, we mean a function that composes with M to give the identity.

First, we turn it into a matrix. The corresponding matrix is

$$A = \begin{bmatrix} 1 & -i \\ i & 2 \end{bmatrix}$$

This matrix has determinant 1, but let's just leave it as $|A|$ for now. The inverse is given by

$$A^{-1} = \frac{1}{|A|} \begin{bmatrix} 2 & i \\ -i & 1 \end{bmatrix}.$$

Turning this back into a Möbius function (which I will presumptiously call M^{-1}) we get

$$M^{-1} = \frac{\frac{1}{|A|}2z + \frac{1}{|A|}i}{\frac{1}{|A|}(-i)z + \frac{1}{|A|}} = \frac{2z+i}{-iz+1}.$$

It remains for you to verify that for all $z \in \mathbb{C} \cup \{\infty\}$, $M \circ M^{-1}(z) = z$ and $M^{-1} \circ M(z) = z$.

The previous example proves that every Möbius transformation has an inverse, and is therefore a bijection (see Appendix B).

Lecture 13 - 04/06 (bijections review, see Appendix B)

Lecture 14 - 06/06

Back to the hyperbolic plane!

Let's return to the study of hyperbolic transformations, and in particular, those hyperbolic transformations that are restrictions of Möbius transformations to $\mathbb{D}$. Recall from above, that if σ_1 is the hyperbolic reflection about the d-line defined by α, and σ_2 is the hyperbolic reflection about hte d-line defined by β, then

$$\sigma_2\sigma_1(z) = \frac{(\beta\overline{\alpha} - 1)z + (\alpha - \beta)}{(\overline{\alpha} - \overline{\beta})z + (\alpha\overline{\beta} - 1)}.$$

This is a Möbius transformation, but it's not just any Möbius transformation. This one is of the form $M(z) = \frac{az+b}{\overline{b}z+\overline{a}}$.

In fact, we can show that the composition of any two hyperbolic reflections is a Möbius transformation of the form $M(z) = \frac{az+b}{\overline{b}z+\overline{a}}$.

Exercise. Let $\sigma(z) = \frac{\alpha\overline{z}-1}{\overline{z}-\overline{\alpha}}$, $\tau_\theta(z) = e^{2i\theta}\overline{z}$ and $\tau_\phi(z) = e^{2i\phi}\overline{z}$. These are all hyperbolic reflections. Show that the hyperbolic transformations $\tau_\phi\tau_\theta, \tau_\phi\sigma$, and $\sigma\tau_\phi$ are all restrictions of Möbius functions of the form

$$M(z) = \frac{az+b}{\overline{b}z+\overline{a}}$$

restricted to $\mathbb{D}$.

Great, so we know that the composition of two hyperbolic reflections is a special kind of Möbius transformation. We also know the composition of an even number of hyperbolic reflections is a Möbius transformation. Let's see what happens if we compose two Möbius transformations which are the composition of two hyperbolic reflections.

Let $M_1(z) = \frac{az+b}{\overline{b}z+\overline{a}}$ and $M_2(z) = \frac{cz+d}{\overline{d}z+\overline{c}}$. We will compose these my multiplying the corresponding matrices. We have

$$\begin{bmatrix} c & d \\ \overline{d} & \overline{c} \end{bmatrix} \begin{bmatrix} a & b \\ \overline{b} & \overline{a} \end{bmatrix} = \begin{bmatrix} ca + d\overline{b} & cd + d\overline{a} \\ \overline{d}a + \overline{c}\overline{b} & \overline{d}b + \overline{ca} \end{bmatrix}.$$

Therefore

$$M_1M_2(z) = \frac{(ca + d\overline{b})z + cd + d\overline{a}}{(\overline{d}a + \overline{c}\overline{b})z + \overline{d}b + \overline{ca}}$$

which is again of the same form! This tells us that the composition of an even number of hyperbolic reflections is a Möbius transformation of the form $M(z) = \frac{az+b}{\overline{b}z+\overline{a}}$.

We can put even more restrictions on the Möbius transformations. After all, we know such a Möbius transformation must have $M(0) \in \mathbb{D}$. since $M(0) = \frac{b}{\overline{a}}$, this puts the further restriction that $|b| < |a|$.

This is all coming together into quite a pretty picture, but it's even prettier! It turns out that if a hyperbolic transformation is a composition of an even number of hyperbolic reflections, then it's the composition of 2 hyperbolic reflections!

Theorem 21. *Let $\tau \in \mathcal{G}_{\mathbb{D}}$. The following are equivalent.*

1. *τ is the restriction to $\mathbb{D}$ of a Möbius transformation of the form $M(z) = \frac{az+b}{\overline{b}z+\overline{a}}$ where $|b| < |a|$.*
2. *τ is a composition of two hyperbolic reflections.*
3. *τ is a composition of an even number of hyperbolic reflections.*

Proof. The discussion preceding the statement of the theorem proves that 3. implies 1. Since two is an even number, it's clear that 2. implies 3. It remains to show that 1. implies 2.

Suppose $M(z) = \frac{az+b}{\overline{b}z+\overline{a}}$ with $|b| < |a|$. Assume first that $M(0) = 0$. Then $M(z) = \frac{a}{\overline{a}}z$ which is a rotation about 0 (since $\left|\frac{a}{\overline{a}}\right| = 1$). We have seen earlier (or it's an exercise now for you to show) that every rotation about 0 is a composition of two hyperbolic reflections across d-lines passing through 0.

Now assume $M(0) \neq 0$, and so that we don't get a heinous clash of notation, let $M(z) = \frac{cz+d}{\overline{d}z+\overline{c}}$. In this case, $d \neq 0$ since $M(0) \neq 0$. Let $\alpha = -(\overline{c}/\overline{d})$, which is legal since $\overline{d} \neq 0$. Furthermore, since $|d| < |c|$, we have $|\alpha| > 1$.

Let σ be the hyperbolic reflection about the d-line defined by α. Then $\sigma = N \circ B$ where $B(z) = \overline{z}$ and N is the Möbius transformation with corresponding matrix

$$\begin{bmatrix} \frac{-\overline{c}}{\overline{d}} & -1 \\ 1 & \frac{c}{d} \end{bmatrix}.$$

Then the Möbius transformation $M \circ N$ has corresponding matrix

$$\begin{bmatrix} c & d \\ \overline{d} & \overline{c} \end{bmatrix} \begin{bmatrix} \frac{-\overline{c}}{\overline{d}} & -1 \\ 1 & \frac{c}{d} \end{bmatrix} = \begin{bmatrix} d - \frac{|c|^2}{\overline{d}} & 0 \\ 0 & -\left(\overline{d} - \frac{|c|^2}{d}\right). \end{bmatrix}$$

Therefore

$$M \circ \sigma(z) = M \circ N \circ B(z) = \frac{\left(d - \frac{|c|^2}{\overline{d}}\right)\overline{z}}{-\left(\overline{d} - \frac{|c|^2}{d}\right)} = -\frac{\gamma}{\overline{\gamma}}\overline{z} = \omega\overline{z}$$

where $|\omega| = 1$. Therefore $M \circ \sigma$ is a hyperbolic reflection. Then $M = (M \circ \sigma) \circ \sigma$, and so M is a composition of two hyperbolic reflections. ■

Lecture 15 - 09/06

3.5 Orientation-preserving hyperbolic transformations

We now have a much better idea of what hyperbolic transformations look like. The ones that are compositions of an even number of hyperbolic reflections (which we now know are compositions of just two hyperbolic reflections) preserve the orientation of angles, whereas the others reverse the orientation of the angles.

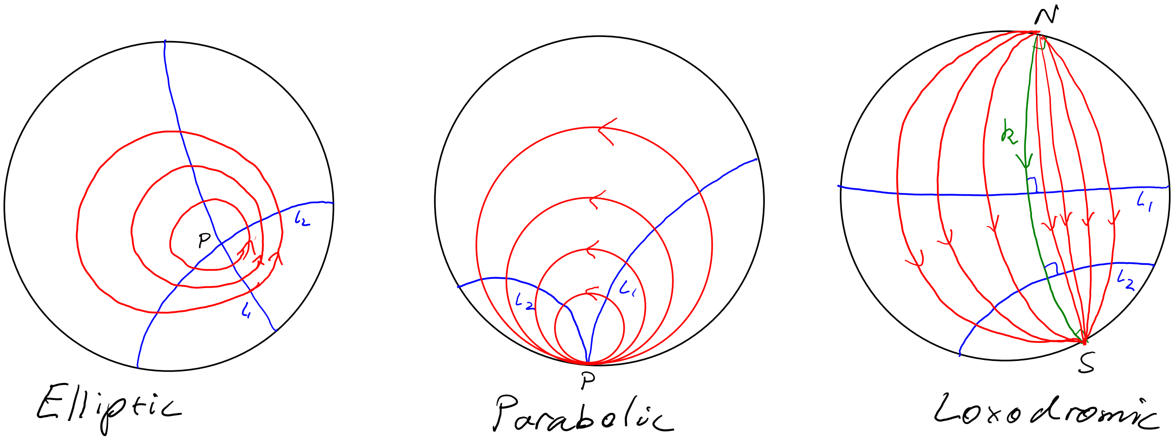


Figure 1: The red circles (and lines) are preserved by τ, and points move along each red circle.

Definition. Let $\tau \in \mathcal{G}_{\mathbb{D}}$. If τ preserves the orientation of angles, then we say τ is an **orientation-preserving** hyperbolic transformation. If τ reverses the orientation of angles, we say τ is an **orientation-reversing** hyperbolic transformation.

We now have the following corollary of Theorem 21.

Corollary 22. *Every orientation-preserving hyperbolic transformation can be written as a composition of two hyperbolic reflections. Every orientation-reversing hyperbolic transformation can be written as a composition of three hyperbolic reflections.*

Note that the identity (the transformation given by $\tau(z) = z$ for all $z \in \mathbb{D}$) is an orientation-preserving transformation, and can be written as $\sigma \circ \sigma$ for any hyperbolic reflection σ. Also, any hyperbolic reflection σ can be written as $\sigma \circ \tau \circ \tau$ for any hyperbolic reflection τ.

Theorem 21 gives us a very interesting picture of what orientation-preserving hyperbolic transformations can look like. In the analogous story for Euclidean geometry, orientation preserving Euclidean transformations are either translations or rotations. Booooriiiiiing.

In the hyperbolic case, we know that every orientation-preserving hyperbolic transformation is a composition of two hyperbolic reflections. So we can analyse how the transformation behaves by how the two d-lines interact with each other.

Let l_1 and l_2 be d-lines, and let $\tau = \sigma_2\sigma_1$ be the corresponding orientation-preserving transformation.

Elliptic transformations

If l_1 and l_2 intersect in $\mathbb{D}$, then the composition of the hyperbolic reflections is called an **elliptic** transformation, or a **hyperbolic rotation**. If P is the intersection point of l_1 and l_2, then P is the only fixed point in $\mathbb{D} \cup \partial\mathbb{D}$. More precisely, if $Q \in \mathbb{D} \cup \partial\mathbb{D}$ is such that $\tau(Q) = Q$, then $Q = P$.

Of course, technically τ is only defined on $\mathbb{D}$, but we can see what happens to $\partial\mathbb{D}$ by considering τ as a composition of the two circle inversions defining σ_1 and σ_2.

A good example to have in mind here is a rotation about the origin. We've seen several times that this is a reflection across two d-lines that contain 0, and the only fixed point is 0. For all points

$Q \in \mathbb{D} \cup \partial\mathbb{D}$ other than 0, they move around 0, staying on the unique Euclidean circle centered at 0 that passes through Q.

In general, if P is the unique fixed point of the elliptic transformation τ, each point other than P rotates around P while remaining in a hyperbolic circle (we'll see what that mean later on) centered at P.

Parabolic transformations

If l_1 and l_2 are parallel, τ is what is called a **parabolic** transformaion, or a **hyperbolic limit rotation**. There are no points $P \in \mathbb{D}$ so that $\tau(P) = P$.

However, there is a fixed point in $\partial\mathbb{D}$. Let C_1 and C_2 be the Euclidean circles (or lines) defining l_1 and l_2. Since l_1 and l_2 are parallel, C_1 and C_2 intersect at a unique point $P \in \partial\mathbb{D}$. The point P is the only point in $\mathbb{D} \cup \partial\mathbb{D}$ so that $\tau(P) = P$.

A parabolic transformation τ acts very interestingly on $\mathbb{D}$. Let H be a Euclidean circle on the interior of $\partial\mathbb{D}$, that is tangent to $\partial\mathbb{D}$ and P. Such a subset of $\mathbb{D}$ is called a **horocycle**, and has the property that $\tau(H) = H$.

The transformation τ moves points along horocycles! In fact, we can say something a little more precise. For every point $Q \in \mathbb{D} \cup \partial\mathbb{D}$, $\lim_{n\to\infty} \tau^n(Q) = P$ and $\lim_{n\to\infty} \tau^{-n}(Q) = P$ (where limits are taken in Euclidean space).

Loxodromic transformations

if l_1 and l_2 are ultra-parallel, then τ is a **loxodromic** transformation, or a **hyperbolic translation**. Again, there are no points in $\mathbb{D}$ that are fixed by τ. However, similar to parabolic transformations, there are points on $\partial\mathbb{D}$ that are fixed.

We will see later on in the course that if two d-lines are ultra-parallel, there is a unique d-line that is orthogonal to both. Let k be the unique d-line orthogonal to both l_1 and l_2. Let N and S (for North and South) be the points of intersection of the circle (or line) defining k and $\partial\mathbb{D}$. It turns out that N and S are the only points in $\mathbb{D} \cup \partial\mathbb{D}$ such that $\tau(P) = P$.

The rest of the points move from N to S, guided in some sense by k. Points that lie on k stay on k and move from N to S. In fact, for every point $P \in \mathbb{D} \cup \partial\mathbb{D}$ other than N or S, $\lim_{n\to\infty} \tau^n(P) = S$ and $\lim_{n\to\infty} \tau^{-n}(P) = N$.

Lecture 16 - 11/06

3.6 Canonical form

We have seen that there is some ambiguity in how we write the formulas for hyperbolic transformations. For example,

$$\sigma(z) = \frac{2z+1}{z+2} = \frac{-2z-1}{-z-2} = \frac{2iz+i}{iz+2i}$$

and so on. In fact, there are many ways to write a particular hyperbolic transformation.

Let's play around with an arbitrary orientation-preserving hyperbolic transformation for a bit. We have

$$M(z) = \frac{az+b}{\bar{b}z+\bar{a}} = \frac{\frac{a}{\bar{a}}z + \frac{b}{\bar{a}}}{1+\frac{\bar{b}}{\bar{a}}z} = \left(\frac{a}{\bar{a}}\right)\frac{z - \frac{-b}{a}}{1 - \frac{-\bar{b}}{\bar{a}}z}.$$

Cool. This is fun. Note that $\left|\frac{a}{\bar{a}}\right| = 1$, so we can write $\frac{a}{\bar{a}}$ as $e^{i\theta}$ for some real number θ. Also notice that since $|b| < |a|$, $m = \frac{-b}{a} \in \mathbb{D}$. We have proved the following.

Theorem 23. *Every orientation-preserving hyperbolic transformation is of the form*

$$M(z) = e^{i\theta}\frac{z-m}{1-\overline{m}z}$$

where $\theta \in \mathbb{R}$ and $m \in \mathbb{D}$.

This form of an orientation-preserving hyperbolic transformation is called **canonical form**. There are two nice things about canonical form. First, it is unique. Well, pretty much:

Exercise. Suppose that for all $z \in \mathbb{D}$, $e^{i\theta}\dfrac{z-m}{1-\overline{m}z} = e^{i\phi}\dfrac{z-n}{1-\overline{n}z}$, where $n, m \in \mathbb{D}$. Prove that $n = m$ and $\theta - \phi = 2\pi k$ for some $k \in \mathbb{Z}$.

The second nice thing is how it behaves when you feed it m. In fact, canonical form always maps m to 0! This can be checked by plugging in m and watching the magic happen. In fact, every orientation-preserving hyperbolic transformation that maps m to 0 is of the form above.

Exercise. Let $M(z)$ be an orientation-preserving hyperbolic transformation so that $M(m) = 0$. Prove that there is a real number θ so that for all $z \in \mathbb{D}$, $M(z) = e^{i\theta}\frac{z-m}{1-\overline{m}z}$.

Geometrically, this is telling us that every orientation-preserving hyperbolic transformation that maps m to 0 is just obtained by one of them followed by some rotation about the origin. A believable fact!

Now, what about orientation-reversing hyperbolic transformations? Well, we start with the following exercise.

Exercise. Let $\sigma \in \mathcal{G}_{\mathbb{D}}$ be an orientation-reversing hyperbolic transformation. Then $\sigma = MB$, where M is an orientation-preserving hyperbolic transformation and $B(z) = \overline{z}$.

As a consequence of the exercise, we have the following corollary of the previous theorem.

Corollary 24. *Every orientation-reversing hyperbolic transformation can be written in the form*

$$\sigma(z) = e^{i\theta}\frac{\overline{z}-m}{1-\overline{mz}}$$

for some $\theta \in \mathbb{R}$ and $m \in \mathbb{D}$.

Note that in the corollary, $\sigma(\overline{m}) = 0$.

3.7 Hyperbolic distance

After much avoiding of actual distance, we will finally bite the bullet and define a hyperbolic distance. Here are some properties we would like a distance function to have. It should be a function

$$d : \mathbb{D} \times \mathbb{D} \to \mathbb{R}_{\geq 0}$$

(that is, something that eats two elements of $\mathbb{D}$ and spits out a non-negative real number) that satisfies the following properties for all $z, w, v \in \mathbb{D}$.

1. $d(z, w) = 0$ if and only if $z = w$.
2. $d(z, w) = d(w, z)$.
3. $d(z, w) + d(w, v) \geq d(z, v)$ (*Triangle inequality*).

4. $d(z,w) + d(w,v) = d(z,v)$ if and only if z, w, v are on a d-line with w between z and v.
5. $d(z,w) = d(\tau(z), \tau(w))$ for all $\tau \in \mathcal{G}_{\mathbb{D}}$.

We will now define such a distance function. It will be a while until we prove it satisfies all these properties, but for now we can take it for granted that it does!

Definition. Let $z \in \mathbb{D}$. The **hyperbolic distance** from 0 to z is

$$\begin{aligned} d(0,z) &= \operatorname{arctanh}(|z|) \\ &= \frac{1}{2}\ln\left(\frac{1+|z|}{1-|z|}\right). \end{aligned}$$

Lecture 17 - 13/06

Now, using the fact (which we're just taking on faith for now) that $d(z,w) = d(\tau(z), \tau(w))$ for all hyperbolic transformations τ, we can derive a formula for the hyperbolic distance between any two points.

Let $\tau(z) = \frac{z-u}{1-\overline{u}z}$. This is a hyperbolic transformation that maps u to 0. Therefore we have for all $u, v \in \mathbb{D}$,

$$d(u,v) = d\left(0, \frac{v-u}{1-\overline{u}v}\right) = \operatorname{arctanh}\left(\left|\frac{v-u}{1-\overline{u}v}\right|\right).$$

So, for example,

$$d\left(\frac{1}{2}, \frac{1}{3}i\right) = \operatorname{arctanh}\left(\sqrt{\frac{13}{37}}\right) \approx 0.6891\ldots$$

Before we go on, let's quickly review hyperbolic trig functions.

3.8 Hyperbolic trigonometric functions

We have

$$\begin{aligned} \sinh(x) &= \frac{1}{2}(e^x - e^{-x}) \\ \cosh(x) &= \frac{1}{2}(e^x + e^{-x}) \\ \tanh(x) &= \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \end{aligned}$$

These functions satisfy the identity $\cosh^2(x) - \sinh^2(x) = 1$.

We are mainly interested in tanh and its inverse, arctanh. Recall that this means $\tanh(\operatorname{arctanh}(x)) = x$ for all $x \in (-1,1)$, and $\operatorname{arctanh}(\tanh(x)) = x$ for all $x \in \mathbb{R}$.

The important thing to note is that the range of $\tanh(x)$ is $(-1,1)$, and so the domain of $\operatorname{arctanh}(x)$ is $(-1,1)$. Even better, as x gets closer and closer to 1, $\operatorname{arctanh}(x)$ shoots off to $+\infty$.

What this tells us is that points that are near the boundary $\partial\mathbb{D}$, are really far away!

3.9 Hyperbolic circles

Now that we have distance, we can talk about circles. After all, a circle is just the set of all points that are the same distance away from some given point.

Definition. The **hyperbolic circle** with radius r and center c is the set $\{z \in \mathbb{D} : d(c,z) = r\}$.

What does that look like to us mortal Euclidean creatures? Well, as always, let's first look at the case when the center is 0. In that case the circle looks like

$$\{z \in \mathbb{D} : d(0,z) = r\} = \{z \in \mathbb{D} : \operatorname{arctanh}(|z|) = r\} = \{z \in \mathbb{D} : |z| = \tanh(r)\}.$$

So the hyperbolic circle centered at 0 with radius r is the Euclidean circle centered at 0 with Euclidean radius $\tanh(r)$! Intriguing.

So, the natural question is now, what does the hyperbolic circle centered at some other point in $\mathbb{D}$ look like. Well, suppose that center is $c \in \mathbb{D}$. We know (by the Origin Lemma for example) that there is some hyperbolic transformation τ so that $\tau(0) = c$. Since all hyperbolic transformations preserve distance, we can take the circle centered at 0 with hyperbolic radius r and apply τ to it!

More precisely, let $C = \{z \in \mathbb{D} : d(0,z) = r\}$. Then

$$\tau(C) = \{\tau(z) \in \mathbb{D} : d(\tau(0), \tau(z)) = r\} = \{\sigma(z) \in \mathbb{D} : d(c, \tau(z)) = r\}$$

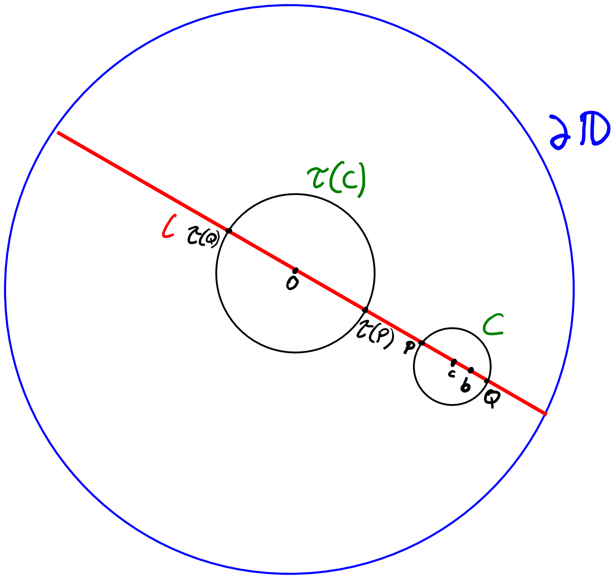
which is exactly the hyperbolic circle centered at c with radius r. Since hyperbolic transformations are just compositions of circle inversions, we know that $\tau(C)$ must be a Euclidean circle (can you see why it can't be a Euclidean line?). In fact, we have the following theorem.

Lecture 18 - 16/06

Theorem 25. *Every hyperbolic circle is a Euclidean circle in $\mathbb{D}$. Every Euclidean circle in $\mathbb{D}$ is a hyperbolic circle.*

Proof. The discussion above shows that every hyperbolic circle is a Euclidean circle in $\mathbb{D}$. Let's prove that every Euclidean circle in $\mathbb{D}$ is a hyperbolic circle.

Let C be a Euclidean circle in $\mathbb{D}$ with center c. Let P and Q be the intersection of the ray $\overleftrightarrow{0c}$ with C. Note that $\overleftrightarrow{Oc} \cap \mathbb{D}$ is the unique d-line passing through any two of the points $0, c, P, Q$. Call this d-line l.



Let b be the hyperbolic midpoint (defined in Assignment 2) of P and Q. Since l is a d-line containing P and Q, b is on l. Let τ be the hyperbolic reflection so that $\tau(b) = 0$. Since $\tau(0) = b$, it follows that $\tau(l) = l$.

Let σ_l be the hyperbolic reflection across l. Since the Euclidean center of C lies on l, C and l are orthogonal. Therefore $\sigma_l(C) = C$. By a result from Assignment 2, $\tau\sigma_l\tau^{-1} = \sigma_{\tau(l)}$. However, $\tau(l) = l$ so $\tau\sigma_l = \sigma_l\tau$. Therefore $\tau(C) = \tau\sigma_l(C) = \sigma_l\tau(C)$. We have that the Euclidean circle $\tau(C)$ is fixed by the hyperbolic reflection σ_l. Since σ_l is the (restriction to $\mathbb{D}$ of a) Euclidean reflection across a Euclidean line, it follows that the Euclidean center of $\tau(C)$ is on l.

Now, $\tau(P)$ and $\tau(Q)$ are on $\tau(C)$, and $d(\tau(b), \tau(P)) = d(\tau(b), \tau(Q))$ so $d(0, \tau(P)) = d(0, \tau(Q))$. Therefore $|\tau(P)| = |\tau(Q)|$. Since l passes through $\tau(P)$ and $\tau(Q)$ and the Euclidean center of $\tau(C)$, it must be the case that the Euclidean center of $\tau(C)$ is the Euclidean midpoint of $\tau(P)$ and $\tau(Q)$. This is 0. Finally, we have that $\tau(C)$ is a circle with Euclidean center 0 and Euclidean radius $|\tau(P)|$. So,

$$\tau(C) = \{z \in \mathbb{D} : |z| = |\tau(P)|\} = \{z \in \mathbb{D} : d(0, z) = \operatorname{arctanh}(|\tau(P)|)\}$$

and $\tau(C)$ is the hyperbolic circle with hyperbolic center 0 and hyperbolic radius $\operatorname{arctanh}(|\tau(P)|)$.

We finally have that $\tau\tau(C) = C$ is a hyperbolic circle. ■

Be warned. Although every Euclidean circle is a hyperbolic circle (and vice versa), the Euclidean center and hyperbolic center are almost never the same, and neither is the Euclidean radius and the hyperbolic radius. For example, we already saw above that the hyperbolic circle centered at 0 with hyperbolic radius r is the Euclidean circle centered at 0 with Euclidean radius $\tanh(r)$. And unless $r = 0$, $r \neq \tanh(r)$.

However, from the sketch of the proof above we can see that the Euclidean and hyperbolic centers of a circle both lie on the same ray emanating from 0.

Exercise. Let $C \subset \mathbb{D}$ be a circle with Euclidean center c and hyperbolic center b. Prove that 0, c, and b lie on a d-line.

Lecture 19 - 18/06

Let's return to our desired properties of hyperbolic distance. We want the distance function

$$d : \mathbb{D} \times \mathbb{D} \to \mathbb{R}_{\geq 0}$$

to satisfy for all $z, w, v \in \mathbb{D}$,

1. $d(z, w) = 0$ if and only if $z = w$.
2. $d(z, w) = d(w, z)$.
3. $d(z, w) + d(w, v) \geq d(z, v)$ (*Triangle inequality*).
4. $d(z, w) + d(w, v) = d(z, v)$ if and only if z, w, v are on a d-line with w between z and v.
5. $d(z, w) = d(\tau(z), \tau(w))$ for all $\tau \in \mathcal{G}_{\mathbb{D}}$.

Exercise. Prove that Properties 1 and 2 hold for hyperbolic distance.

A part of Property 4 is proved on Assignment 3. Property 5 will be proved through a series of exercises in the practice problems. We focus now on Property 3, the **triangle inequality**, assuming we have proved all the other properties.

During the proof, we will use the fact that hyperbolic circles are Euclidean circles, as well as the fact that Properties 3 and 4 hold for Euclidean distance and points lying on Euclidean lines.

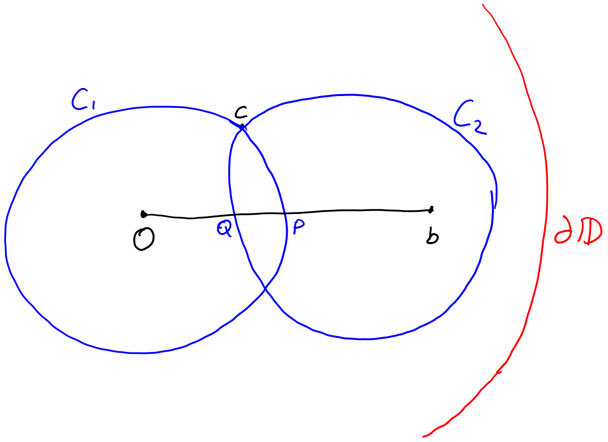
Theorem 26. *For all $u, v, w \in \mathbb{D}$, $d(u, v) + d(v, w) \geq d(u, w)$.*

Proof. Let $\tau \in \mathcal{G}_{\mathbb{D}}$ be such that $\tau(u) = 0$, $\tau(v) = c$ and $\tau(w) = b$. Since τ preserves hyperbolic distance, it suffices for us to show that $d(0, c) + d(c, b) \geq d(0, b)$.

First note that if $d(0, c) \geq d(0, b)$ or $d(c, b) \geq d(0, b)$, then we're done since $d(x, y) \geq 0$ for all $x, y \in \mathbb{D}$. So we may assume $d(0, c) < d(0, b)$ and $d(b, c) < d(0, b)$.

Let l be the d-line containing 0 and b, which is also part of a Euclidean line.

If c lies on l, it must be that c is between 0 and b. Then the desired inequality holds (and is an equality) by Property 4. So, we may assume that c is not on l.



Consider the hyperbolic circles C_1 and C_2 as

$$C_1 = \{z \in \mathbb{D} : d(0, z) = d(0, c)\} \quad \text{and} \quad C_2 = \{z \in \mathbb{D} : d(b, z) = d(b, c)\}.$$

So, C_1 is the hyperbolic circle centered at 0 passing through c, and C_2 is the hyperbolic circle centered at b passing through c. Thus C_1 and C_2 intersect at c.

Since the hyperbolic centers of C_1 and C_2 are on l, so are the Euclidean centers. Therefore C_1 and C_2 are orthogonal to l. Let σ_l by the hyperbolic reflection across l. Then $\sigma_l(C_1) = C_1$ and $\sigma_l(C_2) = C_2$. Then $\sigma_l(c)$ is also an intersection point of C_1 and C_2. Since c is not on l, c and $\sigma_l(c)$ are distinct, and are the two intersection points of C_1 and C_2. In particular, C_1 and C_2 do not intersect each other on l.

Since $d(0, c) < d(0, b)$, C_1 intersects l between 0 and b. Call this intersection point P. Since $d(b, c) < d(0, b)$, C_2 intersects l between 0 and b. Call this intersection point Q.

As a consequence of the fact that Euclidean distance satisfies the triangle inequality and Property 4, the points $0, Q, P, b$ appear on l in that order, and $Q \neq P$. By Property 4 for hyperbolic distance we have

$$\begin{aligned} d(0, c) + d(c, b) &= d(0, P) + d(Q, b) \\ &= d(0, P) + d(Q, P) + d(P, b) \\ &=> d(0, P) + d(P, b) \\ &= d(0, b) \end{aligned}$$

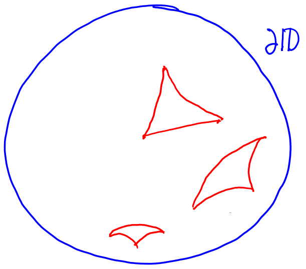
which completes the proof. ■

Lecture 20 - 20/06

3.10 Hyperbolic triangles

Let's do some geometry! At least in the sense that maybe you're familiar with. Like studying triangles, for example.

Definition. A **hyperbolic triangle** is a collection of three points in $\mathbb{D}$ that do not lie on a single d-line, together with the d-line segments joining each pair of points. If A, B, and C are the vertices of a hyperbolic triangle, we may refer to the hyperbolic triangle as ABC.



Some hyperbolic triangles

Exercise. Sketch out the hyperbolic triangle with vertices 0, $\frac{1}{2}$, and $\frac{i}{2}$.

Theorem 27. *The sum of the angles of a hyperbolic triangle is less than π.*

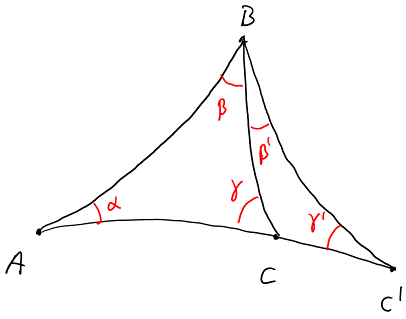
Proof. Let $\tau \in \mathcal{G}_{\mathbb{D}}$ map one of the vertices to 0, and the other two vertices to u and v. Then two of the edges of the triangle $0uv$ are segments of d-lines passing through 0, and the third is part of a Euclidean circle centered outside $\mathbb{D}$.

The Euclidean triangle $\triangle 0uv$ has angle sum π. However, the hyperbolic angles of the hyperbolic triangle $0uv$ at u and v are less than the Euclidean angles $\angle 0uv$ and $\angle 0vu$. The Euclidean angle $\angle u0v$ is equal to the hyperbolic angle at 0. Therefore the sum of the hyperbolic angles of the triangle is less than π. ■

If we draw out a few pictures, it becomes tempting to guess that the larger the triangle is, the smaller its angles are. Small triangles appear to have angle sum close to π, and larger triangles appear to have angle sum close to 0 (imagine, for instance, a hyperbolic triangle with vertices very close in the Euclidean sense to $\partial\mathbb{D}$). Let's see a concrete example of this.

Example. Let A, C, C' be distinct points on a d-line, appearing in that order (so C is between A and C'). Let B be a point not on the d-line. We will compare the hyperbolic triangles ABC and ABC', the former of which is a smaller triangle sitting inside the latter.

Let $\alpha, \beta, \beta', \gamma, \gamma'$ be the angles as indicated in the image.



Since the sum of the angles of the hyperbolic triangle CBC' are less than π, it follows that $\gamma > \gamma' + \beta'$. Then $\alpha + \beta + \gamma > \alpha + \beta + \beta' + \gamma'$. Therefore the sum of the angles of the hyperbolic triangle ABC is greater than the sum of the angles of ABC'.

Exercise. Prove that the sum of the interior angles of a hyperbolic quadrilateral is less than 2π.

Although some things about hyperbolic triangles are different from Euclidean triangles (like angle sum, for instance), some things are the same. Isosceles hyperbolic triangles behave similarly to isosceles Euclidean triangles.

Before we get into the weeds of the result, we need a fact about d-lines.

Lemma 28. *Let $P \in \mathbb{D}$ and let L be a Euclidean line contianing P. Then there is a unique d-line containing P that is tangent to L.*

Proof. Let C be a Euclidean circle orthogonal to $\partial\mathbb{D}$ so that $I_C(P) = 0$ (such a C exists by the origin lemma). Then $I_C(L)$ is a Euclidean circle or line containing 0. Let M be the Euclidean tangent line to $I_C(L)$ at 0 (if $I_C(L)$ is a line, take $M = I_C(L)$). Then $M \cap \mathbb{D}$ is a d-line containing 0. So $I_C(M) \cap \mathbb{D}$ is a d-line containing P tangent to $I_C(I_C(L)) = L$. Uniqueness is an exercise. ■

Lecture 21 - 23/06

The moral of the story is that given a point $P \in \mathbb{D}$ and a direction, there is a unique d-line passing through P, travelling in that direction. This allows us to take a d-line bisecting an angle between two d-lines, for example. The lemma also implies that if two d-lines l_1 and l_2 intersect a third d-line l at a point P, and if the angles that l makes at P with l_1 and l_2 are equal, then l_1 and l_2 are the same d-line.

Theorem 29. *Let ABC be a hyperbolic triangle. Then $d(A, B) = d(A, C)$ if and only if the hyperbolic angles satisfy $\angle ABC = \angle ACB$.*

Proof. Let l_a, l_b, and l_c be the unique d-lines passing through B and C, A and C, and A and B respectively.

Now suppose $\angle ABC = \angle ACB$, which means the angle between l_a and l_c at B is equal to the angle between l_a and l_b at C.

Let m be the hyperbolic midpoint of B and C, so m is on l_a and $d(m, B) = d(m, C)$. Let l be the unqiue d-line passing through m that is orthogonal to l_a. Then if σ_m is the hyperbolic reflection across m, $\sigma_m(l_a) = l_a$ and therefore $\sigma_m(B) = C$.

Since $\angle ABC = \angle ACB$ and since hyperbolic reflections preserve (and reverse orientation of) angles, $\sigma_m(l_b) = l_c$ and $\sigma_m(l_c) = l_b$. Since A is the intersection of l_b and l_c, $\sigma_m(A) = A$.

Therefore, $d(A, B) = d(\sigma_m(A), \sigma_m(B)) = d(A, C)$.

Lecture 22 - 25/06

Conversely, suppose $d(A, B) = d(A, C)$ and let l be the unique d-line passing through A that bisects $\angle BAC$. Let σ_l be the hyperbolic reflection across l. Since angles are preserved by σ_l, and since $\sigma_l(A) = A$, it follows that $\sigma_l(l_a) = l_b$.

Since $B \in l_c$, $\sigma_l(B) \in l_b$. Furthermore we have $d(A, C) = d(A, B) = d(\sigma_l(A), \sigma_l(B)) = d(A, \sigma_l(B))$. Therefore, $\sigma_l(B) = C$. It follows that $\sigma_l(l_a) = l_a$. Since σ_l preserves angles, we have that the angle between l_b and l_a is equal to the angle between $\sigma_l(l_b) = l_c$ and $\sigma_l(l_a) = l_a$, completing the proof. ■

We come now to a major difference in the behaviour of triangles in hyperbolic space versus in Euclidean space. In Euclidean space, you can scale triangles, which gives a whole range of triangles with the same angles, but different side lengths. Such triangles are called similar. In Euclidean geometry, if you can take one triangle to another by a Euclidean transformation (a rotation, translation, or reflection), then the triangles are congruent.

Definition. Let ABC and XYZ be hyperbolic triangles with angles α, β, γ and θ, ω, ζ at the vertices A, B, C and X, Y, Z respectively. If $\alpha = \theta$, $\beta = \omega$, and $\gamma = \zeta$, we say the triangles are **similar**.

If there exists a hyperbolic transformation $\tau \in \mathcal{G}_{\mathbb{D}}$ so that $\tau(A) = X$, $\tau(B) = Y$, and $\tau(C) = Z$, then the triangles are **congruent**.

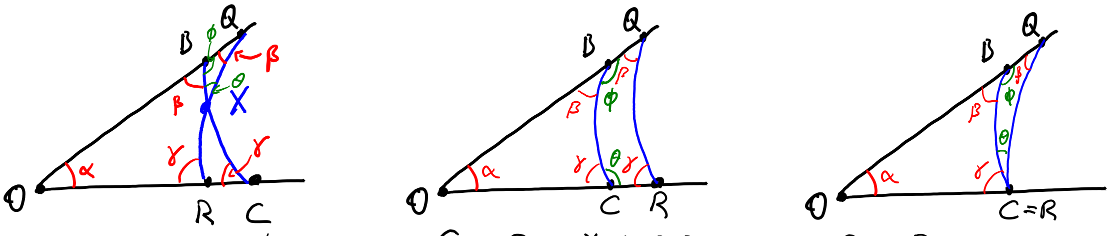
It turns out that in hyperbolic geometry, there are no similar triangles that are not congruent.

Theorem 30. *Similar hyperbolic triangles are congruent.*

Proof. Let ABC and PQR be similar hyperbolic triangles so that α is the measure of the angles at A and P, β is the measure of the angles at B and Q, and γ is the measure of the angles at C and R.

After applying hyperbolic transformations to ABC and PQR independently, we may assume that $A = P = 0$, Q lies on the ray $\overrightarrow{0B}$ and R lies on the ray $\overrightarrow{0C}$. Note that the rays $\overrightarrow{0B}$ and $\overrightarrow{0C}$, when intersected with $\mathbb{D}$, are parts of d-lines. We want to show that $B = Q$ and $C = R$.

Suppose that, towards a contradition, that B is distinct from Q, and lies between 0 and Q on the d-line segment $0Q$. There are three cases to consider.



Case 1: $\beta + \theta + \phi > \pi$ Case 2: $\beta + \gamma + \phi + \theta = 2\pi$ Case 3: $\beta + \phi + \theta > \pi$

Case 1: The point R is between 0 and C on the d-line segment $0C$. Then the d-line segment QR intersects the d-line segment BC at a point X. Then the angle sum of the hyperbolic triangle BQX is larger than π, a contradition.

Case 2: The point C is between 0 and R on the d-line segement $0R$. Then the hyperbolic quadrilateral $BQRC$ has angle sum equal to 2π, a contradiction.

Case 3: $C = R$. Then the angle sum of the triangle BCQ is larger than π, a contradiction.

So, we are forced to conclude that $B = Q$. Now let l be the d-line passing through 0 and B, m the d-line passing through 0 and C, l_a the d-line passing through B and C, and l_p the d-line passing through Q and R. Since $B = Q$ and since the angles of the triangles at B and Q are equal, we must have that $l_a = l_p$. Since C is the intersection of l_a and m, and R is the intersection of l_p and m, we conclude that $C = R$. ■

Lecture 23 - 27/06

As the vertices of a triangle get closer and closer (in the Euclidean distance) to $\partial\mathbb{D}$, the angle at that vertex gets smaller and smaller. It will turn out to be helpful to allow vertices of triangles

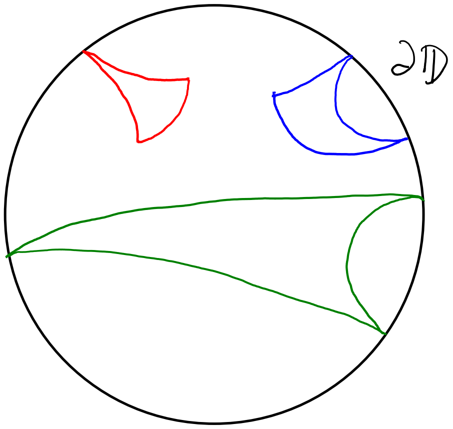


Figure 2: A simply asymptotic triangle in red, a doubly asymptotic triangle in blue, and an ideal triangle in green.

to be on $\partial\mathbb{D}$. Of course, such objects are no longer triangles, but they behave in similar ways, and can be thought of as limits of hyperbolic triangles.

Here are some examples of such "triangles."

In these triangles (which we will call asymptotic triangles), one or more of the vertices are on $\partial\mathbb{D}$.

Definition. A hyperbolic triangle with one or more vertices on $\partial\mathbb{D}$ is called an **asymptotic triangle**.

If one vertex is on $\partial\mathbb{D}$, it is **simply asymptotic**. If two vertices are on $\partial\mathbb{D}$, it is **doubly asmptotic**. If all vertices are on $\partial\mathbb{D}$, it is **trebly asymptotic**, or **ideal**.

The vertices on $\partial\mathbb{D}$ are called **ideal points** of the triangle. The angle at a vertex which is an ideal point of a triangle is 0.

Just like regular hyperbolic triangles, asymptotic triangles are defined by their vertices. Of course, for this to make sense, we need to be sure that any two points in $\mathbb{D} \cup \partial\mathbb{D}$ define a unique d-line (we already know this is true for two points in $\mathbb{D}$).

Exercise. Let $p, q \in \partial\mathbb{D}$, and $A \in \mathbb{D}$.

1. There is a unique d-line defined by a Euclidean circle C so that C intersects $\partial\mathbb{D}$ at p and q.
2. There is a unique d-line l passing through A so that the Euclidean circle C defining l intersects $\partial\mathbb{D}$ at p.

It is worth noting that the ideal points of an asymptotic triangle *are not* part of the triangle. They are simply there to help us define the triangle.

It follows from the definition that the angle sum of an ideal triangle is 0.

Here are some important exercises you can do to get friendly with some asymptotic triangles.

Exercise.

1. Sketch out the doubly asymptotic triangle with vertices $0, i, 1$. Note that i and 1 are ideal points of the triangle.
2. Prove that two doubly asymptotic triangles are congruent if and only if the angles at the non-ideal points are equal.

3. Prove that all ideal triangles are congruent.

Note that asymptotic triangles are never congruenct to hyperbolic triangles, since there is no hyperbolic transformation that maps a point in $\mathbb{D}$ to a point in $\partial\mathbb{D}$.

3.11 Area of hyperbolic triangles

We are now ready to study area of triangles. In Euclidean geometry, we can define area by filling a region with smaller and smaller rectangles. Since we know how to compute the area of a rectangle, we can compute the area of any shape. Unfortunately, rectangles in hyperbolic space are a little more complicated and don't lend themselves nicely to the study of area.

However, assuming some facts we will take for granted, and insisting that area satisfies some properties (much like we did for distance), we can compute the area of triangles.

There are certain properties we want area to have.

- Congruent regions should have the same area.
- If a region is inside another, then the first region should have a smaller area than the second.
- If a region is split up into two regions, the sum of the areas of the smaller region should be the area of the larger region.

We saw earlier that as triangles get larger, their angle sum gets smaller. Since triangles with the same angles are congruent, the area of a triangle will be some function of the angles.

To figure out what this function is, we will first take the following fact for granted.

Fact 31. *The area of an ideal triangle is finite.*

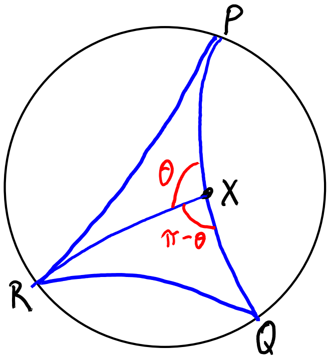
There is a proof of this with certain assumptions, but we won't cover it in this course.

Lecture 24 - 02/07

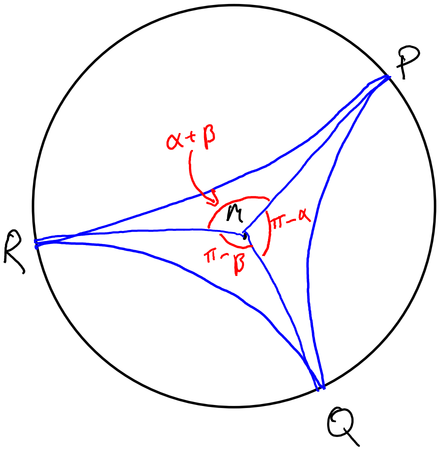
Now, let's see if we can figure out what the area of a hyperbolic triangle (which may be an asymptotic triangle).

Theorem 32. *Let T be a hyperbolic triangle with angles α, β, γ (these angles may be 0 if T is an asymptotic triangle). The area of T is proportional to $\pi - (\alpha + \beta + \gamma)$. If we set the area of an ideal triangle to be π, then the area of T is equal to $\pi - (\alpha + \beta + \gamma)$.*

Proof. Since all ideal triangles are congruent, they all have the same finite area. Call that area k. Consider an ideal triangle PQR and let X be a point on the edge PQ. This divides the ideal triangle into two doubly asymptotic triangles PXR and RXQ. Since doubly asymptotic triangles are congruent exactly when their one non-zero angles are equal, the area of a doubly asymptotic triangle is a function of the angle. Give this function a name, f, so that the area of a doubly asymptotic triangle with angle θ is $f(\theta)$.



Putting all of the names we have given things so far together we get $k = f(\theta) + f(\pi - \theta)$. Consider now an ideal triangle PQR with a point M in the interior. Joining this point to the three ideal points of the triangle splits triangle PQR into three doubly asymptotic triangles, PQM, PRM, and QRM. Name two of the angles $\pi - \alpha$ and $\pi - \beta$, and then the third angle must be $\alpha + \beta$.

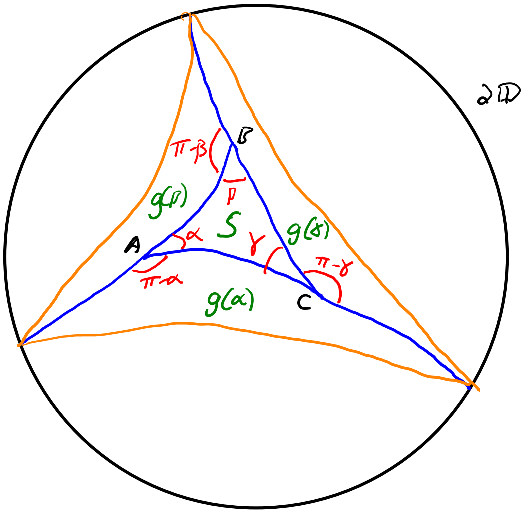


Then we have $k = f(\alpha + \beta) + f(\pi - \alpha) + f(\pi - \beta)$. However, substituting in $k = f(\theta) + f(\pi - \theta)$ with $\theta = \alpha + \beta$ gives $f(\pi - (\alpha + \beta)) = f(\pi - \alpha) + f(\pi - \beta)$.

Let $g(\theta) = f(\pi - \theta)$, so $g(\theta)$ is the area of a doubly asymptotic triangle with angle $\pi - \theta$. Then $g(\alpha + \beta) = g(\alpha) + g(\beta)$. Assuming g is continuous, we can conclude that $g(\theta) = \lambda\theta$ for some real number λ (this requires some justification, but I will skip it here).

Furthermore, since $k = f(\theta) + f(\pi - \theta)$, we have $k = g(\pi - \theta) + g(\theta) = \lambda(\pi - \theta) + \lambda\theta = \lambda\pi$. Therefore, $\lambda = \frac{k}{\pi}$.

Now consider an arbitrary hyperbolic triangle with angles α, β, γ at vertices A, B, C as shown.



Extend AC past C to meet $\partial\mathbb{D}$, extend CB past B to $\partial\mathbb{D}$, and extend BA past A to $\partial\mathbb{D}$. Then the three ideal points introduced form the ideal points of an ideal triangle. Furthermore, the ideal triangle is divided into four triangles, one of which is the original, and the other three being doubly asymptotic triangles with angles $\pi-\alpha$, $\pi-\beta$, and $\pi-\gamma$. Therefore, if S is the area of our original triangle,

$$S = k - g(\alpha) - g(\beta) - g(\gamma) = k - \frac{k}{\pi}(\alpha+\beta+\gamma) = \frac{k}{\pi}(\pi - (\alpha+\beta+\gamma).$$

Finally, if we assume $k=\pi$, $S = \pi - (\alpha+\beta+\gamma)$. ■

3.12 The upper-half plane model

We have been studying hyperbolic geometry so far in what is called the Poincaré disk model $\mathbb{D}$. We have a set of points $\mathbb{D}$, a boundary $\partial\mathbb{D}$, a group of hyperbolic transformations $\mathcal{G}_{\mathbb{D}}$, a collection of lines (d-lines) and a notion of angle and distance.

It turns out that there is another model, called the **upper-half plane model**, of hyperbolic geometry. Transformations are still built up out of circle inversions and reflections across lines, but the set of points and lines are different. The upper-half plane model has its advantages, and disadvantages, over $\mathbb{D}$, as we shall see. It will be helpful to be comfortable in both models as some problems are easier to solve in one but not the other.

Definition. Define the **upper-half plane** as

$$\mathbb{H} = \{a+bi \in \mathbb{C} : b > 0\}$$

with boundary

$$\partial\mathbb{H} = \{a+bi \in \mathbb{C} : b = 0\} \cup \{\infty\}.$$

We can transport everything we know about hyperbolic geometry in $\mathbb{D}$ over to $\mathbb{H}$ with the following Möbius transformation.

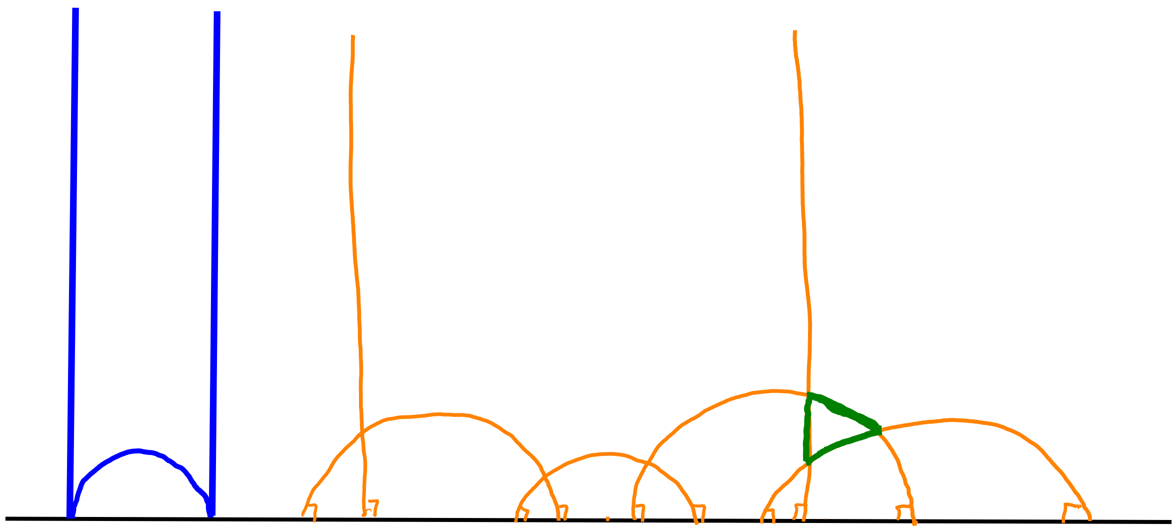


Figure 3: Some d-lines in $\mathbb{H}$ in orange, a hyperbolic triangle in green, and an ideal triangle in blue.

Example. Consider the Möbius transformation $M(z) = \frac{iz+i}{-z+1}$. Then

$$\begin{aligned} M(0) &= i \\ M(1) &= \infty \\ M(-1) &= 0 \\ M(i) &= -1 \\ M(-i) &= 1. \end{aligned}$$

Also, $M(\mathbb{D}) = \mathbb{H}$ and $M(\partial\mathbb{D}) = \partial\mathbb{H}$.

Lecture 25 - 04/07 The Möbius transformation M has inverse given by

$$M^{-1}(z) = \frac{z-i}{z+i}.$$

d-lines in $\mathbb{H}$.

Since Möbius transformations preserve Euclidean circles/lines and angle, we can figure out pretty quickly which objects should play the role of d-lines in $\mathbb{H}$.

Since a d-line in $\mathbb{D}$ is a Euclidean circle/line orthogonal to $\partial\mathbb{D}$, d-lines in $\mathbb{H}$ are Euclidean circles and lines orthogonal to $\partial\mathbb{H}$. These are exactly the set of circles centered on the real line, and the set of vertical lines, intersected with $\mathbb{H}$. Note that the vertical lines have an ideal point at ∞.

Hyperbolic transformations in $\mathbb{H}$.

Once we have d-lines, we can talk about hyperbolic transformations. Hyperbolic transformations in $\mathbb{H}$ are built up in exactly the same way as they are in $\mathbb{D}$, as compositions of hyperbolic reflections. A hyperbolic reflection in $\mathbb{H}$ is a circle inversion across a d-line (if it's defined by a Euclidean circle), or a reflection across a d-line (if the d-line is defined by a vertical Euclidean line).

Here is one place where working in the upper-half plane model has its advantages. A composition of two reflections across vertical lines is a horizontal translation. In fact, every translation of the form $t(z) = z - a$ for a real number a is a hyperbolic transformation in $\mathbb{D}$! Since it's a composition of two reflections, it's an orientation-preserving transformation. Even better, it only fixes one point on $\partial\mathbb{H}$ (the point ∞) and no points in $\mathbb{H}$, so it's a parabolic transformation. Every horizontal line in $\mathbb{D}$ is a horosphere fixed by t!

Exercise. Let $a \in \mathbb{R}$. Find two d-lines in $\mathbb{H}$ so that $r(z) = z - a$ is a composition of reflections about the two d-lines.

In Question 3 on Assignment 1, you showed that the composition of two circle inversions that shared the same Euclidean center is a dilation. Using this fact, we can see that dilations are also hyperbolic transformations! Let l_1 and l_2 be defined by two circles centered at 0. Then if σ_1 and σ_2 are the reflections across l_1 and l_2 respectively, we have $\sigma_2\sigma_1(z) = kz$ for some positive real number k. This is an example of a loxodromic transformation on $\mathbb{H}$. It fixes 0 and ∞ on $\partial\mathbb{H}$, and it has the imaginary axis as its axis.

Exercise. Let k be a positive real number. Find two d-lines in $\mathbb{H}$ so that the composition of their hyperbolic reflections is given by $r(z) = kz$.

Distance, angle, and area.
We can define distance, angle, and area similarly to the way we do in $\mathbb{D}$.

Definition. The **angle** between two intersecting curves in $\mathbb{H}$ is their Euclidean angle at a point of intersection.

We can port over our notion of area from $\mathbb{D}$ to get the following fact:

Fact 33. *The area of a hyperbolic triangle in $\mathbb{H}$ is $\pi - (\alpha + \beta + \gamma)$ where α, β, γ are the interior angles of the triangle.*

For distance, we define it for two points on the same vertical d-line, and the distance between any two other points is computed by first applying a hyperbolic transformation to ensure the two points are on the same vertical d-line.

Definition. Let $a + bi$ and $a + ci$ be two points in $\mathbb{H}$ (so $b, c > 0$). The **hyperbolic distance** between the two points is given by $d(a + bi, a + ci) = |\ln(b) - \ln(c)|$.

Exercise. Let $P, Q \in \mathbb{D}$ be two points that map under $M : \mathbb{D} \to \mathbb{H}$ to two points lying on the same vertical d-line in $\mathbb{H}$. Prove that $d(P, Q) = kd(M(P), M(Q))$ for some positive real number k.

There is much much more to say on the subject of hyperbolic geometry. In the interest of saving some space for some of the other non-Euclidean geometries, we must part ways here.

Lecture 26 - 07/07

4 Spherical geometry

We now shift our attention to doing geometry on a sphere. We will follow the same kinds of ideas as we did for hyperbolic geometry. We first define our points, and then lines, distance, and angles, and include a group of transformations. Let's get to it!

Definition. Define the set $\mathbb{S} = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}$.

So, the points in our sphereical geometry are all the points that have Euclidean distance 1 from the origin in $\mathbb{R}^3$.

Definition. A **line** in $\mathbb{S}$ is a great circle, which is the intersection of a plane through the origin in $\mathbb{R}^3$ with $\mathbb{S}$.

Great circles are the shortest path between any two points while staying on a sphere. It's the path that planes would take on the Earth if they did not have to take into account other aerial traffic, flying over emergency landing locations, and other such aviation-related concerns.

Definition. The **distance** between two points $p, q \in \mathbb{S}$, denoted $d(p, q)$ is the angle (in radians) between the two vectors p and q. We choose $d(p, q)$ so that $0 \leq d(p, q) \leq \pi$.

Recall that the angle θ between two vectors $v, w \in \mathbb{R}^3$ is given by $\|v\| \, \|w\| \cos(\theta) = v \cdot w$ (the dot pruduct of v and w). However, for all vectors on $\mathbb{S}$ have size 1, so for $p, q \in \mathbb{S}$, $d(p, q)$ satisfies $\cos(d(p, q)) = p \cdot q$.

Example. If $p = (1, 0, 0)$ and $q = (0, 1, 0)$, then $p \cdot q = 0$. Therefore $\cos(d(p, q)) = 0$ and $d(p, q) = \frac{\pi}{2}$.

For any $p \in \mathbb{S}$, $p \cdot (-p) = -1$ and so $d(p, -p) = \pi$.

Definition. The angle between two intersecting curves in $\mathbb{S}$ is the angle between the tangent lines in $\mathbb{R}^3$ to the curves at the point of intersection.

We will mainly be concerned with the angle between two lines in $\mathbb{S}$. Recall from your linear algebra days that a plane in $\mathbb{R}^3$ is defined by a point that the plane passes through, and a normal vector to the plane.

Proposition 34. *The angle between two lines at an intersection point is the angle between the normal vectors of the planes defining the lines.*

Proof. This is an exercise. ■

We have a particularly nice way of describing lines in $\mathbb{S}$. Given a point $p \in \mathbb{S}$, we can consider it as the normal vector for some plane passing through the origin. Such a plane defines the line defined by p.

Definition. Let $p = (p_1, p_2, p_3) \in \mathbb{S}$ and let P be the plane through the origin with normal vector p. The **line defined by** p is $\mathbb{S} \cap P = \{(x, y, z) \in \mathbb{S} : p_1 x + p_2 y + p_3 z = 0\}$.

So, for example, the line defined by $(0, 0, 1)$ is the intersection of the xy-plane and $\mathbb{S}$.

Exercise. Show that the angle between the lines defined by p and q is equal to the distance $d(p, q)$ in $\mathbb{S}$.

Lecture 27 - 09/07

4.1 Spherical transformations

Just like in hyperbolic geometry, we will build up the group of transformations from reflections across lines in $\mathbb{S}$. Thankfully, these reflections are a little more familiar to us than the circle inversions from hyperbolic geometry.

Recall that if $\hat{n}$ is a unit e normal vector to a plane passing through the origin in $\mathbb{R}^3$, then the reflection across the plane is given by

$$R(v) = v - 2(v \cdot \hat{n})\hat{n}$$

where $v \cdot \hat{n}$ is the dot product between v and $\hat{n}$.

Definition. Let $p \in \mathbb{S}$ and let l be the line defined by p. The **reflection** σ across l is the map

$$\sigma : \mathbb{S} \to \mathbb{S}$$

given by $\sigma(v) = v - 2(v \cdot p)p$.

Of course, we need to make sure that σ is indeed a map from $\mathbb{S}$ to $\mathbb{S}$.

Exercise. Show that $\sigma(\mathbb{S}) = \mathbb{S}$ where σ is a reflection across a line in $\mathbb{S}$.

Example. Suppose $p = (0, 0, 1)$, and σ is the reflection across the line defined by p. Then σ simply negates the z-coordinate. So, for example, $\sigma(0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) = (0, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$.

Reflections in $\mathbb{S}$ act similarly to reflections in $\mathbb{D}$ or $\mathbb{H}$, with the added bonus that they are linear maps!

Exercise. Let σ be a reflection across a line l in $\mathbb{S}$.

1. Show that σ is defined by a linear map, and therefore can be represented by a 3×3 matrix.
2. Show that $\sigma^2(p) = p$ for all $p \in \mathbb{S}$.

The major attraction of thinking about reflections is that they will turn out to respect all the geometry we have talked about so far: lines, distance, and angle. To prove this, we need the following super-useful lemma.

Lemma 35. *Let $\sigma : \mathbb{S} \to \mathbb{S}$ be the reflection across the line l defined by p. Then for all $v, w \in \mathbb{S}$, $v \cdot w = \sigma(v) \cdot \sigma(w)$.*

Proof. Using properties of the dot product in $\mathbb{R}^3$ we have

$$\begin{aligned}\sigma(v) \cdot \sigma(w) &= (v - 2(v \cdot p)p) \cdot (w - 2(w \cdot p)p) \\ &= v \cdot w - 2(v \cdot p)(p \cdot w) - 2(w \cdot p)(v \cdot p) + 4(v \cdot p)(w \cdot p) \\ &= v \cdot w\end{aligned}$$

completing the proof. ■

Why is this a useful thing to have? Well, the dot product in $\mathbb{R}^3$ gives us all the geomtry in $\mathbb{R}^3$. After all, we can define both distance and angle in terms of the dot product. Similarly, the dot product is intimately intertwined with all our notions in spherical geometry we have seen so far. The points in $\mathbb{S}$ are those whose dot product with themselves is 1. A line in $\mathbb{S}$ defined by a point p is all the points in $\mathbb{S}$ whose dot product with p is 0. The distance d between two points p and q is given by $\cos(d) = p \cdot q$. So, since reflections preserve the dot product, we should expect them to preserve a lot of the geometry in $\mathbb{S}$.

Proposition 36. *Let $\sigma : \mathbb{S} \to \mathbb{S}$ be the reflection across a line in $\mathbb{S}$. Then*

1. *If l is a line in $\mathbb{S}$, then $\sigma(l)$ is a line in $\mathbb{S}$.*
2. *If $v, w \in \mathbb{S}$, then $d(v, w) = d(\sigma(v), \sigma(w))$.*
3. *The angle between lines is preserved by σ.*

Proof. 1. Let l be the line defined by $p \in \mathbb{S}$, so $l = \{v \in \mathbb{S} : v \cdot p = 0\}$. We wish to show $\sigma(l) = \{v \in \mathbb{S} : v \cdot \sigma(p) = 0\}$. To this end, suppose $w \in \sigma(l)$, so $w = \sigma(v)$ for some $v \in l$. Then

$$w \cdot \sigma(p) = \sigma(v) \cdot \sigma(p) = v \cdot p = 0$$

and so $w \in \{v \in \mathbb{S} : v \cdot \sigma(p) = 0$. Therefore $\sigma(l) \subset \{v \in \mathbb{S} : v \cdot \sigma(p) = 0\}$. Conversely, suppose $w \cdot \sigma(p) \in 0$. Then $\sigma(\sigma(w)) \cdot \sigma(p) = 0$ so $\sigma(w) \cdot p = 0$. Therefore $\sigma(w) \in l$ and so $w = \sigma(\sigma(w)) \in \sigma(l)$. Finally, we can conclude $\sigma(l)$ is the line defined by $\sigma(p)$.

2. Since $d(v, w)$ is defined by $\cos(d(v, w)) = v \cdot w$, we have

$$\cos(d(\sigma(v), \sigma(w))) = \sigma(v) \cdot \sigma(w) = v \cdot w = \cos(d(v, w)).$$

Since we insist distance is between 0 and π we can conclude $d(\sigma(v), \sigma(w)) = d(v, w)$.

3. This is an exercise.

■

Lecture 28 - 11/07

We now shift our attention to rotations, which are the orientation-preserving spherical transformations. It will turn out that every rotation is the composition of two reflections.

Example. The rotation about the z-axis by an angle of θ is given by the matrix

$$\begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

We call this tranformation $R(p, \theta)$ where $p = (0, 0, 1)$. Geometrically, it's a rotation by θ counterclockwise, if you're standing at p looking back towards the origin.

Definition. Let $p \in \mathbb{S}$ and $\theta \in \mathbb{R}$. Then

$$R(p, \theta) : \mathbb{S} \to \mathbb{S}$$

is the rotation by θ about the Euclidean line in $\mathbb{R}^3$ containing p and the origin. The angle θ is measured counterclockwise when standing at p and looking towards the origin.

Note that $R(p, \theta) = R(p, \theta + 2\pi k) = R(-p, -\theta)$ for all $k \in \mathbb{Z}$. Let's explore the rotations that fix the x-, y-, and z-axes in a little more depth.

Example. Let $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, and $e_3 = (0, 0, 1)$. Then $R(e_1, \theta)$, $R(e_2, \theta)$, and $R(e_3, \theta)$ are given by the matrices

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{bmatrix}, \quad \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

respectively. So, for example,

$$R(e_2, \frac{\pi}{2}) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{bmatrix}.$$

As with hyperbolic transformations, every spherical transformation has an inverse, and we can compose them.

Exercise. Prove that the inverse of $R(p,\theta)$ is $R(p,-\theta)$.

Example. The composition $R(e_1,\frac{\pi}{2})\circ R(e_2,\frac{\pi}{2})$ is given by the matrix

$$\begin{bmatrix}1 & 0 & 0\\ 0 & \cos(\frac{\pi}{2}) & -\sin(\frac{\pi}{2})\\ 0 & \sin(\frac{\pi}{2}) & \cos(\frac{\pi}{2})\end{bmatrix}\begin{bmatrix}\cos(\frac{\pi}{2}) & 0 & \sin(\frac{\pi}{2})\\ 0 & 1 & 0\\ -\sin(\frac{\pi}{2}) & 0 & \cos(\frac{\pi}{2})\end{bmatrix}=\begin{bmatrix}0 & 0 & 1\\ 1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix}.$$

This linear map moves e_1 to e_2, e_2 to e_3, and e_3 to e_1. This is a rotation! In fact,

$$R(e_1,\frac{\pi}{2})\circ R(e_2,\frac{\pi}{2})=R(q,\frac{2\pi}{3})$$

where $q=\frac{1}{\sqrt{3}}(1,1,1)$.

We have another way of expressing the points in $\mathbb{S}$. Just like on the surface of the Earth, we can determine a point by specifying a line of longitude and a line of latitude.

We can do a similar thing on $\mathbb{S}$ and we get **spherical coordinates**.

Definition. For a point $p=(p_1,p_2,p_3)\in\mathbb{S}$, the **colatitude** θ is the angle p makes with $(0,0,1)$, measured from $(0,0,1)$. We insist that the colatitutde satisfies $0\le\theta\le\pi$.

The **longitude** ϕ of p is the angle the point $(p_1,p_2,0)$ makes with $(1,0,0)$, measured counterclockwise from $(1,0,0)$. We insist that the longitude satisfies $0\le\phi<2\pi$.

If the colatitude of $p\in\mathbb{S}$ is θ and the longitude is ϕ, then the coordinates of p can be recovered by

$$p=(\cos(\phi)\sin(\theta),\sin(\phi)\sin(\theta),\cos(\theta)).$$

Note that we call θ the colatitude because the regular latitude (as we're used to on Earth) would be given by $\frac{\pi}{2}-\theta$.

Example. The point $p=(0,1,0)$ has longitude $\frac{\pi}{2}$ and colatitude $\frac{\pi}{2}$. The point $p=(1,0,0)$ has longitude 0 and colatitude $\frac{\pi}{2}$.

Lecture 29 - 14/07

Example. Let $p=\left(\frac{\sqrt{3}}{2\sqrt{2}},\frac{\sqrt{3}}{2\sqrt{2}},\frac{1}{2}\right)$. Then $p\cdot e_3=\frac{1}{2}$ so $\cos(\theta)=\frac{1}{2}$ and $\theta=\frac{\pi}{3}$. In the xy-plane, the x and y coordinates are equal (and positive) so $\phi=\frac{\pi}{4}$. Therefore the colatitude of p is $\theta=\frac{\pi}{3}$ and the longitude is $\theta=\frac{\pi}{4}$.

The main reason to care about polar coordinates is that it gives us a convenient way to describe where the north pole (that is the vector $e_3=(0,0,1)$) gets sent to under certain rotations. The following result is analogous to the origin lemma from hyperbolic geometry.

Proposition 37. *Let $p=(\cos(\phi)\sin(\theta),\sin(\phi)\sin(\theta),\cos(\theta))$. Then $R(e_3,\phi)R(e_2,\theta)(e_3)=p$.*

Proof. Looking at the matrices for these rotations we have

$$\begin{bmatrix}\cos(\phi) & -\sin(\phi) & 0\\ \sin(\phi) & \cos(\phi) & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\cos(\theta) & 0 & \sin(\theta)\\ 0 & 1 & 0\\ -\sin(\theta) & 0 & \cos(\theta)\end{bmatrix}\begin{bmatrix}0\\0\\1\end{bmatrix}=\begin{bmatrix}\cos(\phi)\sin(\theta)\\ \sin(\phi)\sin(\theta)\\ \cos(\theta)\end{bmatrix}.$$

Tehrefore $R(e_3,\phi)R(e_2,\theta)(e_3)=p$. ■

Great! A direct computation, but where did it come from? Well, imagine trying to get the north pole to a point with colatitude θ and longitude ϕ. The first thing to do is to rotate from the north pole down towards the positive x-axis by θ (which involves rotating around the y-axis). Then we rotate counterclockwise from the positive x-axis by ϕ, around the z-axis. This composition of rotations is precisely $R(e_3, \phi)R(e_2, \theta)$.

Note that a rotation that sends p to e_3 is given by $R(e_2, -\theta)R(e_3, -\phi)$.

Our next goal is to investigate what happens when we compose two reflections or rotations.

Example. Let $p = (-\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}})$. Let $\sigma_p : \mathbb{S} \to \mathbb{S}$ be the reflection across the line defined by p. Let's consider the composition $\sigma_p \sigma_{e_3}$. We have

$$\begin{aligned}\sigma_p\sigma_{e_3}(e_1) &= e_3\\ \sigma_p\sigma_{e_3}(e_2) &= e_2\\ \sigma_p\sigma_{e_3}(e_3) &= -e_1\end{aligned}$$

and therefore the matrix of the composition is given by

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

which is thematrix for $R(e_2, \frac{-\pi}{2})$!

So, we just saw that two reflections composed together is a rotation. In fact, this is always true, and should remind you of what happens when you compose two hyperbolic reflections in $\mathbb{D}$ across d-lines that pass through 0. Here are some fun facts that we will use without proof for the rest of the course.

Fact 38. *The composition of two reflections in $\mathbb{S}$ is a rotation. The composition of two rotations is a rotation.*

Exercise. Let $p, q \in \mathbb{S}$, σ_p, σ_q the reflections across the lines defined by p and q. Show that $\sigma_q\sigma_p = R(\widehat{p \times q}, \theta)$ where θ can be computed from $p \times q$.

Lecture 30 - 16/07

Definition. A **spherical transformation** is a finite composition of reflections in $\mathbb{S}$. The group of all spherical transformations is called the **spherical transformation group**, denoted $\mathcal{G}_{\mathbb{S}}$, or the **orthogonal group**, denoted $\mathrm{O}(3)$.

A transformation $\tau \in \mathrm{O}(3)$ is **orientation-preserving** if it is the composition of an even number of reflections. It is **orientation-reversing** otherwise. The set of all orientation-preserving spherical transformations is the **special orthogonal group**, denoted $\mathrm{SO}(3)$.

As a consequence of Fact 38 we have that every $\tau \in \mathrm{SO}(3)$ is the composition of two reflections (the same reflection if τ is the identity), and every $\tau \in \mathrm{O}(3)$ is the composition of three reflections.

Exercise. Let $A \in \mathrm{O}(3)$ represent a reflection. Show it has eigenvalues $1, 1, -1$ and therefore $\det(A) = -1$.

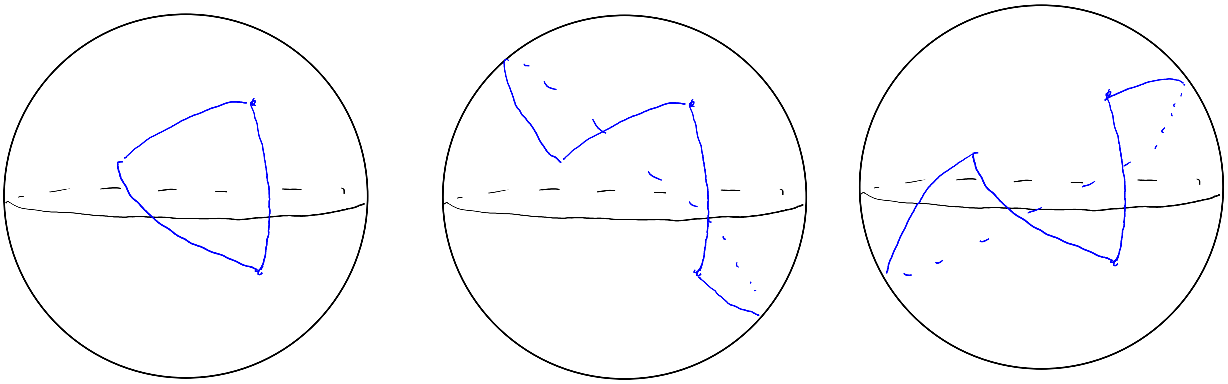
With this exercise in mind, and remembering that for matrices $\det(A)\det(B) = \det(AB)$, we have that $\tau \in \mathrm{SO}(3)$ if and only if the matrix representing τ has determinant 1.

Lecture 31 - 18/07

4.2 Spherical triangles

We now shift gears to studying triangles in $\mathbb{S}$. Given $p, q, r \in \mathbb{S}$, there are many triangles with vertices p, q, r! This is due to the fact that we can choose two line segments between any two points. These two line segments correspond to different parts of the great circle passing through two points (assuming they're not antipodal of course). The two line segments have lengths that sum to π.

But the situation gets even worse! For each triangle you draw, you can choose one of two regions in the sphere that is the interior of the triangle. Take a look at the following pictures, which give several different triangles with the same three vertices.



Exercise. Given three distinct points $p, q, r \in \mathbb{S}$, no two of which are antipodal, how many possible triangles are there?

The main topic of study with these triangles is the computation of their area. Let's start with an example, and remember, the surface area of $\mathbb{S}$ is 4π.

Example. Consider a triangle with vertices e_1, e_2, and e_3. There is a small(ish) triangle with accounts for exactly one-eigth of the surface area of $\mathbb{S}$. Therefore the area of T is $\frac{\pi}{2}$. However, there are several other triangles with these vertices. One of them has area $4\pi - \frac{\pi}{2}$ (the complement of the small(ish) one). Another has area $2\pi - \frac{\pi}{2}$. Can you find it?

In order to come up with a formula for the area of a spherical triangle, we first need to talk about lunes.

Definition. A **lune** is a 2-sided polygon defined by two great circles. The angle of the lune is the angle between the two great circles on the interior of the lune. Such an angle may be greater than π.

Note that given two great circles, there is a choice as to which region gives the lune.

Lemma 39. *The area of a lune with angle α is 2α.*

Proof. The lune covers $\frac{\alpha}{2\pi}$ of the surface of the sphere. Therefore the area is $4\pi \frac{\alpha}{2\pi} = 2\alpha$. ■

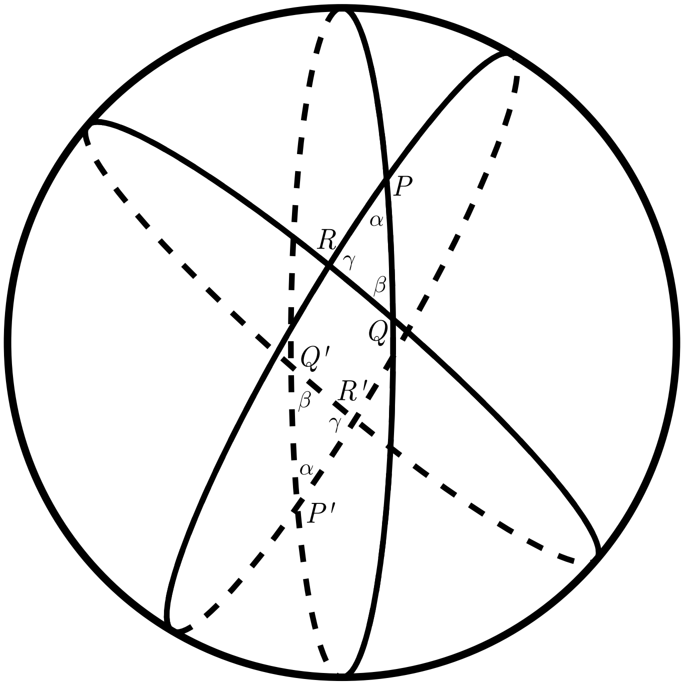
Before we get to the area of a spherical triangle, we need the following fact about spherical transformations. Again, we will state it without proof.

Fact 40. *If R is a region in $\mathbb{S}$ and $\tau \in \mathrm{O}(3)$, then $\tau(R)$ and R have the same area.*

Theorem 41. *A spherical triangle with internal angles α, β, γ has area $\alpha + \beta + \gamma - \pi$.*

Proof. We begin by first proving the result for triangles where every side length is $\leq \pi$. Consider the triangle T with vertices P, Q, and R, and area A. Let α, β, and γ be the internal angles of T at P, Q, and R respectively. Extend each edge of the triangle to the entire great circle. The three great circles also intersect at $P' = -P$, $Q' = -Q$, and $R' = -R$.

Let $\tau \in \mathrm{O}(3)$ be given by $\tau(x) = -x$ for all $x \in \mathbb{S}$. Then the three great circles divide $\mathbb{S}$ up into eight regions, one of which is T and one of which is $\tau(T)$ (the copy of the original triangle at the back of the circle).



The three great circles form two lunes of each angle α, β, and γ. Every point on $\mathbb{S}$ is either in T, $\tau(T)$ or exactly one of the six lunes. Furthermore, each of T and $\tau(T)$ is the intersection of exactly one lune of each angle.

So, if we add up the areas of all six lunes, we will account for the surface area of the entire sphere, as well as overcounting the area of T and $\tau(T)$ twice. Therefore

$$4\pi = 4\alpha + 4\beta + 4\gamma - 4A$$

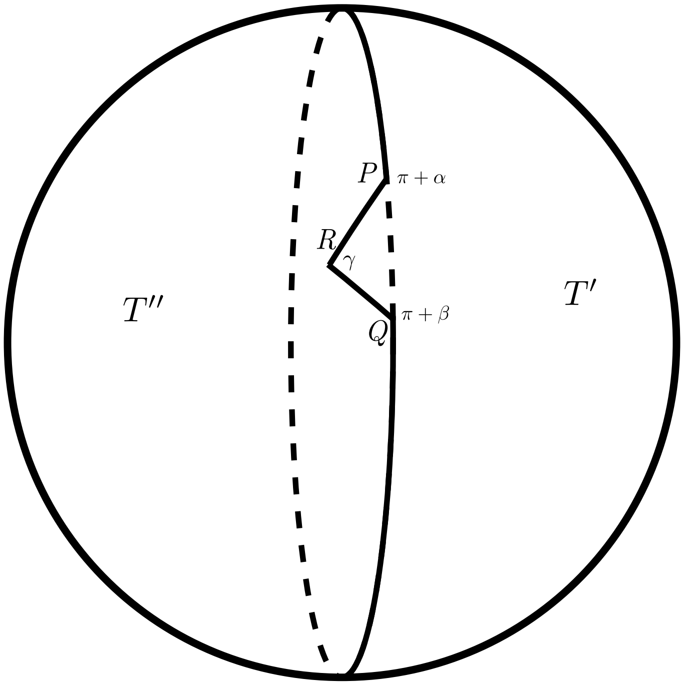
and rearranging gives $A = \alpha + \beta + \gamma - \pi$.

Now we have to deal with the other triangles with vertices P, Q, and R. The first one is the complement of the triangle T. This triangle has area $4\pi - A$, and internal angles $2\pi - \alpha, 2\pi - \beta$, and $2\pi - \gamma$. We have

$$(2\pi - \alpha) + (2\pi - \beta) + (2\pi - \gamma) - \pi = 5\pi - (\alpha + \beta + \gamma) = 4\pi - A$$

which verifies the area formula for the triangle that is the complement of T.

All other triangles with vertices P, Q, and R have exactly one edge with length $\geq \pi$ (as an exercise, figure out why two of the lengths cannot be $> \pi$). Without loss of generality, let the edge joining P and Q have length $\geq \pi$. There are two triangles here, call them T' and T'' as shown.



The area of T' is $2\pi + A$ and the area of T'' is $2\pi - A$. The interior angles of T' are $\pi+\alpha, \pi+\beta$, and γ. Checking the formula we have

$$(\pi+\alpha) + (\pi+\beta) + \gamma - \pi = 2\pi + (\alpha+\beta+\gamma-\pi) = 2\pi + A$$

which shows that T' satisfies the formula. The triangle T'' has angles $\pi-\alpha$, $\pi-\beta$, and $2\pi-\gamma$. We have

$$(\pi-\alpha) + (\pi-\beta) + (2\pi-\gamma) - \pi = 2\pi - (\alpha+\beta+\gamma-\pi) = 2\pi - A$$

completing the proof. ■

Hooray! We now have the formula for the area of a spherical triangle, which, just like the hyperbolic case, depends only on the sum of the interior angles.

Corollary 42. *The sum of the interior angles of a triangle is*

- $<\pi$ *if and only if it is a hyperbolic triangle,*
- $=\pi$ *if and only if it is a Euclidean triangle,*
- $>\pi$ *if and only if it is a spherical triangle.*

Lecture 32 - 21/07

5 Projective Geometry

Definition. The **real projective plane** $\mathbb{RP}^2$ is the set of all lines through the origin in $\mathbb{R}^3$. A line through the origin is a **point** in $\mathbb{RP}^2$.

Definition. The expression $[a, b, c]$ where a, b, c are not all 0, represents the point in $\mathbb{RP}^2$ given by the line through $(a, b, c) \in \mathbb{R}^3$. We say that $[a, b, c]$ are the **homogeneous coordinates** of the point P passing through (a, b, c).

As a warning, homogenous coordinates are not unique! In fact, for any $\lambda \in \mathbb{R}$, $\lambda \neq 0$, we have $[a, b, c] = [\lambda a, \lambda b, \lambda c]$.

5.1 Projective lines

Definition. A **projective line** (or **line**) in $\mathbb{RP}^2$ is a plane in $\mathbb{R}^3$ passing through the origin. Points in $\mathbb{RP}^2$ are **collinear** if they lie on a line.

Recall that the general equation for a plane passing through the origin in $\mathbb{R}^3$ is $ax + by + cz = 0$, where not all a, b, c are 0. Just like in Euclidean and hyperbolic geometry, every pair of points determines a unique line.

Proposition 43. *Any two distinct points in $\mathbb{RP}^2$ lie on a unique line.*

Proof. Let $P = [v_1, v_2, v_3]$ and $Q = [w_1, w_2, w_3]$ be distinct points in $\mathbb{RP}^2$. Since they are distinct, the vectors (v_1, v_2, v_3) and (w_1, w_2, w_3) in $\mathbb{R}^3$ are linearly independent. Therefore the cross product $(v_1, v_2, v_3) \times (w_1, w_2, w_3) = (a, b, c)$ is not the zero vector (so a, b, c are not all 0). Then the plane in $\mathbb{R}^3$ defined by $ax + by + cz = 0$ contains (v_1, v_2, v_3) and (w_1, w_2, w_3), and so the projective line defined by the plane contains P and Q. Uniqueness is left as an exercise. ■

Recall that three vectors in $\mathbb{R}^3$ are linearly dependent if and only if they lie in the same 2-dimensional subspace. This gives us a quick computational way to check whether or not three points in $\mathbb{RP}^2$ are collinear.

Exercise. Let $P = [2, 1, 3]$, $Q = [1, 2, 1]$, $R = [-1, 4, -3]$ in $\mathbb{RP}^2$. Compute $\begin{vmatrix} 2 & 1 & 3 \\ 1 & 2 & 1 \\ -1 & 4 & -3 \end{vmatrix}$ and determine whether or not P, Q, and R are collinear.

We now come to an interesting property of projective geometry, which sets it apart from hyperbolic and Euclidean geometry (but shows it's closer to spherical geometry!). In Euclidean geometry (and hyperbolic geometry) there are parallel lines. In projective geometry, every two lines intersect. Uniquely!

Proposition 44. *Any two distinct lines in $\mathbb{RP}^2$ intersect at a single point.*

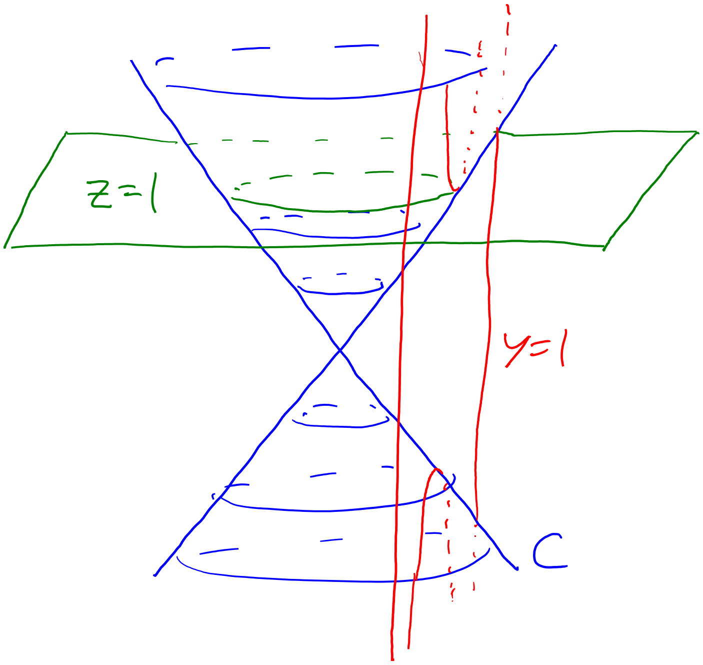
Proof. Two distinct projective lines correspond to two distinct 2-dimensional subspaces of $\mathbb{R}^3$ (since the planes defining projective lines must pass through the origin in $\mathbb{R}^3$). In $\mathbb{R}^3$, the intersection of two planes is either a plane (if the planes are equal) or a line. Since the projective lines are distinct, the planes defining them intersect at a line in $\mathbb{R}^3$ through the origin. This line corresponds to a point, which is the unique intersection point of the two projective lines. ■

Lecture 33 and beyond!

5.2 Embedding planes

Given a subset of $\mathbb{RP}^2$, we can imagine it being projected onto a plane that does not intersect the origin in $\mathbb{R}^3$. Depending on which plane we choose, the image can look very different.

Consider for example the subset C of $\mathbb{RP}^2$ defined by the equation $x^2 + y^2 = z^2$. The set of points in $\mathbb{R}^3$ satisfying this equation is a cone.



Let's check that these points actually define a set of points in $\mathbb{RP}^2$. To do this, we just need to check that if (a, b, c) is in the region, then $(\lambda a, \lambda b, \lambda c)$ is also in the region, for all $\lambda \in \mathbb{R} \setminus \{0\}$. To that end, if (a, b, c) satisfies the equation, then $a^2 + b^2 = c^2$. For a non-zero real number λ we have $(\lambda a)^2 + (\lambda b)^2 = \lambda^2(a^2 + b^2) = \lambda^2 c^2 = (\lambda c)^2$. Therefore $(\lambda a, \lambda b, \lambda c)$ also satisfies the equation.

We can conclude that the set of points in $\mathbb{R}^3$ satisfying the equation $x^2 + y^2 = z^2$ is a union of lines through the origin in $\mathbb{R}^3$, and therefore is a union of points in $\mathbb{RP}^2$.

Now, we can choose different planes in $\mathbb{R}^3$ with which to intersect this region, and we get different perspectives on the same subset C of $\mathbb{RP}^2$.

For example, consider the plane $z = 1$. This gives us the set of points $(x, y, 1) \in \mathbb{R}^3$ satisfying $x^2 + y^2 = 1$, which is a circle. In fact, every point in C intersects $z = 1$, so we get the full set of points in C represented in the plane $z = 1$.

Consider now the plane $y = 1$. This gives us the set of points $(x, 1, z)$ satisfying $z^2 - x^2 = 1$, which is a hyperbola. In this case there are two points in C that do not intersect $y = 1$. These points are $[1, 0, 1]$ and $[-1, 0, 1]$. These two points are examples of ideal points for the plane $y = 1$.

Definition. Let Π be a plane in $\mathbb{R}^3$ that does not contain the origin. An **ideal point** of Π is a point in $\mathbb{RP}^2$ represented by a line that does not intersect Π. An **embedding plane** is a plane Π in $\mathbb{R}^3$ that does not contain the origin, together with all its ideal points.

A choice of embedding plane is a way of turning points in $\mathbb{RP}^2$ into points in the embedding plane. Intiuitively, we can think of the ideal points as points at infinity of the embedding plane, one for each direction.

5.3 Projective transformations

When we studied hyperbolic and spherical geometry (and Euclidean geometry in a past life), there is the all important group of transformations. This group of transformations moves the plane around in such a way that preserves certain properties (distance and angle usually).

For us, the property we are interseted in preserving is incidence. That is, points of intersection between two projective lines, and projective lines that connect two points. However, before incidence properties are preserved, we need to first make sure that points get sent to points, and lines get sent to lines.

Points in $\mathbb{RP}^2$ are lines through the origin in $\mathbb{R}^3$, which are one-dimensional subspaces. Lines in $\mathbb{RP}^2$ are planes through the origin in $\mathbb{R}^3$, which are two-dimensional subspaces. So, whatever our transformations are, they better map one-dimensional subspaces to one-dimensional subspaces, and two-dimensional subspaces to two-dimensional subspaces. Luckily, we know a certain set of transformations of $\mathbb{R}^3$ that do this: invertible linear maps!

First some notation. For a vector $v = (a, b, c) \in \mathbb{R}^3$, we denote by $[v]$ the point $[a, b, c]$ in $\mathbb{RP}^2$.

Definition. A **projective transformation** $\tau : \mathbb{RP}^2 \to \mathbb{RP}^2$ is a function of the form $\tau([v]) = [Av]$ where A is an invertible 3×3 matrix. The collection of all projective transformations is called the **group of projective transformations** or the **projective linear group**, and is denoted $\mathcal{G}_{\mathbb{RP}^2}$ or $\mathrm{PGL}_3(\mathbb{R})$.

It's worth noting that two different matrices can define the same projective transformation. For example, let $A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$. Then $[Av] = [2v] = [v]$ in $\mathbb{RP}^2$, and therefore the projective map τ given by $\tau([v]) = [Av]$ is the same as the identity map (given by the identity matrix)!

Exercise. Let A and B be invertible real 3×3 matrices. Prove that $[Av] = [Bv]$ for all $v \in \mathbb{R}^3$ if and only if $B = \lambda A$ for some non-zero real number λ.

Lecture 34 - 25/07

Now, we want to make sure that projective transformations satisfy some basic properties. The following properties follow from the fact that the map $v \mapsto Av$ for a matrix A is a linear map of $\mathbb{R}^3$ to itself. The details of the proof are left as an exercise, and are a good refresher in exploiting properties of linear maps.

Proposition 45. *Let $\tau \in \mathrm{PGL}_3(\mathbb{R})$.*

1. *If $l \subset \mathbb{RP}^2$ is a projective line, then $\tau(l) \subset \mathbb{RP}^2$ is a projective line.*
2. *If l_1 and l_2 are projective lines intersecting at P, then $\tau(l_1)$ and $\tau(l_2)$ intersect at P.*
3. *If l is the unique projective line containing the distint points $P, Q \in \mathbb{RP}^2$, then $\tau(l)$ contains $\tau(P)$ and $\tau(Q)$.*

Proof. Exercise. ■

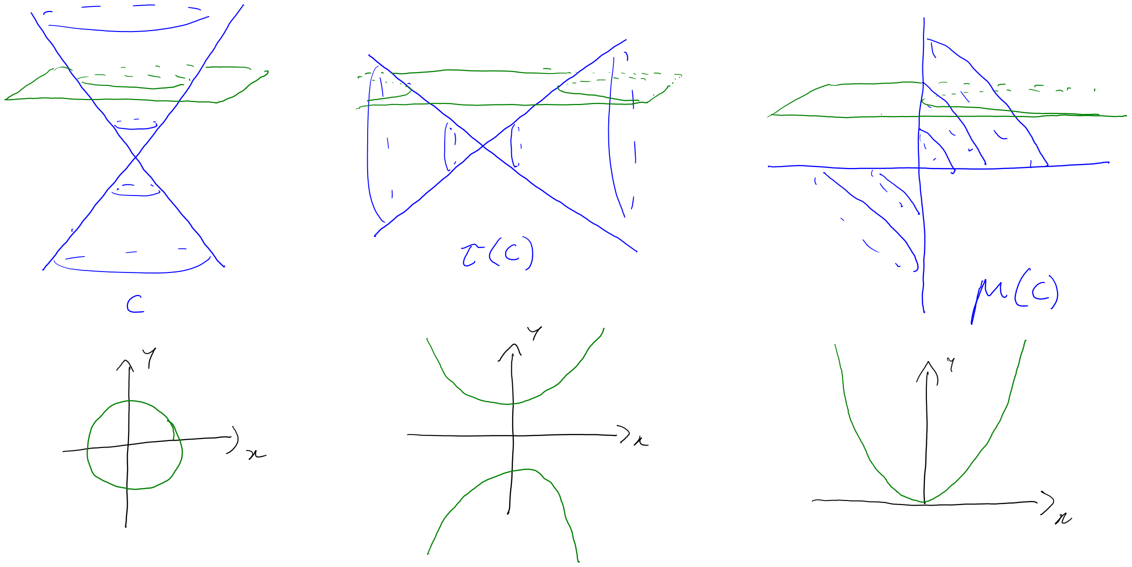
Example. Consider the subset $C = \{[x, y, z] \in \mathbb{RP}^2 : x^2 + y^2 = z^2\}$, which is the cone from the previous lecture. When look at C in the embeddding plane Π given by $z = 1$, we have the circle $x^2 + y^2 = 1$. Now let's apply some projective transformations and see what happens.

Let $\tau \in \mathrm{PGL}_3(\mathbb{R})$ be given by $\tau([v]) = [Av]$ for $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}$. If you've been paying attention, you will recognise this as a rotation about the x-axis by $\frac{\pi}{2}$. As an exercise, you can check that $\tau(C) = \{[x, y, z] \in \mathbb{RP}^2 : x^2 + z^2 = y^2\}$. In Π, $\tau(C)$ is all the points $(x, y, 1)$ on Π so that $y^2 - x^2 = 1$, along with the two ideal points $[1, 1, 0]$ and $[1, -1, 0]$. This is a hyperbola!

Let $\mu \in \mathrm{PGL}_3(\mathbb{R})$ be given by $\mu([v]) = [Bv]$ where B is the rotation matrix $R(e_1, -\frac{\pi}{4})$. As another exercise, you can check that

$$\mu(C) = \{[x, y, z] \in \mathbb{RP}^2 : x^2 = 2yz\}.$$

In Π, $\mu(C)$ is all the points $(x, y, 1)$ satisfying $y = \frac{1}{2}x^2$, along with the single ideal point $[0, 1, 0]$. This is a parabola!



So, what we have seen is that the circle, hyperbola, and parabola above all differ from each other by a projective transformation. That is, they are **projective-congruent**.

Let Π be any embedding plane. Recall that a subset of Π defines a subset of $\mathbb{RP}^2$ by taking the set of all lines in $\mathbb{R}^3$ through points in the subset (along with any ideal points in the subset). Conversely, given a subset of $\mathbb{RP}^2$, we get a subset of Π by intersecting the lines representing the points in the subset with the plane.

Definition. Two subsets of $\mathbb{RP}^2$ (or equivalently, two subsets of an embedding plane Π) are **projective-congruent** (or just **congruent** if the context is clear) if there is a projective transformation mapping one to the other.

Example. Suppose we have three points $[u], [v], [w] \in \mathbb{RP}^2$ that are not collinear (that is, they do not all lie on the same projective line). If none of the three points are ideal points of an embedding plane Π, then in the plane defining Π, the three points form the vertices of a triangle.

Since $[u], [v], [w]$ are not collinear, the set of vectors $\{u, v, w\} \subset \mathbb{R}^3$ is linearly independent (since they don't lie on the same 2-dimensional subspace of $\mathbb{R}^3$. Then we know there is an invertible 3×3

matrix A so that $Ae_1 = u$, $Ae_2 = v$ and $Ae_3 = 2$ (recall the matrix is given by taking the columns to be the vectors, that is, $A = \begin{bmatrix} u & v & w \end{bmatrix}$).

Therefore the projective transformation τ given by $\tau([v]) = [Av]$ has the property that $\tau([e_1]) = [u]$, $\tau([e_2]) = [v]$ and $\tau([e_3]) = [w]$.

Now, given two sets of three non-collinear points in $\mathbb{RP}^2$, there are projective transformations taking each set to the set $\{[e_1], [e_2], [e_3]\}$ and therefore both sets are projective-congruent.

On Π this corresponds to the statement that any two sets of vertices of triangles on the plane are projective-congruent.

Example. Let's make the previous example even more specific. Let $P = [1, 2, 3]$, $Q = [4, -1, 0]$, $R = [1, 1, 1]$ in $\mathbb{RP}^2$. Let

$$A = \begin{bmatrix} 1 & 4 & 1 \\ 2 & -1 & 1 \\ 3 & 0 & 1 \end{bmatrix}.$$

Then $\tau([v]) = [Av]$ has the property that $\tau([e_1]) = P$, $\tau([e_2]) = Q$, and $\tau([e_3]) = R$. However, this is not the only such projective transformation. There are infinitely many! To see this, we can scale each column by our favourite non-zero real number. For example, if

$$B = \begin{bmatrix} 2 & 4 & -1 \\ 4 & -1 & -1 \\ 6 & 0 & -1 \end{bmatrix},$$

then $\mu([v]) = [Bv]$ has the property that $\mu([e_1]) = P$, $\mu([e_2]) = Q$, and $\mu([e_3]) = R$. Furthermore, you can check (and you should!) that $\tau([1, 1, 1]) \neq \mu([1, 1, 1])$ so these two projective transformations are distinct.

What the previous example shows us is that three points is not enough to determine a projective transformation. We will now see that four points will be exaclty the right amount of information to determine a projective transformation.

Lecture 35 and beyond!

Theorem 46 (The fundamental theorem of projective geometry)**.** *Let $a = (a_1, a_2, a_3)$, $b = (b_1, b_2, b_3)$, $c = (c_1, c_2, c_3)$, $d = (d_1, d_2, d_3)$ be vectors in $\mathbb{R}^3$ so that no three of $[a], [b], [c], [d] \in \mathbb{RP}^2$ are collinear. Then there exists a unique projective transformation $\tau \in \mathrm{PGL}_3(\mathbb{R})$ so that $\tau([e_1]) = [a]$, $\tau([e_2]) = [b]$, $\tau([e_3]) = [c]$, and $\tau([1, 1, 1]) = [d]$.*

Proof. For non-zero real numbers α, β, γ, consider the matrix

$$A = \begin{bmatrix} \alpha a_1 & \beta b_1 & \gamma c_1 \\ \alpha a_2 & \beta b_2 & \gamma c_2 \\ \alpha a_3 & \beta b_3 & \gamma c_3 \end{bmatrix}.$$

Note that $Ae_1 = \alpha a$, $Ae_2 = \beta b$ and $Ae_3 = \gamma c$. Therefore, if $\tau \in \mathrm{PGL}_3(\mathbb{R})$ is defined by A, then $\tau([e_1]) = [a]$, $\tau([e_2]) = [b]$, and $\tau([e_3]) = [c]$. So, to show that our desired projective transformation exists, we need to show there are non-zero real numbers α, β, γ so that $\tau([1, 1, 1]) = [d]$.

To that end, we want

$$\begin{bmatrix} \alpha a_1 & \beta b_1 & \gamma c_1 \\ \alpha a_2 & \beta b_2 & \gamma c_2 \\ \alpha a_3 & \beta b_3 & \gamma c_3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix}.$$

Expanding this out gives the following linear system of equations, in the variables α, β, γ:

$$\begin{aligned} a_1\alpha + b_1\beta + c_1\gamma &= d_1 \\ a_2\alpha + b_2\beta + c_2\gamma &= d_2 \\ a_3\alpha + b_3\beta + c_3\gamma &= d_3. \end{aligned}$$

This corresponds to solving the matrix equation

$$\begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix}.$$

Since $[a], [b], [c]$ are not collinear, the coefficient matrix is invertible. Therefore there are real numbers α, β, γ so that $A \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = d$. The fact that α, β, γ are all non-zero follows from the assumption that no three of $[a], [b], [c], [d]$ are collinear.

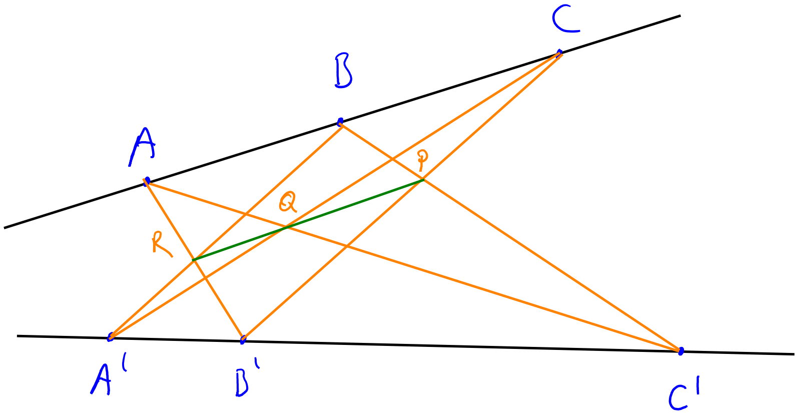
Uniqueness of τ is left as an exercise. ■

Exercise. 1. Suppose $\tau \in \mathrm{PGL}_3(\mathbb{R})$ satisfies $\tau([e_1]) = [e_1]$, $\tau([e_2]) = [e_2]$, $\tau([e_3]) = [e_3]$, and $\tau([1,1,1]) = [1,1,1]$. Prove that τ is the identity map on $\mathbb{RP}^3$.

2. Prove that the element of $\mathrm{PGL}_3(\mathbb{R})$ constructted in the proof of the fundamental theorem of projective geometry is unique.

We are now ready to benefit from some of the machinery built up in the world of projective geometry. Any theorem in Euclidean geometry that is about lines intersecting, or points lying on a line, as the potential to be proved using projective geometry. The statement of the next theorem (Pappus's hexagon theorem) takes place entirely in $\mathbb{R}^2$, but the proof will take place entirely in $\mathbb{RP}^2$.

Theorem 47 (Pappus's hexagon theorem)**.** *Let $A, B, C, A', B', C' \in \mathbb{R}^2$ be six distinct points so that A, B, C lie on a line, and A', B', C' lie on a different line. Assume further that none of A, B, C, A', B', C' are the intersection point between the two lines. Suppose $\overline{BC'}$ and $\overline{B'C}$ intersect at P, $\overline{AC'}$ and $\overline{A'C}$ intersect at Q, and $\overline{AB'}$ and $\overline{A'B}$ intersect at R. Then P, Q, R lie on a line.*



Proof. In this case, we will identify $\mathbb{R}^2$ with the plane T in $\mathbb{R}^3$ defining some embedding plane Π. First we have an important observation that allows us to transfer statements about T to statements about $\mathbb{RP}^2$. Notice that two points in T are collinear (in T) if and only if the corresponding points in $\mathbb{RP}^2$ are collinear in $\mathbb{RP}^2$. This is because projective lines are planes in $\mathbb{R}^3$ passing through the origin, and the intersection of such a plane with T is a line on T.

So, with that in mind, we will treat the nine points in the statement of the theorem as points in $\mathbb{RP}^2$, and it will suffice to show that the points $P, Q, R \in \mathbb{RP}^2$ are collinear.

The assumptions in the statement of the theorem imply that the four points A, A', R, and P in $\mathbb{RP}^2$ are such that no three are collinear (you should justify this!). Therefore by the fundamental theorem, there is a $\tau \in \mathbb{RP}^2$ so that $\tau(A) = [1,0,0]$, $\tau(A') = [0,1,0]$, $\tau(R) = [1,1,1]$, and $\tau(P) = [0,0,1]$.

Now, let's figure out what the projective coordinates of the other five points should be. Since $\tau(B')$ lies on the projective line passing through $\tau(A) = [1,0,0]$ and $\tau(R) = [1,1,1]$, $\tau(B') = [s+t,t,t]$ for some real numbers s and t. Since $\tau(B') \neq \tau(A)$, we have that $t \neq 0$. Therefore we have $\tau(B') = [r,1,1]$ for some $r \in \mathbb{R}$.

The point $\tau(B)$ is on the projective line passing through $\tau(A') = [0,1,0]$ and $\tau(R) = [1,1,1]$. Similarly, to $\tau(B')$, we can conclude that $\tau(B) = [1,s,1]$ for some $s \in \mathbb{R}$.

We have $\tau(C)$ on the line passing through $\tau(A) = [1,0,0]$ and $\tau(B) = [1,s,1]$, and also on the line passing through $\tau(B') = [r,1,1]$ and $\tau(P) = [0,0,1]$. The first line is defined by a plane with normal vector $(1,0,0) \times (1,s,1) = (0,-1,s)$. The second line is defined by a plane with normal vector $(1,-r,0)$. The intersection of these two planes is given by the line in $\mathbb{R}^3$ through the origin and the vector $(0,-1,s) \times (1,-r,0) = (rs,s,1)$. Therefore $\tau(C) = [rs,s,1]$. A similar argument gives $\tau(C') = [r,rs,1]$.

Finally, $\tau(Q)$ is the intersection of the projective line l_1 passing through $\tau(A) = [1,0,0]$ and $\tau(C') = [r,rs,1]$, and the projective line l_2 passing through $\tau(A') = [0,1,0]$ and $\tau(C) = [rs,s,1]$. The plane in $\mathbb{R}^3$ defining l_1 has normal vector $(1,0,0) \times (r,rs,1) = (0,-1,rs)$. The plane in $\mathbb{R}^3$ defining l_2 has normal vector $(0,1,0) \times (rs,s,1) = (1,0,-rs)$. The intersection of l_1 and l_2 is therefore represented by the line in $\mathbb{R}^3$ passing through $(0,-1,rs) \times (1,0,-rs) = (rs,rs,1)$. Therefore, $\tau(Q) = [rs,rs,1]$.

Now, $\tau(P) = [0,0,1]$, $\tau(Q) = [rs,rs,1]$, $\tau(R) = [1,1,1]$. These three points lie on the projective line defined by $x = y$. Applying the projective transformation τ^{-1} allows us to conclude that P, Q, R lie on the same projective line. Alas, we may conclude that in the embedding plane Π, P, Q, R are collinear. ■

The take-home message here is that we can use the fundamental theorem to change coordinates in a sense, and turn our problem into one that's much more tractable computationally. There were lots of ones and zeros in the computations in the proof of Pappus's hexagon theorem, and it was easy to check at the end that $\tau(P)$, $\tau(Q)$, and $\tau(R)$ were collinear.

The advantage of a proof like this that uses projective geometry is that it actually proves a slightly more general result. For example, the proof takes into account the case where, for example, $A'B$ and AB' are parallel (so they intersect at an ideal point of the embedding plane Π).

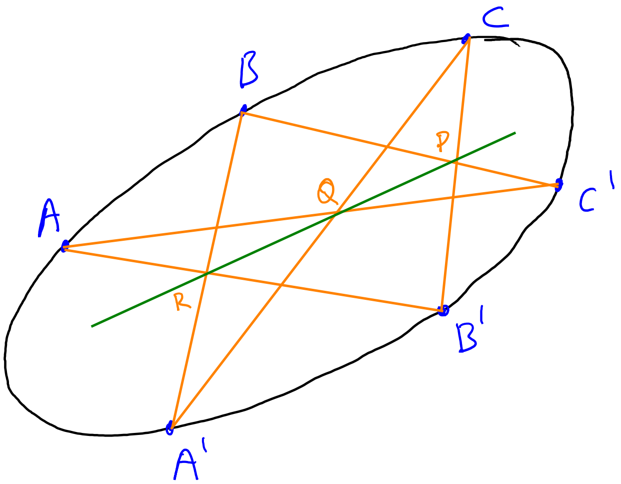
Exercise. What is the statement in Euclidean geometry that arises when $A'B$ and AB' in the statement of Pappus' theorem are parallel?

What is the statement in Euclidean geometry that arises when C is taken to be the ideal point for the line passing through A and B?

Lecture 36 - 30/07

It turns out that Pappus's hexagon theorem is a special case of something more general, called Pascal's theorem, which we shall not prove here. In what follows, a conic section is any of an ellipse, hyperbola, or parabola in $\mathbb{R}^2$.

Theorem 48 (Pascal's theorem)**.** *Let A, B, C, A', B', C' be six distinct points on a conic section. Let $\overline{BC'}$ and $\overline{B'C}$ intersect at P. Let $\overline{AC'}$ and $\overline{A'C}$ intersect at Q. Let $\overline{AB'}$ and $\overline{A'B}$ intersect at R. Then P, Q, R are collinear.*



It turns out that the union of two lines is a special example of a conic section, called a **degenerate conic section**. Pascal's theorem can be proved by first proving that every conic section is projective-congruent to the projective conic $C = \{[x, y, z] \in \mathbb{RP}^2 : xy + xz + yz = 0\}$, and then proving the result for six points on C in $\mathbb{RP}^2$. That however, is a story for another time.

5.4 Duality

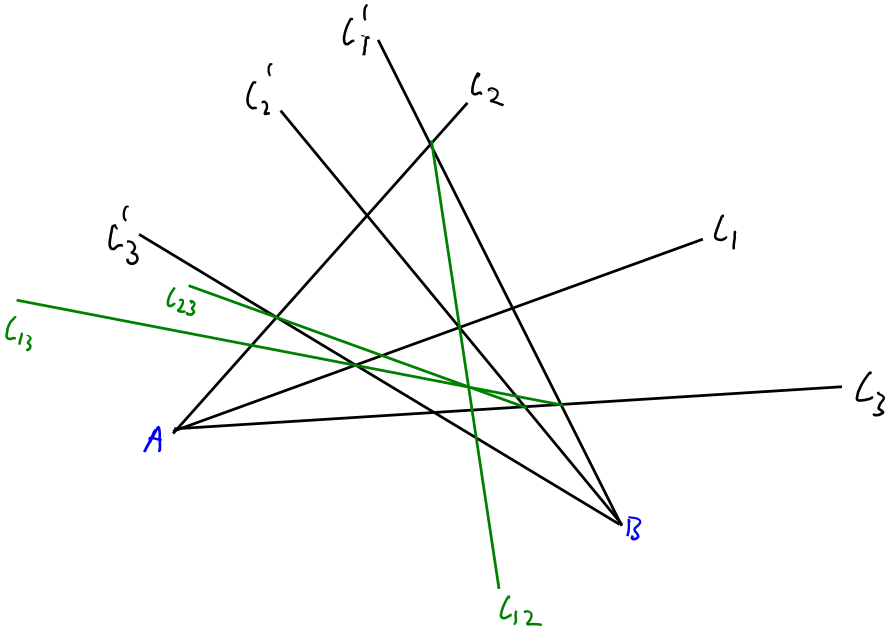
Towards the beginning of the projective geometry part of the course, we saw that lines and points act kind of similarly to each other. One example of this is that two points lie on a unique line, and every two lines intersect at a unique point. In fact, we can interchange the roles of points and lines in $\mathbb{RP}^2$ in such a way that the property of two points being collinear is exchanged with the property that two lines intersect.

Exercise. Let $\mathcal{L}$ be the set of lines in $\mathbb{RP}^2$. Let $\phi : \mathcal{L} \to \mathbb{RP}^2$ be the function defined by $\phi(\{[x, y, z] \in \mathbb{RP}^2 : ax + by + cz = 0\}) = [a, b, c]$. Prove that ϕ is a bijection with the property that for $l_1, l_2 \in \mathcal{L}$, l_1 and l_2 intersect at $P \in \mathbb{RP}^2$ if and only if $\phi(l_1)$ and $\phi(l_2)$ lie on the line $\phi^{-1}(P)$.

With this in mind, for every theorem in projective geometry that only deals with lines, points, which lines intersect, and which points are collinear, there is likely to be a dual theorem. For example, the dual to the statement "there is a unique line passing through any two points" is "every two lines intersect at a unique point." Here is the dual statement to Pappus's hexagon theorem.

Theorem 49. *Let l_1, l_2, l_3 be three lines in $\mathbb{R}^2$ that all intersect at a point A. Let l'_1, l'_2, l'_3 be three other lines in $\mathbb{R}^2$, none of which contain A, that all intersect at B. Assume that none of the six lines are parallel. Let l_{12} be the line passing through the intersection of l_1 and l'_2, and the intersection of l'_1 and l_2. Let l_{13} be the line passing through the intersection of l_1 and l'_3, and the intersection of l'_1*

and l_3. Let l_{23} be the line passing through the intersection of l_2 and l_3', and the intersection of l_2' and l_3. Then l_{12}, l_{23}, and l_{13}, all intersect at a common point.



Proof. This is an exercise. Either use the previous exercise combined with Pappus's hexagon theorem, or prove it directly! ■

A Crash course in complex cnumbers

The **complex numbers** are the set $\mathbb{C} = \{a + bi : a, b \in \mathbb{R}\}$. The complex numbers come with addition and multiplication, which is just usual addition and multiplication, except where you see i^2 you replace it with -1. For example,

$$(2 + 3i) + (1 - i) = 3 - 2i \quad \text{and} \quad (2 + 3i)(1 - i) = 2 + 3i - 2i - 3i^2 = 5 + i.$$

In general we have

$$(a + bi) + (c + di) = (a + c) + (b + d)i \quad \text{and} \quad (a + bi)(c + di) = (ac - bd) + (ad + bc)i.$$

The **real** and **imaginary** parts of $z = a + bi$ are $\operatorname{Re}(z) = a$ and $\operatorname{Im}(z) = b$. The **conjugate** of $z = a + bi$ is $\overline{z} = a - bi$. Conjugation plays nicely with addition and multiplication, and you can check that $\overline{zw} = (\overline{z})(\overline{w})$ and $\overline{z + w} = \overline{z} + \overline{w}$.

To find the inverse of $z = a + bi \neq 0$ we compute

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

The quantity $a^2 + b^2$ associated to $a + bi$ is an important one, and its square root is called the **modulus** of z, denoted $|z|$. So, if $z = a + bi$, its modulus is given by

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

Putting things together so far we have that for $z \neq 0$, $z^{-1} = \frac{\overline{z}}{|z|}$.

A.1 The complex plane

Just like we can put the real numbers on the number line, we can plot out the complex numbers in $\mathbb{R}^2$, and we call this the **complex plane**. It's just like $\mathbb{R}^2$ except the x-coordinate contains the real part of the complex number, and the y-coordinate contains the imaginary part. So, for example, $1 + 2i$ would be plotted at the point $(1, 2)$. The complex number 0 is at the origin, and i and $-i$ both lie on the y-axis. The x-axis is the familiar real number line.

[IMAGE HERE]

We can geometrically intepret some of the things that came up at the beginning of this appendix. For $z \in \mathbb{C}$, the conjugate $\overline{z}$ is the point in the complex plane obtained by reflecting z across the x-axis. The modulus $|z|$ is the distance in the complex plane that z is away from the origin.

B Injections, Surjections, and Bijections

Intuitively we know the definitions of an injection, surjection and bijection. An injection from S to T is a function that doesn't send any two elements of S to the same element of T. A surjection from S to T is a function that sends something to everything in T, or a function that hits everything in T. A bijection is a perfect matching, kind of like a dictionary, between elements of S and elements of T. That is, every element of S has an element of T associated to it, and vice versa. This is the same as saying that f is both surjective and injective. Let's make these intuitions formal.

Definition. Let $f : S \to T$ be a function.

- We say f is **injective** (or f is an **injection**) if whenever $f(s_1) = f(s_2)$, we have $s_1 = s_2$.

- We say f is **surjective** (or f is a **surjection**) if for all $t \in T$, there exists an $s \in S$ such that $f(s) = t$.
- We say f is **bijective** (or f is a **bijection**) if f is both injective and surjective.

This definition is all well and good, but there is another way to think about injections, surjections, and bijections. The idea is as follows.

If $f : S \to T$ is an injection, then every element in T that gets hit has a unique preimage (a unique element $s \in S$ such that $f(s) = t$) so we can define a $g : T \to S$ such that if we do f first and then g, we can return every element in S to itself.

If $f : S \to T$ is a surjection, then every $t \in T$ has at least one preimage, so we can define $g : T \to S$ to be a function that sends t to one of its preimages. Since every t has a preimage, this function has the property that if we do g first and then f, every element of t ends up back where it started.

If $f : S \to T$ is a bijection, then we can do what we did for the injections and surjections in a unique way to get a $g : T \to S$ such that $fg(t) = t$ for all $t \in T$ and $gf(s) = s$ for all $s \in S$.

These ideas are formalised in the following proposition. Here is a bit of notation we will use for the proposition and throughout the notes above. Let S be a set and define the identity function $\mathrm{id}_S : S \to S$ by $\mathrm{id}_S(s) = s$ for all $s \in S$.

Proposition 50. *Let $f : S \to T$ be a function between sets.*

- *f is an injection if and only if there exists a function $g : T \to S$ such that $gf = id_S$.*
- *f is a surjection if and only if there exists a function $g : T \to S$ such that $fg = id_T$.*
- *f is a bijection if and only if there exists a function $g : T \to S$ such that $fg = id_T$ and $gf = id_S$. Furthermore, such a g is unique and we denote it $g = f^{-1}$.*

Proof. Suppose f is an injection. Pick an $x \in S$ and for every $t \in f(S)$, let $s_t \in S$ be the unique element of S such that $f(s_t) = t$. Recall $f(S) := \{t \in T : \text{there exists } s \in S \text{ such that } f(s) = t\}$. Define $g : T \to S$ by

$$g(t) = \begin{cases} s_t & \text{if } t \in f(S) \\ x & \text{otherwise.} \end{cases}$$

Since every $s \in S$ is of the form s_t for some $t \in T$, we see $gf(s_t) = g(t) = s_t$ for all $s_t \in S$ so $gf = \mathrm{id}_S$.

Conversely suppose $f(a) = f(b) = t_0$ where $a \neq b$ in S. Suppose $g : T \to S$ is such that $gf = \mathrm{id}_S$. If $g(t_0) \neq a$, then $gf(a) \neq a$, so we must have $g(t_0) = a$. Then we have $gf(b) = a \neq b$, so such a g cannot exist.

Suppose f is a surjection. Define $g : T \to S$ by $g(t) = s_t$ where $f(s_t) = t$. Note that since f is surjective, we can always do this. Then $fg(t) = f(s_t) = t$ for all $t \in T$, so $fg = \mathrm{id}_T$.

Conversely, if f is anot a surjection there is some $t_0 \in T$ such that there is no $s \in S$ such that $f(s) = t_0$. Let $g : T \to S$ be a candidate function such that $fg = \mathrm{id}_T$. Then $fg(t_0) \neq t_0$ since there is no element in S such that $f(s) = t_0$. Therefore there is no function $g : T \to S$ such that $fg = \mathrm{id}_T$.

Finally, if f is a bijection, then define $g : T \to S$ to be $g(t) = s_t$ where $s_t \in S$ is the unique element such that $f(s_t) = t$. Note that every element in S is of the form s_t for some t. Then

$$gf(s_t) = g(t) = s_t \quad \text{and} \quad fg(t) = f(s_t) = t$$

for all $s_t \in S$ and $t \in T$, so $gf = \mathrm{id}_S$ and $fg = \mathrm{id}_T$. Conversely, if f is not injective or surjective, the same arguments above show that there cannot exist a $g : T \to S$ such that $fg = \mathrm{id}_S$ or $gf = \mathrm{id}_T$ respectively.

It remains to show that in the case when f is a bijection, the inverse g is unique. Suppose there is another map $h : T \to S$ such that $fh = \mathrm{id}_S$. Then $f(h(t)) = t = f(g(t))$ fur all $t \in T$. Since f is injective, we must have $h(t) = g(t)$ for all $t \in T$, completing the proof. ■

Whenever you have an if and only if statement, you can use either property as the definition in your head. For example, we can now think of an injection has a map with a left inverse, or as a map which sends different elements to different elements. Whichever definition is easier or more helpful in a particular situation should be the one you use.

It is worth noting here that the above proof relies on the axiom of choice, but that is a discussion for another time and course.

Let's look at the main thing we have studied in the course, circle inversions!

Example. Let $I_C : \mathbb{C} \cup \{\infty\} \to \mathbb{C} \cup \{\infty\}$ be a circle inversion about some circle C in $\mathbb{C}$. We know that $I_C I_C(p) = p$ for all $p \in \mathbb{C} \cup \{\infty\}$. This exactly tells us that I_C is a bijection with $I_C^{-1} = I_C$. Neat!

In the hyperbolic geometry section of the course, we focus a lot on hyperbolic transformations.

Exercise. Let $\sigma : \mathbb{D} \to \mathbb{D}$ be a hyperbolic reflection. Prove that σ is a bijection with $\sigma^{-1} = \sigma$.

Example. Even better, we can compose hyperbolic reflections to get hyperbolic transformations. For example, suppose $\tau \in \mathcal{G}_{\mathbb{D}}$ is $\tau = \sigma_1\sigma_2$ where $\sigma_1\sigma_2$ are hyperbolic reflections. Then for all $x \in \mathbb{D}$ we have

$$\begin{aligned}\sigma_1\sigma_2\sigma_2^{-1}\sigma_1^{-1}(x) &= \sigma_1\sigma_2\sigma_2(\sigma_1(x))\\ &= \sigma_1\sigma_1(x)\\ &= 1\end{aligned}$$

and similarly,

$$\sigma_2^{-1}\sigma_1^{-1}\sigma_1\sigma_2(x) = \sigma_2^{-1}\sigma_2(x) = x$$

for all $x \in \mathbb{D}$. Therefore $\tau^{-1} = (\sigma_1\sigma_2)^{-1} = \sigma_2^{-1}\sigma_1^{-1} = \sigma_2\sigma_1$. In fact we can generalise this argument to a composition of any finite number of hyperbolic reflections to get that every hyperbolic transformation is a bijection from the hyperbolic plane to itself!

The example with the hyperbolic transformation hints at a more general statement about bijections, which is that you can compose them to get more bijections!

Proposition 51. *Let $f : X \to Y$ and $g : Y \to Z$ be bijections. Then gf is a bijection.*

Proof. We have $(gf)^{-1} = f^{-1}g^{-1}$, so gf is a bijection (you should fill in the details here). ■