GEOMETRY

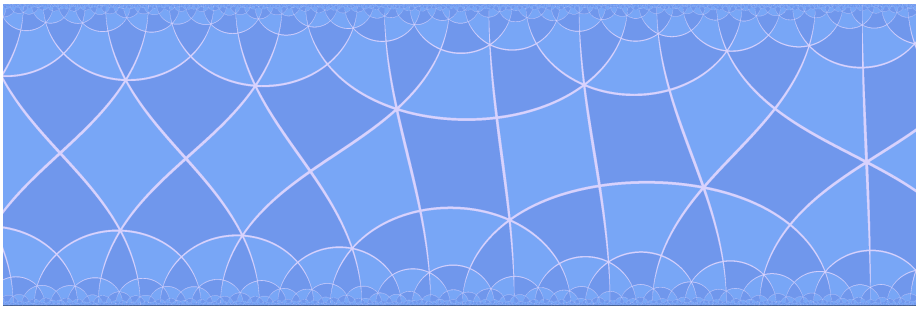
A MODERN VIEW

Steve Trettel

Table of contents

Dedication

This is the first draft of a future textbook in modern geometry at the advanced undergraduate level. It was used as the course text for the course *Foundations of Geometry* at USF in the Fall of 2023.



This very first edition is dedicated the students of that course: Alana, Daniel, Emelia, Frances, Katie & Quinn.

Thank you for your patience as I wrote these notes live over the course of the semester, and your enthusiasm which motivated me to continue writing each evening, even when I began to realize I committed to a much bigger project than I should have for the term.

Part I.

The Greeks

1. Euclid

Geometry began deep in antiquity, arising from our need to measure properties of the physical world as we came out of the last ice age and began to collect ourselves into cities. Much of the subject's long prehistory is lost to the depths of preliterate time, but signs of the ancients knowledge remain in their surviving artworks.

Figure 1.1.: An Archimedean spiral on Neolithic Pottery, Romania 7000BCE

With writing and the ensuing civilization growth, recognizably modern geometry was practiced by bronze-age peoples the world over, with notable examples in Babylon, Egypt and China.

Figure 1.2.: A collection of Babylonian homework exercises, written in Akkadian, 1700BCE

Geometry at this stage in our history was more a compendium of true facts about space, than it was a coherent theory of *how* space behaved. One could derive new facts from known facts, or see that new observations were consistent with existing knowledge, but early mathematicians had not yet found an order to this chaos.

1.1. Euclid's Postulates

By the time Euclid was born in the 300s BCE, Greek civilization was reaching the twilight of its golden age, and serious geometry had been practiced by its mathematician-philosophers for centuries. While not much is known of Euclid's personal life, his fame survives into modernity as the author of the three book series *The Elements*, collecting and systematizing much of the Greek's knowledge of geometry. Much of its content originates from earlier mathematicians, including Eudoxus, Hippocrates of Chios, Thales and Theaetetus, while other theorems in *The Elements* are previously mentioned by Plato and Aristotle.

But this is not to take anything away from Euclid's incredible achievement: while the *theorems* were not all original, it is his *exposition* that became a model for all mathematics written over the following twenty centuries. Euclid took the mass of knowlege humans had discovered, and built it into a logical, self-contained, and understandable body of knowledge. He turned a collection of truths into a *mathematical theory*.

Euclid reduced the entirety of Greek geometry to the logical consequences of just five statements. Five! Everything that had ever been measured about the nature of space sprung forth from pure logic and five basic truths. These truths, which Euclid called his Αξιώματα (Axiomata), we usually call *Euclid's Postulates* in English.

Definition 1.1 (Αξιώματα του Ευκλείδη: Euclid's Postulates)**.**

- Η κατασκευή μιας ευθείας γραμμής από ένα σημείο σε οποιοδήποτε άλλο
- Μια πεπερασμένη ευθεία μπορεί να επεκταθεί απεριόριστα
- Ένας κύκλος ορίζεται από ένα κέντρο και μια απόσταση(ακτίνα)
- Όλες οι ορθές γωνίες είναι ίσες
- Έστω δύο ευθείες που τέμνονται με μια τρίτη. Οι ευθείες αυτές θα έχουν ένα σημείο τομής από την μεριά που οι εσωτερικές γωνίες που σχηματίζονται με την τρίτη ευθεία έχουν άθροισμα μικρότερο από δύο ορθές γωνίες.

In English:

- A straight line segment can be drawn joining any two points.
- Any straight line segment can be extended indefinitely in a straight line.
- Given any straight line segment, a circle can be drawn having the segment as radius and one endpoint as center.
- All right angles are congruent.

- If two lines are drawn which intersect a third in such a way that the sum of the inner angles on one side is less than two right angles, then the two lines inevitably must intersect each other on that side if extended far enough.

These five statements were chosen to be as directly observable and intuitive as possible: and for the most part they do an excellent job at that. Drawing a line between any two points? Sounds reasonable. Making a line longer? Also reasonable. Rotating any line segment around to make a circle? Alright. Any two right angles being congruent? Of course.

But the fifth one, that one you have to read a couple times and draw a picture before you're convinced of its truth. We will have much more to say about this 5th postulate in the next chapter.

1.2. The Idea of Proof

The real power of Euclid's postulates comes from the ability to *prove* things from them.

Definition 1.2 (Proof)**.** A proof of a statement S is a sequence of logical claims, starting from a collection of foundational statements (like a list of axioms and previously proven statments) and ending with the statement S that you wanted to verify is true.

As a quick example, if we take as foundational the definition of even number being a integer that is 2 times another integer and the rules of arithmetic, we can prove that for any integer x, the integer $x + x$ is even. Our proof goes as follows:

- We can use the fact that $1x = x$ to rewrite $x + x$ as $1x + 1x$.
- We can factor out the x to get $1x + 1x = (1 + 1)x$
- We can use that $1 + 1 = 2$ to get $(1 + 1)x = 2x$
- $2x$ is twice the integer x, so it is even.
- Thus, since $x + x = 2x$, we see $x + x$ is even.

This proof is not particularly *exciting*, but it is *clear* and *unambiguously true.* Each bit of the proof made just a small step at a time, and explained why each step held using the foundational material. *Anyone* who understands the foundational material and knows how to read would be convinced by this argument that $x + x$ must be even.

Arguments in greek geometry are exactly of the same style, except we replace the rules of arithmetic with the Postulates of Euclid, and definitions like "evenness" with geometric definitions like "triangle".

Definition 1.3 (Triangle)**.** A triangle is a figure in the plane composed of three points p, q, r which do not all lie on a common line. The *sides* of the triangle are the line segments pq, pr and qr. The *angles* of the triangle are defined by these sides at the points p, q, r themselves.

Indeed, given just this definition and the five postulates of Euclid, we can follow him in proving his first Proposition

Theorem 1.1 (Equilateral Triangles Exist: Elements Prop I)**.** *There exists a triangle in the plane all of whose sides are the same length.*

Proof.

- Choose two points p, q in the plane. Draw the line segment between them, of length L. (Postulate 1)
- Now form the circle of radius L centered at p, and the circle of radius L centered at q (Postulate 3)
- These two circles intersect in two points. Select one of the intersection points, call it r.
- Use postulate 1 again to draw a line from p to r and from q to r. Together with the original line, these form a triangle.
- The line segment from p to r is length L, as is the length of the segment from q to r as they are both radii of circles. But the distance from p to q was L too: so this triangle has three sides of length L.

□

Remark 1.1. That the circles intersect seems intuitively obvious, but how would one actually prove this? This is the beginning hints that defining things in terms of calculus will prove useful. From our modern perspective, this seems to have something to do with the *continuity* of the circle, and perhaps the *intermediate value theorem.*

A video demonstration of this proof is below:

https://youtu.be/zdofiH5HncU

Propositions, themselves being verified from the list of known facts, are then "legal" to be used in the justification of future facts. Euclid is very intentional in his development of the subject, and makes sure that every proposition only uses in its justification facts that have been previously proven.

1.3. Absolute Geometry

Euclid delays using Postulate 5 as long as possible, and proves the first 28 propositions of Book I using only Postulates 1-4. (Indeed, in Proposition I we used only Postulates 1 and 3!) These days, we call such results theorems of *absolute geometry*.

Definition 1.4 (Absolute geometry)**.** Absolute geometry is the set of theorems which can be proven using *only* Euclid's postulates 1 through 4.

Below we embark on a brief, and incomplete tour of Euclid's work in absolute geometry to get better acquainted with the Greek notion of geometric proof.

The early propositions focus mostly on increasing our toolkit: they prove that its possible to do certain useful geometric constructions from the axioms, so that we can in the future use these directly where convenient.

Proposition 1.1 (Copying a Line Segment: Elements Prop 2)**.** *Given a segment L and a point p not on that line, its possible to draw a new line segment starting at p whose length is the same as L.*

The proof of this proposition uses only Proposition I together with the 5 postulates, and so must be quite ingenious: there isn't much to work with! Indeed, its best understood through an animation, so there is a youtube video below.

https://youtu.be/aBCkBJoXMlo

Proposition 1.2 (Cutting a Line Segment to Size: Elements Prop 3)**.** *Given two line segments of unequal lengths, its possible to cut the longer line segment so that the remaining piece has the same length as the shorter.*

Proof. Start with a line segment AB, and another line segment CD, and without loss of generality let's say that CD is the longer segment.

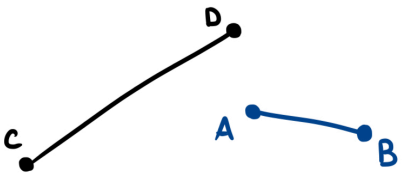
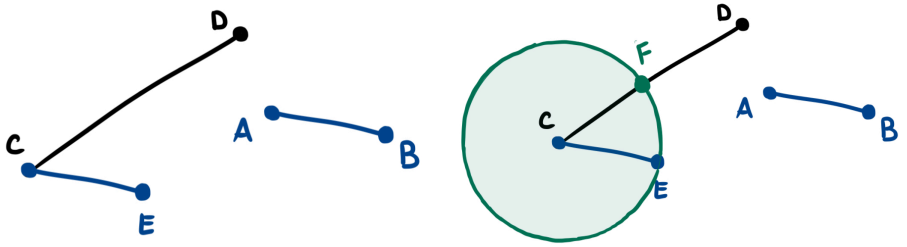


Figure 1.3.: .

- Use Proposition 2 to build a segment with the same length as AB but now starting at the point C.



- Now use Postulate 3 to spin this new line around to form a circle centered at C.

- This new line intersects CD in a new point, call it F.
- Now the line CF is a radius of the circle, and so has the same length as the original copied segment: the length of AB.

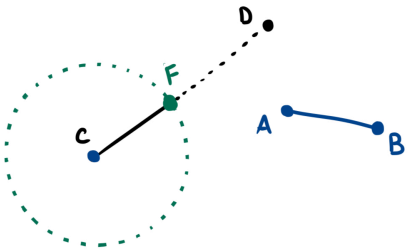


Figure 1.4.: .

□

Exercise 1.1 (Constructing an Isoceles Triangle)**.** Start with a line segment of length a. Prove that you can construct a triangle with one side of length a, and two sides of length $2a$, using the postulates 1-5 and the propositons 1-3 given so far.

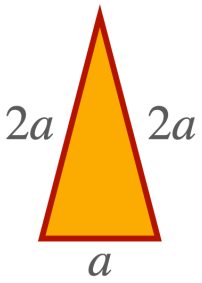


Figure 1.5.: An isosceles triangle with two sides double the other.

Now that we can copy and cut line segments, Euclid is ready to prove his first theorem about telling when two shapes are the same (or *congruent*). You may recall this as *side-angle-side* congruence from elementary geometry: Euclid proves it by using what we just proved to copy one triangle on top the other, to see they are equal.

Theorem 1.2 (Side-Angle-Side Congruence: Elements Prop 4)**.** *If two triangles share a pair of sides with the same lengths, and those sides form an angle of equal measure on each, then the two triangles are congruent.*

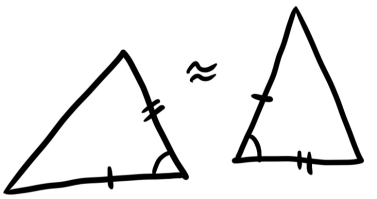


Figure 1.6.: Side-Angle-Side Congruence: sides with the same length and angles with the same measure are marked alike on the two triangles.

https://youtu.be/sk2dL_kitcE?si=IdB32dJQqdSneuuz

Euclid continues on this way for some time, proving more theorems theorems about triangles, including side-side-side congruence.

Theorem 1.3 (Side-Side-Side Congruence: Elements Prop 8)**.** *If two triangles have corresponding sides of the same three lengths, then the two triangles are congruent.*

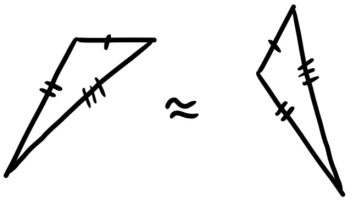


Figure 1.7.: Side-Side-Side Congruence: sides with the same length are marked alike on the two triangles.

As a quick application we use this to show that the angles of an equilateral triangle are all equal to one another.

Proposition 1.3. *The three angles of an equilateral triangle are equal.*

Proof. Let ABC be an equilateral triangle. Then as all of its sides are the same length, it is side-side-side congruent to any rotated copy of itself. Concretely, we see that ABC is congruent to BCA.

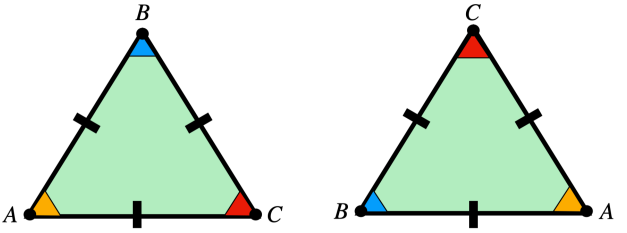


Figure 1.8.: Equilateral triangles have three equal angles.

This sets up an equality between the angles:

$$A = B \qquad B = C \qquad C = A$$

Thus all the angles are equal to one another. □

These congruences are both theorems of absolute geometry, meaning that they are true in *any* world where Postulates 1-4 hold. Being able to tell when two triangles are the same gives Euclid a new power - to verify that one can cut an angle precisely in half.

Proposition 1.4 (Bisecting an Angle: Elements Prop 9)**.** *If θ is any angle, it is possible to draw a line dividing it into two angles each of measure $\theta/2$.*

Before watching the video on this one (or looking back at your notes) try to draw the diagram from the instructions!

Proof.

- Start with an arbitrary angle at a point A, and choose a point along D one of the angle's rays.
- Use the segment from A to D, with Postulate 3, to construct a circle centered at A.
- This circle intersects the angles other ray at some third point, E
- Use Proposition 1 to construct an equilateral triangle on the segment DE, which goes across the angle.
- Call the vertex of the equilateral triangle F. Now use Postulate 1 to draw a line from the angle's vertex A to F.
- This creates two triangles, using the new line AF, and then using one side of the equilateral triangle, and one side ray of the original angle.
- These two triangles, AEF and ADF have all three pairs of sides the same length: thus, by Side-Side-Side congruence, they are equal.
- This means their angles are also equal. So the two angles we have split the original angle at A into are equal, and so each must be half the original angle's measure.

□

This proof is a bit involved too - so it may be helpful to watch a video for future reference!

https://youtu.be/HUv0I96vH34

Applying this to a straight angle allows one to bisect this into two equal angles. Since half a straight angle is a right angle, a corollary of this propositon is that it is possible to construct a right angle along any line segment.

Proposition 1.5 (Constructing Right Angle: Elements Prop 11)**.** *Given a line segment L and any point p along that line segment, it is possible to construct a perpendicular line T to L passing through p.*

At this point, we now (finally!) know that right angles must exist! Of course we had an axiom about right angles, but it did not tell us that there *were* any: it just said *IF* you had two right angles, then they are congruent. But it never gave you a means of making a right angle yourself! Now that we have one, we can do several interesting constructions: for example, we can prove that right triangles exist:

Proposition 1.6. *A right triangle exists can be created with any two leg lengths a, b.*

Proof. Begin with a line segment S_1 of length a, and another line of length b.

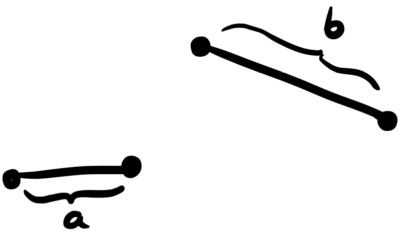


Figure 1.9.: Lines of the lengths we want as legs of the right triangle.

- Using Proposition 1.5, construct a segment at a right angle to S_1 at one of its endpoints, p.
- Use postulate 2 to extend this line segment indefinitely (in case the original segment you constructed was shorter than b!)

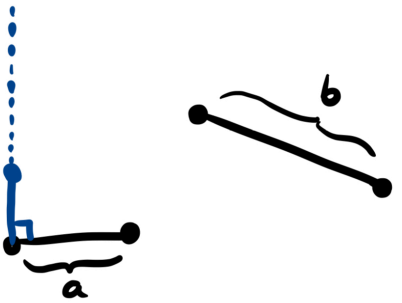
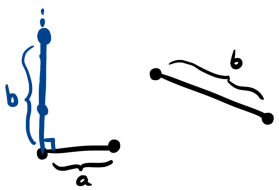


Figure 1.10.: Constructing a long perpendicular to one segment.

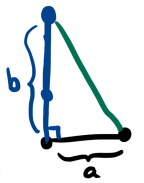
- Use Euclid's Proposition 3 (Cutting a Line Segment to Size) to trim this new segment until it is length b. Call the result S_2.
- Now, use Postulate 1 to connect the endpoints of S_1, S_2 by a straight line. Together, these three line segments form a triangle, with one right angle at p, and side lengths a, b as required.

□

Knowing that right triangles exist, its natural to ask next how we can tell when two right triangles are congruent. But we already have that tool: we can use Euclid's proposition asserting Side-Angle-Side congruence to conclude right triangles are congruent if and only if their legs are the same length.



(a) Cutting to size



(b) The final triangle

We've already learned quite a bit about triangles in absolute geometry, though we haven't quite exhausted the possible knowledge. Euclid goes on to prove a few more propositions, before reaching the final general congruence test for triangles in the 26th:

Proposition 1.7 (Angle-Side-Angle Congruence: Elements Prop 26)**.** *Two triangles are congruent if they have two equal pairs of angles, and an equal corresponding side.*

Again, we give a quick application of this triangle congruence, and complete the converse of Proposition 1.3.

Proposition 1.8 (Equilateral if Equiangular)**.** *Prove that a triangle with three equal angles has three equal sides. This proves that a triangle is equilateral if and only if it is equiangular.*

Proof. Let ABC be an equiangular triangle, so the angle measures at A, B and C are all equal.
Choose one of the angles - say B - and bisect it with a line. This line divides the triangle into two smaller triangles, which we see are congruent (they share a side: the new bisecting line, as well as two angles since they each have one of the original angles, and one of the bisected halves). Thus, the remaining pairs of sides of this triangle are also congruent, so BA equals BC.

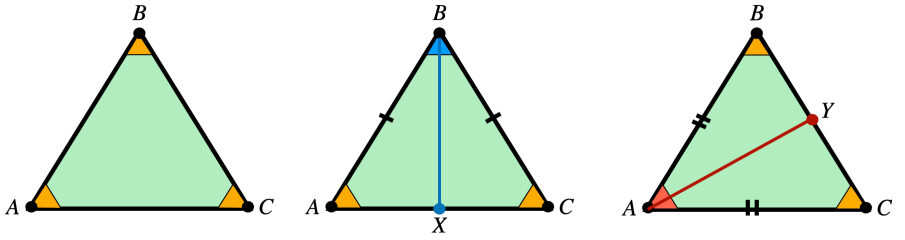


Figure 1.12.: Equilateral

There was nothing special about the angle B, so we may also do the same construction at another angle - say A. This again gives a pair of congruent triangles, from which we can conclude that AB equals AC.

Stringing these equalities together, we see that $AB = AC = BC$ so all three sides are equal, and the triangle is equilateral. □

Exercise 1.2. Prove that inside of an equilateral triangle, you can inscribe an upside down equilateral triangle of exactly half the side length, as in the figure below. In your proof, feel free to use any of the Postulates, as well as any proposition stated above this point.

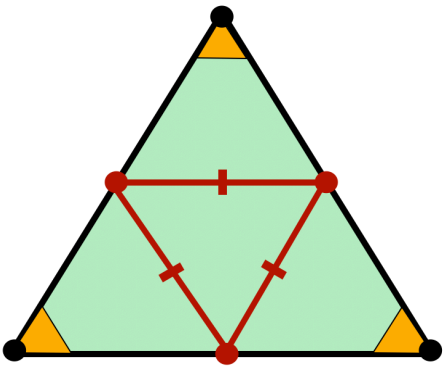


Figure 1.13.: Nested equilateral triangles.

2. Parallels

The fifth postulate of Euclid is often called the *parallel postulate*, as it gives a condition that can be checked for whether or not two lines are parallel.

Definition 2.1 (Parallel)**.** Two lines are parallel if they do not intersect.

Why the parallel postulate was suspicious to take as an axiom to the ancients.

One reason often cited: its complexity! Its a long statement.

But better reason, it implies the EXISTENCE of something (an intersection) at an arbitrary distance. The other postulates only assure the existence of things on scales that arleady show up in problem (given a segment, it can be extended finitely. Given a length, you can make a circle with it.)

2.0.1. Equivalents to Postulate 5

Definition 2.2 (Equivalence to Parallel Postulate)**.** A postulate P is equivalent to the parallel postulate if - P can be proven from postulates 1-5 - The combination of postulates 1-4 and P can prove Postulate 5.

One cleaner statement equivalent to Postulate 5 was already known to Proclus in antiquity, but became widely recognized after John Playfair's 1795 commentary on the Elements:

Definition 2.3 (Playfair's Axiom)**.** In a plane, given a line and a point not on it, a unique line parallel to the given line can be drawn through the point.

Remark 2.1. Often Playfair's axiom is stated more generally, and only asserts that *at most one* line parallel to a given line can be drawn. However, it is possible to prove directly from Euclids Postulates 1-4 that parallel lines exist (Book I Proposition 31): so our formulation is equivalent.

This does away with much of the seeming complexity of the original postulate, replacing the condition of precise angle measures and intersections with the stipulation that parallel lines are *unique*. However this does not help with the more substantive point of unease, that Postulate 5 says something about what is going on *arbitrarily far away*. After all, how do you check that a parallel line is unique other than to check that *all other lines* make some intersection, even though many of those intersections will be unobservably far away.

In 1733, a Jesuit Priest and geometer by the name of Giovanni Saccheri made a useful, and prescient observation. He asked himself, what are all the logical possible statements that could take the place of Euclid's 5th postulate, or the (now-called) Playfair's Axiom? Well, if the axiom states that there exists a *unique* parallel through a given point, the other logical choices are that there are *none* or there are *many*.

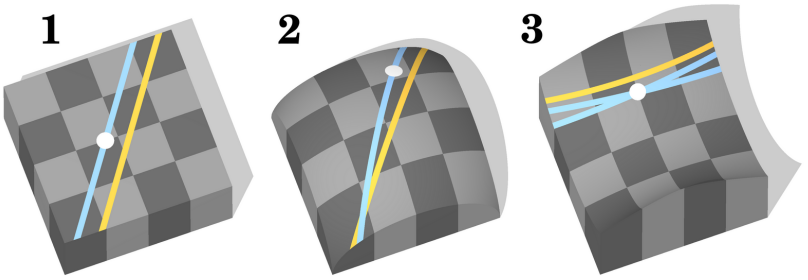


Figure 2.1.: The three possibilities of Giovanni Saccheri, image by Søren Peo Pedersen (Wikicommons)

Saccheri attempted to draw a contradiction from the other cases with Euclid's other four postulates, and while his investigations did not quite succeed, they led in some very interesting directions we will return to later on.

In the millennium and a half span from Proclus until the 1800s, many other foundational theorems of Euclidean geometry were also shown to be equivalent to the 5th. Among these are the following short list:

Theorem 2.1 (Some Equivalents to the Parallel Postulate)**.** *The following postulates are equivalent to the parallel postulate:*

- *All triangles have angle sum π.*
- *At least one rectangle exists.*
- *There exist triangles of arbitrarily large area*
- *Circumference/Radius is a constant for circles*
- *Area/Radius Squared is a constant for circles*
- *Equidistant curves to a line are lines*
- *There exists a pair of triangles which are similar, but not congruent*

- *The pythagorean theorem is true*
- *Given any 3 non-collinear points, there is a circle passing through them.*
- *Any two parallel lines have a common perpendicular.*

All of these properties are true of the world around us, and some of them (like the statement that *rectangles exist*) are even finitely checkable: it seems inconceivable to imagine a world where they were false! Yet, for over two millennia mathematicians the world over tried - and failed - to prove any single one of these statements from the first four postulates of Euclid alone.

The reason for this is grander than any of them could imagine, until Gauss, Lobachevsky and Bolyai entered the scene in the early 1800s. No one could prove any of these statements from the first four because they are *not implied by the first four*! There are logically possible, consistent mathematical worlds which *act* very similar to the geometry we find around us on earth, but for which the Pythagorean theorem is false. We will encounter these worlds (hyperbolic geometries) in the second half of this course.

2.1. Proofs: Triangles

To recover all of geometry in its full glory, he first invokes the 5th postulate in Proposition 29, which is restated below.

Proposition 2.1 (Alternate Angles are Equal (Euclid Prop 29))**.** *A straight line falling on parallel straight lines makes the alternate angles equal to one another*

Before we can prove this however, we need to talk a little bit about how the greeks measured angles. Euclid has several criteria called *common notions* he uses to axiomatize the means of measuring quantities such as angle, length, and area. For us in this short introduction, we will summarize some of these in the following "angle axioms".

Definition 2.4 (Angle Axioms)**.**

- Any two congruent angles have the same measure.
- If an angle A is divided into two disjoint angles B and C, the measure of A is the sum of the measures of B and C.

To get accustomed to these, we will first prove a practice result comparing angles, which does not need the parallel postulate.

Proposition 2.2 (Opposite Angles are Equal)**.** *If two lines intersect at a point in* $\mathbb{E}^2$*, the each of the two opposite pairs of angles have equal measure to one another.*

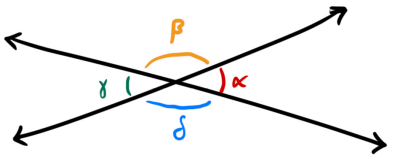


Figure 2.2.: Opposite angles are equal in measure

Remark 2.2. Here we have used that any two angles which sum to a straight line equal two right angles, as two right angles also form a straight line. Euclid proves this separately as Proposition 13.

Proof. Let two lines cross at a point, determining the four angles labled $\alpha, \beta, \gamma, \delta$ in the diagram above. Any two angles which together from a straight line are congruent to one another (a straight line is two right angles) and so we have the following equalities

$$\alpha + \beta = \beta + \gamma = \gamma + \delta = \delta + \alpha$$

Take the first equality, $\alpha + \beta = \beta + \gamma$ and subtract β from both sides: this gives $\alpha = \gamma$. Similarly, subtracting γ from the second equality $\beta + \gamma = \gamma + \delta$ yields $\beta = \delta$. □

Next, we'll prove our first lemma that *does* invoke the parallel postulate:

Proposition 2.3 (Corresponding Angles are Equal)**.** *If a line crosses a pair of parallel lines, the angles it makes with each of the parallel lines are equal in measure.*

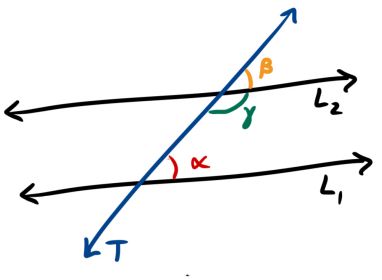


Figure 2.3.: Corresponding angles are equal in measure

Proof. Let L_1 and L_2 be two parallel lines, and T a third line crossing them transversely. Let α, β, γ be the three angles determined by these lines as labeled in the corresponding diagram.

Because T is a straight line, we know the sum $\gamma + \beta$ is two right angles. And, by Postulate 5, we know that as L_1 and L_2 do not intersect, the sum of α and γ must also be two right angles. Thus

$$\alpha + \gamma = \gamma + \beta$$

Subtracting γ from each side yields $\alpha = \beta$. □

Now we have enough information assembled to complete the task at hand.

Proving Euclid 29

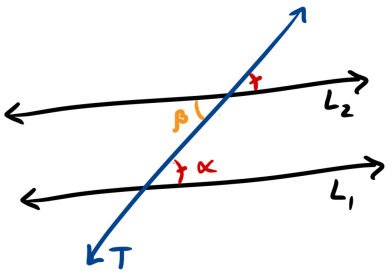


Figure 2.4.: Alternating angles are equal in measure

Proof. Again let L_1 and L_2 be lines crossed by a transverse line T. Denote by α, β the opposite interior angles labeled in the corresponding diagram. By Proposition 2.3, we know that the angle corresponding to α is of equal measure. But this angle is opposite to β, so by Proposition 2.2, we know this angle also equals β. Thus, $\alpha = \beta$. □

We have barely begun our use of the parallel postulate (we so far have used it precisely once, in a lemma about corresponding angles), but even just letting touch our theory is enough to have profound consequenes.

Theorem 2.2 (Triangles have Angle-Sum π)**.** *If T is any triangle in $\mathbb{E}^2$ with angles α, β, γ, then*

$$\alpha + \beta + \gamma = \pi$$

Below I give a proof using not the Parallel Postulate directly, but using the equivalent Playfair's axiom, that parallel lines *exist* and are *unique*. Check Euclid Proposition 32 for the original proof.

Proof. Consider an arbitrary triangle Δ, and choose one side S of the triangle, and let p denote the vertex of Δ opposite S.

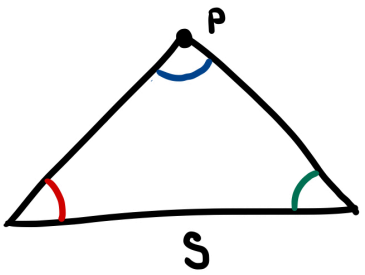


Figure 2.5.: An arbitrary triangle.

By Playfair's Axiom (Definition 2.3), there is a unique line through p which is parallel to the line containing S. Draw this line, and extend all the sides of the triangles to lines (Postulate 2).

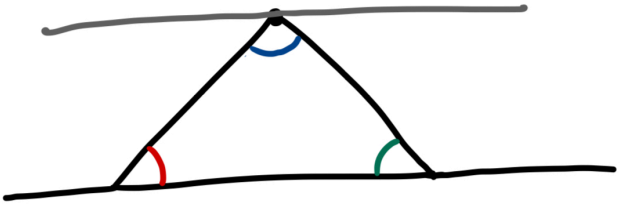


Figure 2.6.: A parallel to one side.

Note that the opposite interior angles formed by two sides of the triangle with the pair of parallel lines are equal (Proposition 2.1).

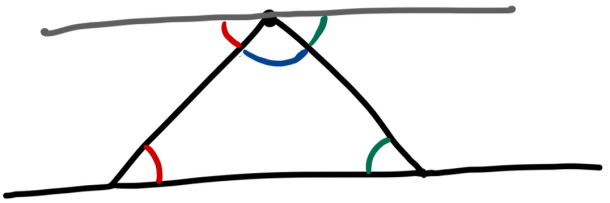


Figure 2.7.: Alternate angles to the bottom fill out a straight line.

Thus, the straight line at the top is the *sum of all three angles of the triangle!* In radians, this sum is π, so the sum of the three angles of Δ is equal to π. □

Exercise 2.1 (Polygon Angle Sum)**.** A polygon is *convex* if all of its angles are less than 180°, so that it has no "indents". Equivalently, a *convex* polygon is one where any line segment with endpoints on the boundary of the polygon lies *inside* the polygon.

Prove that the angle sum of convex quadrilaterals is a constant, for all quadrilaterals. Prove the angle sum of convex pentagons is also a constant. What are these constants?

What do you think the formula is for the sum of angles in a convex n-gon? (Optional: If you have seen mathematial induction, prove your guess!)

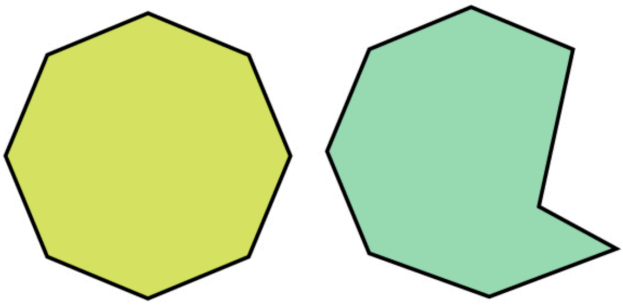


Figure 2.8.: Convex vs Non-Convex Octagon.

2.2. Quadrilaterals

Definition 2.5 (Quadrilaterals)**.** A quadrilateral is a polygon with four straight line sides. If all four angles are right angles, it is called a *rectangle*. A rectangle with all sides the same length is a *square*. If opposing sides are segments of parallel lines, its called a *parallelogram*.

Finally, we've uncovered enough to move beyond triangles a bit!

Theorem 2.3 (Rectangles Exist)**.** *There exists a quadrilateral in the plane, all of whose angles are right angles.*

Exercise 2.2. Prove Theorem 2.3 using Euclid's Postulates (and also Playfair's Axiom, if you like it), and the propositions given so far in this section.

Hint - we know how to make right angles now, and parallel lines through points. Start making some!

In fact (you may or may not have concluded this in your proof, depending on how general you were), for any lengths a and b, one can construct a rectangle with these as the side lengths. Now that we know rectangles exist, we can start asking questions about them: what patterns can we find?

Proposition 2.4 (Rectangles have Congruent Opposite Sides)**.** *Let $ABCD$ be a rectangle. Then the opposite sides AB, CD have the same length, as do the other pair BC, DA.*

Proof.

- Start with an arbitrary rectangle, and extend its sides into lines.
- Looking at one of the sides, it crosses the other pair in two right angles (by definition: its a rectangle). Thus the angle sum is a straight line, and so this other pair of sides is parallel, by the 5th postulate.

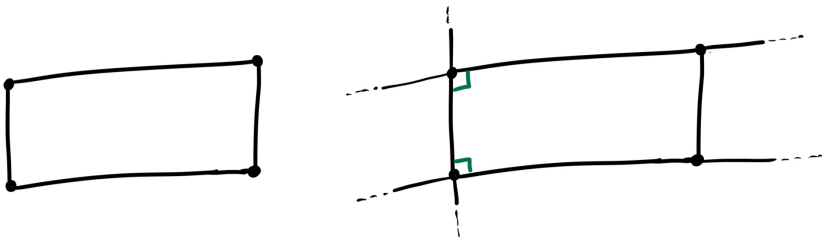


Figure 2.9.: A rectangles opposing sides are parallel.

- Draw a diagonal connecting an opposing pair of vertices of the rectangle.
- This diagonal divides the rectangle into two triangles, which both share a common side.
- But, since the this diagonal cuts across a pair of parallel lines, its alternate angles are equal.

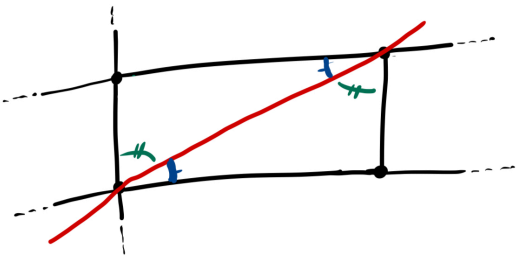


Figure 2.10.: A diagonal cuts the rectangle into two congruent triangles.

- Thus, the two triangles that have been formed are congruent to one another, by Angle-Side-Angle.
- But if the triangles are congruent, then they have the same side lengths.
- Thus, each pair of opposing sides of the rectangle must have the same length.

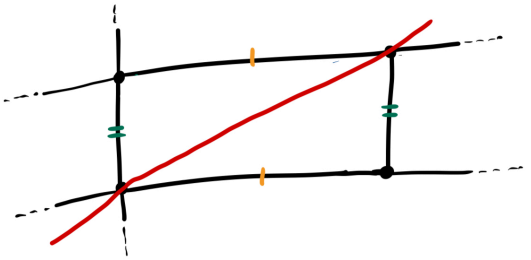


Figure 2.11.: Opposing sides of a rectangle are equal, since congruent triangles have the same side lengths.

□

Running the same argument with the other diagonal also gives a pair of triangles congruent to these, thus the diagonals of a rectangle must be equal in length to one another. In fact, more is true: the point where the diagonals intersect one another divides each of the diagonals in half - the Greeks would say their intersection *bisects* both of the diagonals.

A good way to get a feel for Euclidean geometry is to try and play around with properties like this that you discover. So, rectangles diagonals bisect one another, but is this all? Playing around a bit, its easy to see there should be more examples (draw any two line segments that cut each other in half making some sort of x, then connect up their vertices). But can we give any sort of order to this collection of shapes?

Exercise 2.3 (Bisecting Diagonals = Parallel Sides)**.** If the diagonals of a convex quadrilateral bisect one another, then that quadrilateral is a parallelogram.

3. Pythagoras

The pythagorean theorem needs no introduction, and is perhaps the most well known formula in mathematics from antiquity (and perhaps, only rivaled by $E = mc^2$ in the modern era).

Theorem 3.1. *On a right triangle with legs of length a, b and hypotenuse of length c, these lengths satisfy*

$$a^2 + b^2 = c^2$$

This theorem is fundamental to almost all real-world applications of geometry because it is the Greek foundation for the *distance function* it lets us measure distance between two points in the plane if we only know their horizontal and vertical separations.

Pythagoras' Theorem was not first discoverd by Pythagoras - and its first origins are lost to history though surviving tablets show that it was already in common use in Babylon by 3700 years ago.

Figure 3.1.: A babylonian tablet engraved with Pythagorean triples: three whole numbers for which the sum of the squares of the first two equals the square of the third (3, 4, 5); (8, 15, 17); and (5, 12, 13) are visible here. These were likely used to help determine land boundaries.

The same theorem was discovered by many mathematicians the world over, from ancient india to china and the middle east. But it was the Greeks who, in building the Elements, first realized its essential reliance on the theory of parallels.

While we are these days accustomed to algebraic expressions occurring in the midst of geometric arguments (and so, think of the Pythagorean theorem primarily as an *equation*) the origin of this equation deals fundamentally with the area of squares. Indeed, Euclid's original statement was:

In right-angled triangles the square on the side opposite the right angle equals the sum of the squares on the sides containing the right angle.

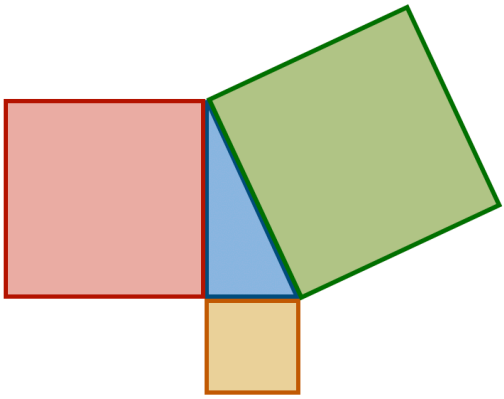


Figure 3.2.: The pythagorean theorem, illustrated.

And so, before we can proceed, we need to study area.

3.1. Areas

Just like measuring angles, Euclid needed some additional rules to specify how areas are to be measured. Here I've summarized these in a modern phrasing for us to use.

Definition 3.1 (Area Axioms)**.**

- The area of a square of side length x is x^2.
- Any two congruent shapes have the same area.
- If a shape R is the disjoint union of two shapes S and T, the area of R is the sum of the areas of S and T.

Proposition 3.1 (Area of a Rectangle)**.** *The area of a rectangle is equal to the product of its two side lengths.*

Proof. Create a square of side lengths $a + b$, and divide it up into a square with side lengths a, one with side lengths b and two rectangles with sides a, b, as shown in the following picture.

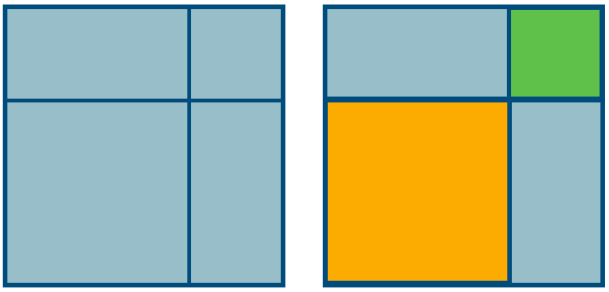


Figure 3.3.: Computing the area of a rectangle given the area of squares

Let A_R denote the unknown area of the rectangle. Using the area axioms, we can write the area of the large square as a sum of the areas of its components

$$(a + b)^2 = a^2 + b^2 + 2A_R$$

From here; its just algebra. Expanding the left hand side and cancelling like terms we find

$$A_R = ab$$

□

Exercise 3.1. Starting with segments of lengths a, b and using the postulates and what we have proven or stated in this text so far, construct the diagram used in the proof of Proposition 3.1, and justify it has the properties claimed of it (made of two squares and two rectangles of the correct dimensions).

We can now use this result to deduce the area of right triangles as well.

Proposition 3.2 (Area of a Right Triangle)**.** *The area of a right triangle is half the product of its two legs.*

Proof. In the proof of **?@prp-rect-opposite-sides** (showing that the opposing sides of a rectangle are the same length) we saw that the diagonal divides a rectangle into two congruent triangles.

Call the area of the rectangle A_R and the area of each of these triangles A_{T_1} and A_{T_2}. Since the triangles are congruent their areas are equal (Area Axiom 2), and since they together make up the entire rectangle, $A_R = A_{T_1} + A_{T_2}$ (Area Axiom 3). Putting these together

$$A_R = A_{T_1} + A_{T_2} = 2A_{T_1} \implies A_{T_1} = \frac{1}{2}A_R$$

Since we know A_R is the product of its two side lengths, and these are the legs of each triangle, we see that the area of the right triangle is half the product of its side lengths, as claimed. □

And finally we can extend this back to general triangles, recovering the familiar formula from high school geometry, that a triangle's area is given by $A = \frac{1}{2}bh$

Proposition 3.3 (Area of a Triangle)**.** *The area of a triangle is half the product of its base and its height.*

Exercise 3.2. Prove Proposition 3.3.

Hint - consider two cases: does the top vertex of the triangle lie over the base, or does it not? In both cases, try to use what we know about right triangles to help.

3.2. Proving the Pythagorean Theorem

This theorem has been proved many times! Indeed, there is an entire textbook by Elisha Scott Loomis devoted to distinct proofs of the Pythagorean theorem, collecting 367 in all, and this website gives 119 distinct proofs for you to peruse.

Euclid even proved this proposition in two distinct ways in *The Elements*, first in Book I, Proposition 47 and much later, in Book 6, Proposition 31.

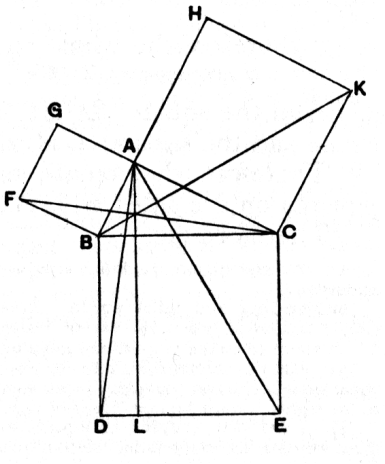
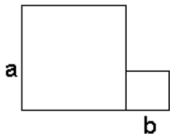


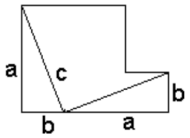
Figure 3.4.: The configuration used by Euclid to prove Pythagoras' Theorem in Book 1, Proposition 47.

However, as with many things in mathematics - time brings new insights and clairty even to the oldest of problems. The styles of proof that I personally find most elegant are all *rearrangement proofs*: starting with one collection of shapes and moving the pieces around in a way that forces the truth of the theorem.

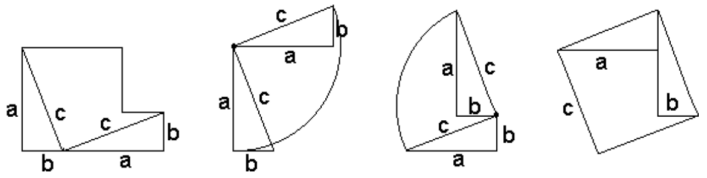
A particularly ingenious rearrangement proof was devised in the mid 800s CE by Thābit ibn Qurra (full name ,(a polymath from Baghdad who made contributions to mathematics, astronomy and medicine.



We start with two squares with sides **a** and **b**, respectively, placed side by side. The total area of the two squares is a^2+b^2.



The construction did not start with a triangle but now we draw two of them, both with sides **a** and **b** and hypotenuse **c**. Note that the segment common to the two squares has been removed. At this point we therefore have two triangles and a strange looking shape.



As a last step, we rotate the triangles 90°, each around its top vertex. The right one is rotated clockwise whereas the left triangle is rotated counterclockwise. Obviously the resulting shape is a square with the side c and area c^2.

Proof. □

Exercise 3.3. Justify that the resulting final shape here (in the proof by Qurra above) is indeed a square.

My favorite proof of the pythagorean theorem needs no words to be convincing: it's core idea is contained in the following diagram.

Proof. The blue areas in these two pictures are the same, as all we have done is slide around the triangles.

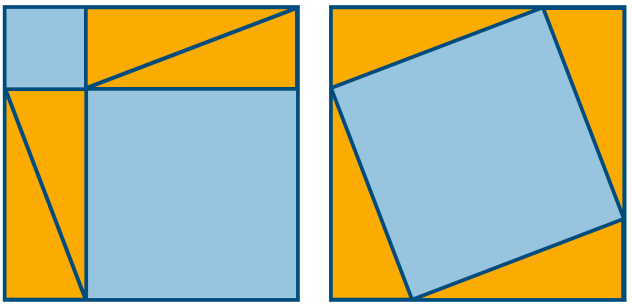
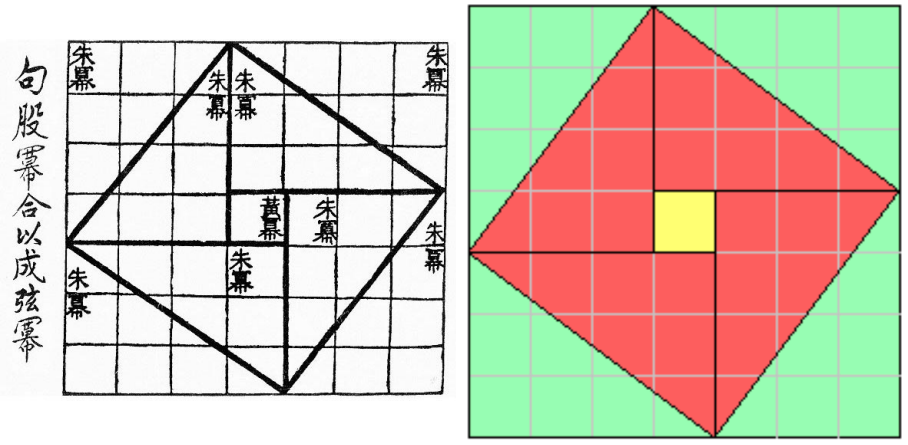


Figure 3.5.: A re-arrangement proof

Of course, to make this rigorous we have to explain why the re-arrangement really is the same square. □

Variants of this proof were discovered by the 12th century Hindu mathematician Bhaskara (Bhaskara II), and even much earlier, appearing in the Chinese astronomical text Zhoubi Suanjing (□□□□) from the second century BCE (and, claiming to record works from the 11th century BCE).



Exercise 3.4. Explain how the diagram from the Zhoubi Suanjing shows that $a^2 + b^2 = c^2$ using the formulas we've derived for the areas of squares and right triangles, and some algebra.

3.3. The Irrationality of $\sqrt{2}$

Traditionally the realization that irrational numbers must exist is attributed to Pythagoras (or his followers, the *Pythagoreans*). While the usual story is likely apocryphal (where the Pythagoreans kept as a strict secret the existence of irrational numbers, and murdered Hippasus for divulging it), their discovery nonetheless revealed a tension between two pillars of Greek mathematics

- Lengths constructed in Euclidean geometry are 'real'
- Any two lengths can be measured by a common ratio.

The first of these is merely stating that the Axioms of euclidean geometry are *constructive* - a proof in Euclidean geometry is a step-by-step recipe to really construct some line segment, polygon, or circle. So if the axioms say you can make something, you really can make it!

Exercise 3.5. Use the constructions of the previous section together with the Pythagorean theorem to prove $\sqrt{2}$ exists.

The second was born out by centuries of experience, in both mathematics and music all known quantities came in common ratios. Notes could be measured relative to each other, as could lengths. Ratios ruled the cosmos (today, we would say the Greeks believed in the number line $\mathbb{Q}$ of rational numbers).

Many proofs of the irrationality of $\sqrt{2}$ have been devised during the 2500 years since its discovery, with perhaps the most famous still being that recorded by Euclid, often phrased algebraically as proof by contradiction using fractions in lowest terms. But there are also purely geometric proofs of this fact. Below is a relatively modern one, devised by Stanley Tennenbaum around 1950.

Theorem 3.2. *The Square Root of 2 is not a rational number.*

Assume that $\sqrt{2} = m/n$ is the ratio of two whole numbers, so $2 = m^2/n^2$, or $2n^2 = m^2$. Geometrically, this means there is a square with integer side lengths (m), whose area is exactly twice the area of another integer-side-length square (n).

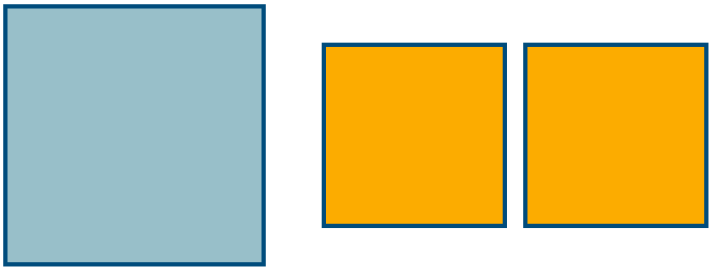


Figure 3.6.: A square with integer side lengths (blue) whose area is twice that of another integer square (yellow).

Now take the two smaller squares and position them inside the larger. They don't fit disjointly (remember, the sum of their areas is the entire square! So if they leave any of the square unfilled they must overlap somewhere else to make up for it).

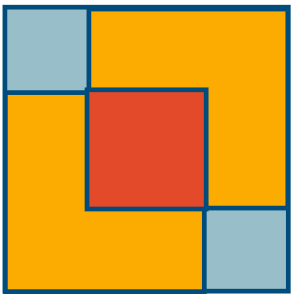


Figure 3.7.: Placing two of the smaller inside the larger must cause an overlap.

Actually, this picture determines three new squares, along the diagonal - two unfilled and one "double covered" by the overlapping yellow ones.

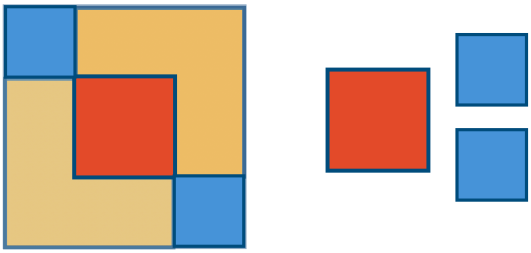


Figure 3.8.: This determines three new squares down the diagonal.

What do we know about these three squares? Well, they all have integer side lengths, as they are differences of integer side length squares (thus, they're *smaller*) than the side lengths of our original squares. But what can we say about their areas?

The two original yellow square's areas exactly sum to the blue squares area: this means the amount they miss (the two smaller blue squares) equals exactly the amount of overlap (the red square). So these *new* squares have the property that twice the area of the smaller add up to the area of the larger! This is just the start of what could easily become an infinite process: we now have a procedure that takes *any integer square solution* and produces a new solution *with smaller squares.* We could repeat this process again and again, getting ever smaller squares.

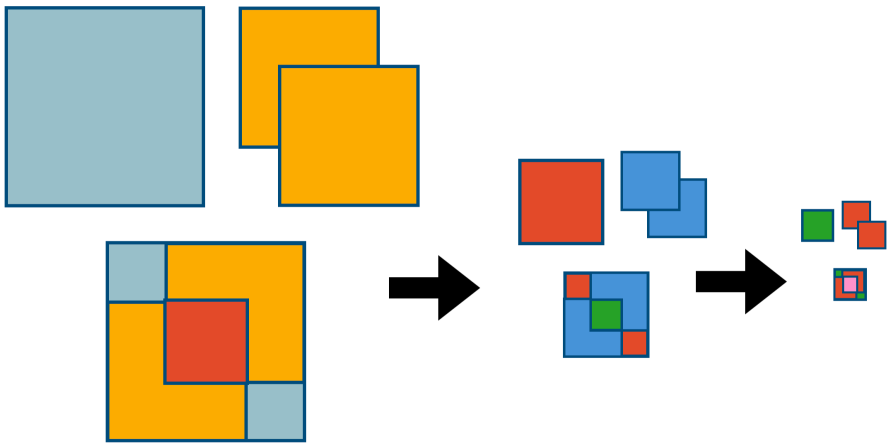
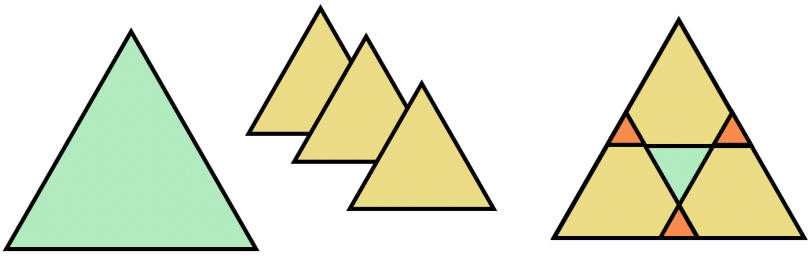


Figure 3.9.: From this, we can do an infinite regress.

But this clearly cannot be! It is impossible to make a list of ever decreasing positive integers, as there is a smallest positive integer: one! Assuming there was *any* rational solution to $\sqrt{2} = m/n$ gave us an infinite procedure to make smaller and smaller integer solutions forever, which cannot happen. Thus there cannot be any solutions at all!

And, the square root of 2 must be irrational.

Exercise 3.6 (The Square Root of 3)**.** Construct a similar argument showing that it is impossible to find two integer side-length equilateral triangles where one has three times the area of the other.



4. Archimedes

In all six books and 465 propositions of *The Elements*, Euclid never attempts to measure the length of a single curve, nor the area of a non-polygonal figure. Pristine, Greek, axiomatic geometry was about lines, and figures made out of them.

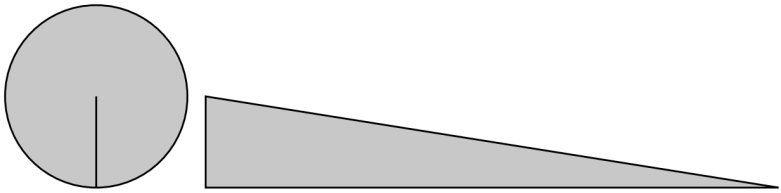
Curves were a different sort of object, a different category or type of thing in the mind of some of the ancients. Much as we would never try to measure the length of a gallon, they would never try to measure the length of a curve.

But to Archimedes, it was not a matter of kind, but of technology. Curves could be measured, if only the correct tools for the job could be developed.

4.1. Measurement of the Circle

In 250BCE Archimedes wrote a mathematical text entitled Κύκλου μέτρησις, or "Measurement of the Circle". While likely much of the text has been lost to time, an important theorem remains

Theorem 4.1 (Area of the Circle: Archimedes)**.** *The area of any circle is equal to the area of a right triangle with one side equal to the circle's radius, and the other side to the circle's circumference.*



This is the first time in greek mathematics that a curved object has been equated to a straight one. The idea of archimedes' argument is both beautiful and ingenious, but the difficulties of following it through using the mathematics of the time were considerable.

Archimedes began by approximating the circle by a polygon with a large number of sides.

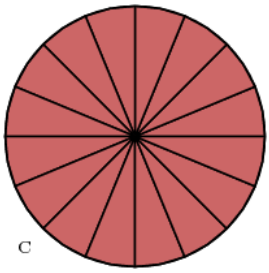


Figure 4.1.: A polygonal approximation of the circle.

He then cut the polygon into triangles, and "unrolled it" along its perimeter, into a sawtooth of wedges. This unrolling has not changed the total area, so this line of triangles has the same area as the original polygon.

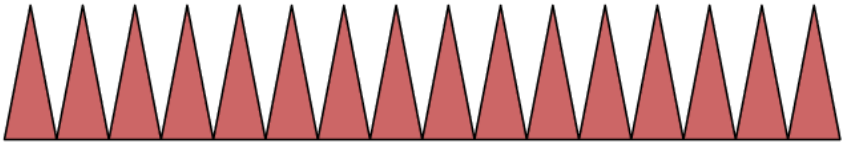


Figure 4.2.: Unrolling a polygon into triangles.

Then, Archimedes recalls that the area of a triangle is given by half its base times its height: that means if you *shear* a triangle, the area is unchanged as both the base and height are not altered by the procedure.

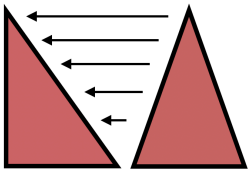


Figure 4.3.: Shearing a triangle leaves area invariant.

So, Archimedes shears all of the triangles along the sawtooth to the left, until all of their vertices coincide atop the perpendicular to the leftmost edge.

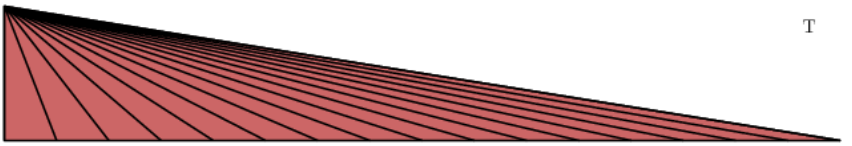
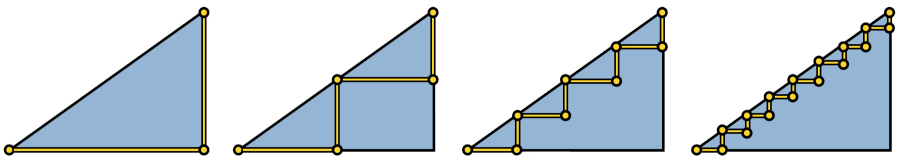


Figure 4.4.: Shearing the sawtooth produces a right triangle.

Thus, for every regular n-gon, archimedes can find an *exact* right triangle which has the same area, and whose height is the radius of the polygon and whose base is the perimeter. Archimedes then *very carefully* argues that as the number of sides of the polygon goes to infinity, the difference of its area from the circle, goes to zero, and the difference of its perimeter from the circumference of the circle does as well. Thus, the circle must share the same property as the polygons, it must have the same area as a right triangle made from its radius and circumference!

It is important one does not come away with the impression that it must just 'obviously' work, and that once Archimedes had his argument for polygons he was essentially done. Perhaps the best way to see this is to consider for yourself a seemingly analogous argument, which completely fails.

Exercise 4.1 (Convergence to the Diagonal)**.** Consider a simpler analog of Archimedes' situation, where instead of trying to measure a curve using straight lines, we are trying to measure a straight diagonal line using only horizontal and vertical segments. The following sequence of paths converges pointwise to the diagonal of the square, but what happens to the lengths?



If you believed that because this sequence of curves limits to the diagonal, its sequence of lengths must limit to the length of the diagonal, what would you have conjectured the pythagorean theorem to be?

To compute the convergence of areas, Archimedes was able to make clever use of the already existing *area axioms* (of Euclid's common notions, our Definition 3.1). However, to measure the length of a curved segment, Archimedes had to introduce two new axioms (as the measurement of curves is not possible in Euclid's framework).

Definition 4.1 (Archimedes Axiom I)**.** If p and q are distinct points in the plane, the line segment from p to q is the shortest of all paths connecting p to q.

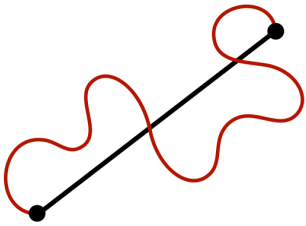


Figure 4.5.: The straight line is shorter, per Axiom I

Definition 4.2 (Archimedes Axiom II)**.** If there are two convex paths from p to q, and one lies inside the convex region defined by the other, than that one is shorter.

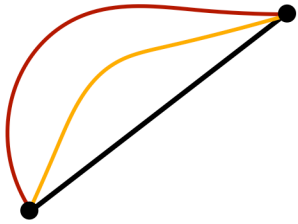


Figure 4.6.: The inside curve is shorter, per Axiom 2

Archimedes could not prove these axioms, as there was no deeper fundamental theory of lengths to rely on. However by formulating his argument axiomatically, he located any possible uncertainty in the axioms themselves: if these two plausible statements were true, then his striking conclusion *necessarily followed*.

Like much of Archimedes' work, these axioms were incredibly prescient and hit on deep truths of mathematics. In our modern re-building of geometry we will in fact take Archimedes' axiom 1 as the *definition* of a straight line. And Archimedes' restriction to only considering *convex curves* was also essential: we've already seen in Exercise 32.15 how delicate arguments can be. But it's even worse than this: when you drop the convexity requirement its not even true anymore that all curves *must have a length* (see the bit at the end of this section for an example).

The surviving text of *Measurement of the Circle* is only fragmentary, but if you wish to read some of the argument in its original (translated) form you can find it here.

Should you open this text you may be at first shocked by the quantity of numbers you see - in Greek works usually geometry reigns supreme and there is essentially no algebra to be found. But here Archimedes takes the opportunity to deduce a practical consequence from the theoretical development discussed above.

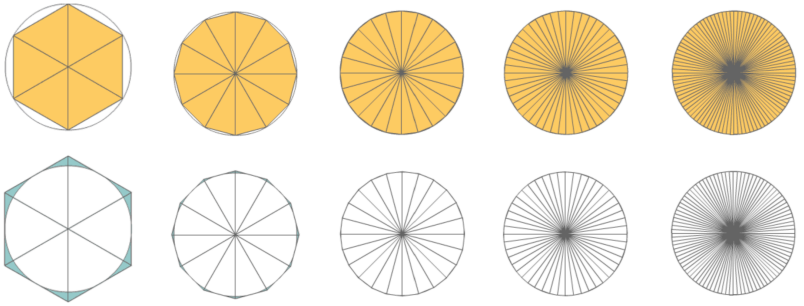


Figure 4.7.: Archimedes' calculation of π using the first several stages of the method of exhaustion: he computed provably over- and under-estimates (by #def-archimedes-axiom-2) starting with hexagons, and iteratively doubling the number of sides

Sides	Inscribed	Circumscribed
6	3.0	3.4641
12	3.1058	3.2154
24	3.1326	3.1597
48	3.1394	3.1461
96	3.1410	3.1427

A Fractal in the Plane

The Koch Snowflake is a *fractal*, defined as the limit of an infinite process starting from a single equilateral triangle. To go from one level to the next, every line segment of the previous level is divided into thirds, and the middle third replaced with the other two sides of an equilateral triangle built on that side.

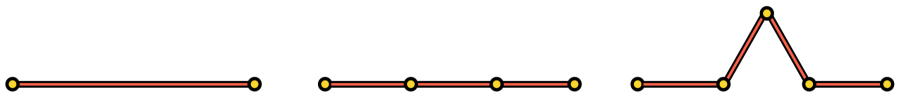


Figure 4.8.: The Koch subdivision rule: replace the middle third of every line segment with the other two sides of an equilateral triangle.

Doing this to *every line segment* quickly turns the triangle into a spiky snowflake like shape, hence the name. Denote by K_n the result of the n^{th} level of this procedure.

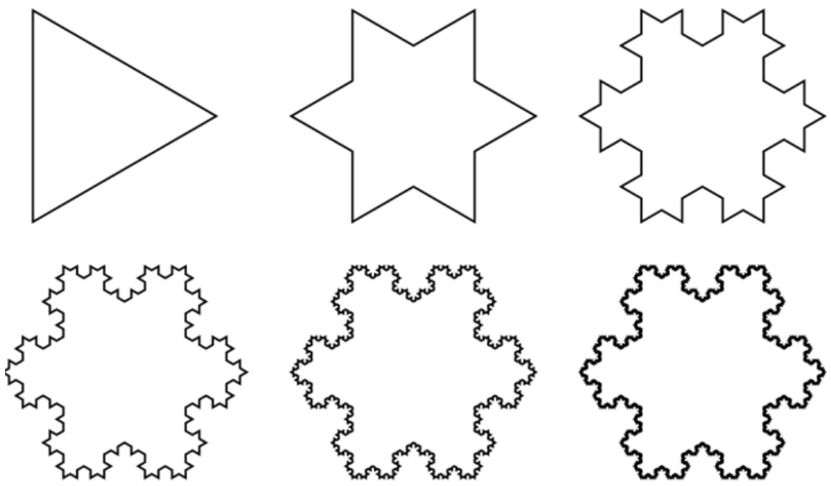


Figure 4.9.: The first six stages K_0, K_1, K_2, K_3, K_4 and K_5 of the Koch snowflake procedure. K_∞ is the fractal itself.

Say the initial triangle at level 0 has perimeter P, and area A. Then we can define the numbers P_n to be the perimeter of the n^{th} level, and A_n to be the area of the n^{th} level..

Exercise 4.2 (The Koch Snowflake Length)**.** What are the perimeters P_1, P_2 and P_3? Conjecture (and prove by induction, if you've had an intro-to-proofs class) a formula for the perimeter P_n.

Explain why as $n \to \infty$ this diverges (using the type of reasoning you would give in a calculus course): thus, the Koch snowflake fractal cannot be assigned a length!

Before doing the next problem: ask yourself what happens to the area of an equilateral triangle when you shrink its sides by a factor of 3? Can you draw a diagram (similar to that from last week's Exercise 32.9 but larger) to see what the ratio of areas must be?

Exercise 4.3 (The Koch Snowflake Area)**.** What are the areas A_1, A_2 and A_3 in terms of the original area A?

Find an infinite series that represents the area of the n^{th} stage A_n (if you've taken an intro to proofs class or beyond - prove it by induction!). Use calculus reasoning to sum this series and show that while the Koch snowflake does not have a perimeter, it *does* have a finite area!

4.2. Quadrature of the Parabola

Archimedes also found the area enclosed by a segment of a parabola and a straight line through an ingenious infinite process. His theorem relates the area of this parabolic segment to the area of the largest triangle that can be inscribed within.

Theorem 4.2. *The area of the segment bounded by a parabola and a chord is* $4/3^{rd}$*s the area of the largest inscribed triangle.*

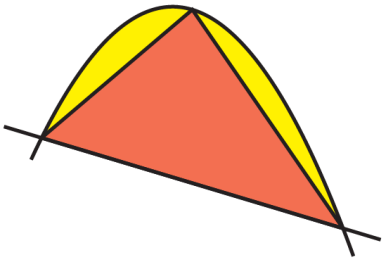


Figure 4.10.: A parabolic region and its largest inscribed triangle

After first describing how to find the largest inscribed triangle (using a calculation of the *tangent lines* to a parabola), Archimedes notes that this triangle divides the remaining region into two more parabolic regions. And, he could fill these with their largest triangles as well!

These two triangles then divide the remaining region of the parabola into *four new parabolic regions*, each of which has their own largest triangle, and so on.

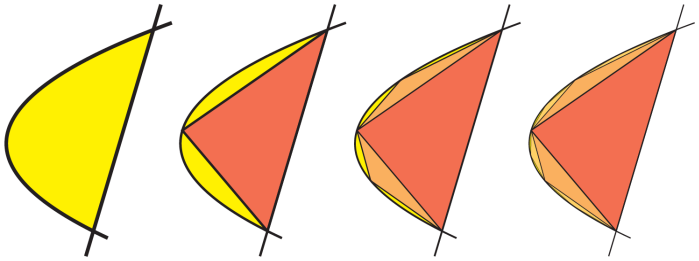


Figure 4.11.: Archimedes' infinite construction of the parabolic segment from triangles

Archimedes proves that in the limit, after doing this infinitely many times, the triangles *completely fill* the parabolic segment, with zero area left over. Thus, the only

task remaining is to add up the area of these infinitely many triangles. And here, he discovers an interesting pattern.

We will call the first triangle in the construction *stage 0* of the process. Then the two triangles we make next comprise *stage 1*, the ensuing four triangles *stage 2*, and the next eight *stage 3*.

Proposition 4.1 (Area of the n^{th} stage)**.** *The total area of the triangles in each stage is* $1/4$ *the total area of triangles in the previous stage.*

If A_n is the area in the n^{th} stage, Archimedes is saying that $A_{n+1} = \frac{1}{4}A_n$. Thus

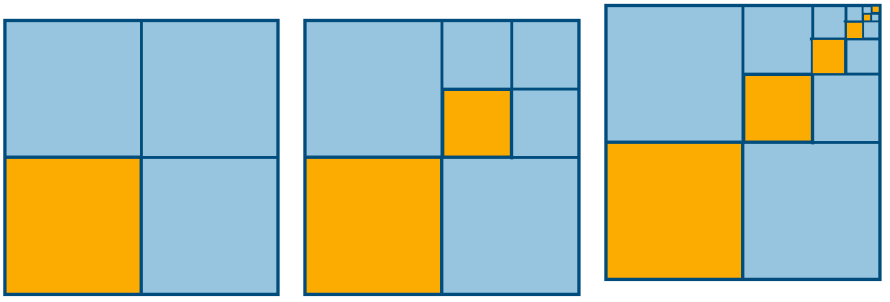
$$A_0 = T \quad A_1 = \frac{1}{4}T \quad A_2 = \frac{1}{16}T \quad A_3 = \frac{1}{64}T \ldots$$

And the total area A is the infinte sum

$$\begin{aligned} A &= T + \frac{1}{4}T + \frac{1}{16}T + \frac{1}{64}T + \cdots \\ &= \left(1 + \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \cdots\right)T \end{aligned}$$

Now Archimedes only has to sum this series. For us moderns this is no trouble: we recognize this immediately as a geometric series

But why is it called *geometric*? Well (this is not the only reason, but...) Archimedes was the first human to sum such a series, and he did so completely geometrically. Ignoring the leading 1, we can interpret all the fractions as proportions of the area of a square. The first term $1/4$ tells us to take a quarter of the square, the next term says to take a quarter of a quarter more, and so on. Repeating this process infinitely, Archimedes ends up with the following figure, where the highlighted squares on the diagonal represent the completed infinite sum.



(a) The first term: $1/4$ (b) The second term: $1/4+1/16$ (c) The infinite process: $1/4 + 1/16 + 1/64 + \cdots$

He then notes that this is precisely one third the area of the bounding square, as two more identical copies of this sequence of squares fill it entirely (just slide our squares to the left, or down). Thus, this infinite sum is precisely $1/3$, and so the total area is 1 plus this, or $4/3$

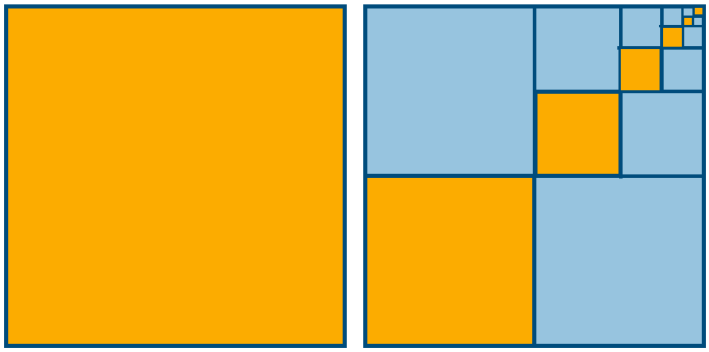


Figure 4.13.: The area of the parabola is the yellow shaded region in these squares

Exercise 4.4. Use the result of Exercise 32.9 (that you can inscribe an equilateral triangle with half the side lengths) to produce an alternative proof of Archimedes sum, but dividing up a triangle instead of a square.

Exercise 4.5. Construct an argument in the same spirit as archimedes to show the following equality:

$$\sum_{n=1}^{\infty} \left(\frac{1}{3}\right)^n = \frac{1}{2}$$

Can you cut a shape iteratively into thirds? It may not be as pretty as Archimedes', but thats oK!

4.3. The Sphere and the Cylinder

Archimedes continued his investigations of curves into the third dimension, where he proved fundamental results about the sphere.

Theorem 4.3 (Sphere Volume: Archimedes)**.** *The volume of the sphere is equal to the volume of its enclosing cylinder, minus the right circular cone with the same base and height.*

That is, the sphere's volume is $2/3$*rds that of its enclosing cylinder.*

Like before, Archimedes required careful use of the method of exhaustion to prove that equality held in the limit. Modern calculus allows us to reach the same result directly from Archimedes' insight much faster.

Exercise 4.6 (Sphere and Cone Slices)**.** Using Calculus, find the volume of both of these shapes as volumes of revolution and show they are equal.

But Archimedes was not content to understand the *volume* of the sphere, he also wanted to relate its surface area to the area of a known shape. He succeeded via an absolutely ingenious argument, to prove the following

Theorem 4.4 (Sphere Surface Area: Archimedes)**.** *The surface area of a sphere is equal to that of its enclosing cylinder.*

Mathematics yutuber 3Blue1Brown has made an excellent video discussing Archimedes proof, which I encourage you to watch: we would have a movie day in class if I we did not have many other interesting places to go!

https://youtu.be/GNcFjFmqEc8?feature=shared

Archimedes himself was so proud of this argument that he instructed that a sphere and cylinder be engraved on his tombstone. After he was killed during the Roman invasion of Syracruse in 212BCE, his tomb was quickly forgotten, only to be found centuries later when the great roman orator Cicero searched it out in 75BCE. In his own words:

> "Once, while I was superintendent in Syracuse, I brought out from the dust Archimedes, a distinguished citizen of that city. In fact, I searched for his tomb, ignored by the Syracusans, surrounded on all sides and covered with brambles and weeds. The Syracusan denied absolutely that it existed, but I possessed the senari verses written on his tomb, according to which on top of the tomb of Archimedes a sphere with a cylinder had been placed. But I was examining everything with the eyes … And shortly after I noticed a small hill not far emerged from the bushes. On it there was the figure of a sphere and a cylinder. And I said immediately to the Syracusans "That's what I wanted!" > Cicero, 75 BC

If you are interested in reading Archimedes' original work a translation of the paper The Sphere and the Cylinder is available here.

5. Modern Axioms

Euclid's postulates were chosen with care to be both *self-evident* and *useful.* But they are by no means the only *possible* axiom set one could choose to base Euclidean geometry off of. Just like it is possible for other statements to be equivalent to Postulate 5, it's also possible for another set of axioms to be Equivalent to Euclid's:

Definition 5.1 (Equivalent Axiom Systems)**.** Two axiom systems $\mathscr{A}$ and $\mathscr{B}$ are *equivalent* if you can both use the axioms of $\mathscr{A}$ to prove the axioms of $\mathscr{B}$, and vice versa.

In modern math, when defining something axiomatically we often prefer to choose axioms whose *meaning* is clear. Can we formulate a collection of axioms equivalent to Euclid's, that capture the *essence* of the geometry of the plane?

5.0.1. Axioms 1,2 & 3: Space is Complete & Infinite

The first three axioms of Euclid focus on the ability to draw lines (between any points, and of any length) and circles (of any radius). All of these together work to capture the property that space doesn't have any holes, and goes on forever.

Definition 5.2. A space X is *complete* if it does not have any holes, gaps or boundaries. Intuitively, a space is complete if you can continue walking straight in any direction, for as long as you like.

It is easiest to explain this notion by giving *non*-examples. The unit disk $D = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$ is not complete because if you start at the center you only have to walk one unit before you have to stop: you've reached the edge of space!

The punctured plane (all of $\mathbb{R}^2$, except the origin has been removed) is also not complete: any line segment passing through the origin in $\mathbb{R}^2$ cannot exist in this space, if you were to try and walk along it you would have to stop when you hit the missing point!

But being complete does not imply that space is infinite: indeed, the surface of the earth is complete, but finite in size! Anyone who starts walking in any direction on the earth's surface can continue walking forever without falling off the world, they'll just come back to their old location over and over.

The other property that us moderns would see as implicitly underlying the first three axioms of Euclid is the infinitude of space.

Definition 5.3. A space X is *infinite* if there are pairs of points arbitrarily far apart from one another.

The way to check if a space is infinite is to ask, "*for every natural number N, can I find a pair of points farther apart than N?" From this reasoning, we can see that the real line $\mathbb{R}$ is infinite, as we can look at the points 0 and $N + 1$: they're at distance $N + 1$ apart, which is greater than N. The same argument applies to the plane or 3-dimensional space, or any $\mathbb{R}^m$.

But this fails for the sphere: while it is complete its *finite in size* the farthest two points can possibly be from one another is if they are opposites on the sphere (like the north and south pole). And these points are only distance π apart, so there are no points on the unit sphere at distance greater than 4.

5.0.2. Axiom 4: Space is Homogeneous and Isotropic

Euclids fourth postulate is short and intuitive: all right angles are equal. But it's actually doing a lot of work! To see this, we must unpack what Euclid meant. Two angles are *equal* (in their measure) if they are congruent: that is, if there is a rigid motion of space that carries one to the other. Thus, Euclid here is claiming that you can always translate and rotate space so that any right angle is carried to any other.

Us moderns would naturally separate this into two actions: you can *translate* space to carry any point to any other, and then you can separately *rotate* space about any point, carrying any direction to any other. These properties are called *homogenity* and *isotropy* respectively.

Definition 5.4 (Homogeneous Space)**.** A space is *homogeneous* if for every pair of points in the space, there is a rigid motion taking one to the other.

Definition 5.5 (Isotropic Space)**.** A space is *isotropic* if for any point p and any two directions leaving p, there is a rotation of the space taking one direction to the other.

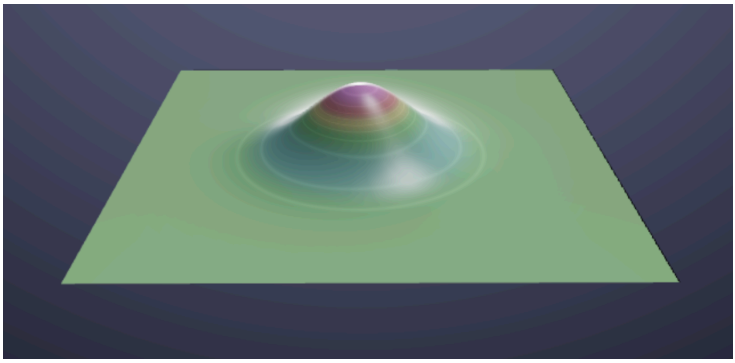


Figure 5.1.: A space that is *not homogenous.* Space looks different near points on the hill to points away from the hill. This space is also not isotropic (on the sides of the hill there's no symmetry between up and down), even though it *does* have rotational symmetry around the very apex of the hill.

We will be able to make these notions much more precise shortly, when we come back and redevelop geometry from calculus. But even before having everything rigorously defined, its useful to see these properties in action.

Example 5.1. Isotropy implies homogenity.

Proof. Let p and q be distinct points of a space X, and draw the line segment between them. Say this line segment is of length L, and mark the point m which is of length $L/2$ along it: the midpoint. Since X is isotropic there are rotations about m of any angle we wish.

Rotate about m by 180 degrees: this exchanges the points p and q. Thus there is a motion of X taking p to q, so X is homogeneous as claimed. □

In two dimensions, it turns out that homogenity also implies isotropy: if a space looks the same at every point then it also looks the same in every direction. But this is *false* in higher dimensions! Indeed, some of my favorite spaces are three dimensional worlds which are homogeneous but *not* isotropic.

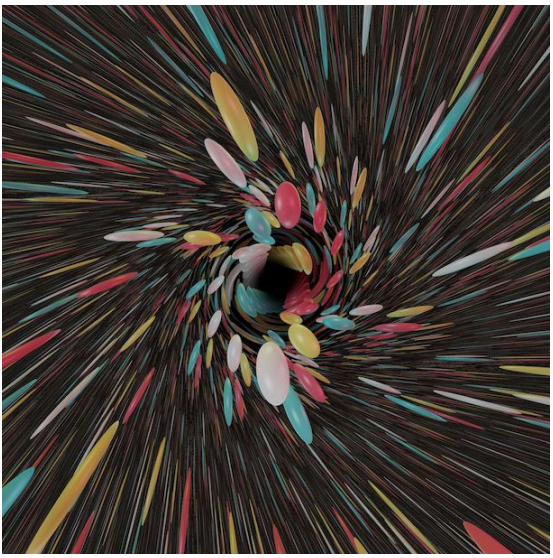


Figure 5.2.: A quick peek at a geometry called SL_2, which is homogenous but not isotropic. You can tell, because the center of your field of view looks *very different* than the view off to the side.

5.0.3. Axiom 5: Space is Flat

The fifth axiom, and all of its equivalents, capture something about space above and beyond the fact that it is infinite in extent and looks the same at every point.

By the list of equivalents to postulate 5, this additional bit of information has a lot of effects on the space: it determines how lines, circles, and triangles behave and it forces the Pythagorean theorem to be true!

It is difficult for us to give a full definition of "flatness" here, but we will in due time. Indeed - much of this course's purpose is to specifically get acquainted with this notion. For now, we'll make do with the following intuition: the plane is flat, and any surface you can make by bending the plane without stretching is also flat. Thus, the surface of a cylinder is flat, as you can roll up a sheet of paper without stretching it, as is the surface of a cone.

Definition 5.6 (Modern Axioms for Euclidean Geometry)**.** The Euclidean plane is

- Complete
- Infinite
- Homogeneous
- Isotropic
- Flat

There are spaces which are *not* flat - the surface of a sphere, for one. Our definition of flatness (and the lack thereof - curvature) will require mathematics beyond the Greeks, and we will return in detail once we have construction our geometric foundations from calculus.

Part II.

Calculus

6. Fundamental Strategy

Calculus is the story of humanity's quest to understand infinity: dealing with the infinitely small (differentiation), infinitely large (convergence), and processes that occur infinitely often (sequences, and infinite series). Out of this philosophical quandries grew an extremely useful set of mathematical tools that radically changed our world.

No longer were we constrained by the straight line geometry of the Greeks, or even the algebra of polynomials from the Middle East. The new mathematical tools provided a means of calculating - or *a calculus* with arbitrary curves.

> **The Fundamental Strategy of Calculus** Upon zooming in far enough, functions appear linear. At this level of zoom, you can replace difficult (nonlinear) problems with simple (linear) ones

This was Archimedes' fundamental insight. In trying to compute the area of a circle, he divided it into small circular wedges. Of course, it was no easier to calculate the area of a wedge than it was to calculate the area of the circle as a whole - as each wedge still had a curved (nonlinear) side. But - as the number of wedges grew - each wedge *shrank*, and allowed us to zoom in on a smaller and smaller piece of the curve. The farther we zoom, the closer this small curved arc is approximated by a *straight line* - making our problem into a *linear one*: we replace the area of a curved sector with the area of a triangle!

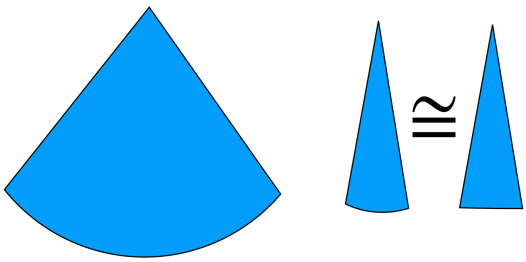


Figure 6.1.: Large circular sectors are not well approximated by triangles But *small* circular sectors only have a tiny piece of circular arc. Tiny arcs are approximately linear, so small sectors *are* well approximated by triangles.

This insight was discovered time and again over the following twenty centuries, by mathematicians the world over. As this is a geometry course - and not one on the history of calculus - we will not have the time to treat many of these amazing insights with the respect and awe they deserve. (Though, for those interested in such matters - consider taking Real Analysis with me in the spring!).

6.1. Infinitesimal Space

In the original, pre-rigorous formulations of calculus, mathematicians had the correct picture in their minds - that if you zoom in far enough on any smooth graph, it should look linear. But they had difficulty putting this intuition on a firm mathematical footing. Of course, at any *finite* level of zoom the curve is *not* linear, but still curves very slightly! It's only in the limit of *infinite zoom* that this approximation becomes exact.

But at the level of infinite zoom, what is the resulting line made out of? It can't be made out of regular points (x, y) on the plane: as these points are what makes up the curve. It must instead be made out of some new, *infinitesimally small* numbers. The intuition was that infinitely near every number x on the line, there were also *infinitesimal numbers* nearer to x than any other other "normal" (finite) real number. And infinitesimally near any point p in the plane, there were *infinitesimally small points*.

But immediately from this idea sprung forth many questions: how many of these infinitely small numbers must there be? If ϵ is one of these infinitesimals near x, then what about 2ϵ? That must still be infinitesimally near to x, as ϵ is so so so small! Similarly, $k\epsilon$ must be infinitesimally near x for any k: there's an entire number line of infinitesimal numbers near every real number!

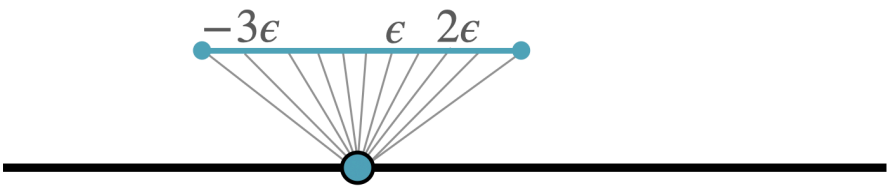


Figure 6.2.: Tangent space to a point on the line.

What about in the plane? If v and w are two points infinitesimally close to p, then what about $v+w$, or $kv+cw$ for scalars k, c? These must also be infinitesimally near p: so it appears there is an entire plane of infinitesimal points that must be near every finite point!

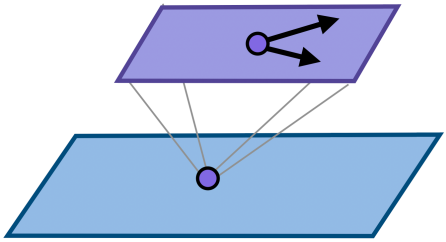


Figure 6.3.: Tangent space to a point on the plane.

This mental picture seemed to make perfect sense, but there were some deep questions. What are the rules of arithmetic for infinitesimally small numbers? Using arithmetic for just infinitesimal numbers alone, or just finite (normal) numbers alone posed no trouble, but strange things happened if you tried to combine them. If ϵ is infinitely close to zero, what is the number $1/\epsilon$? It must be bigger than all normal numbers: so, is it infinity? But then what is $2/\epsilon$? Another infinity larger than the first? Luckily, we don't have to worry about these questions, as during the development of Real Analysis mathematicians proved an important theorem:

Theorem 6.1. *The real line does not contain any infinitesimal numbers.*

This tells us as geometers that we should not be trying to combine the arithmetic of finite and infinitesimal numbers, and should instead keep them separate! This in fact makes things *easier*: at each point of the plane we have a plane of infinitesimal numbers, but different infinitesimal planes do not mix together, or with the points of the space. Because we often use these infinitesimal numbers to describe *tangents* to curves, the modern terminology for these infinitely zoomed in spaces are *tangent spaces*

Definition 6.1 (Tangent Space to the Line)**.** To every point $x \in \mathbb{R}$, there is attached a separate real line of infinitesimal numbers, denoted $T_x\mathbb{R}$ and called the *tangent space to $\mathbb{R}$ at x. Inside a fixed tangent space we can write down linear equations, but points of $T_x\mathbb{R}$ cannot be combined with points of $\mathbb{R}$ itself, or $T_y\mathbb{R}$ for any other point y.

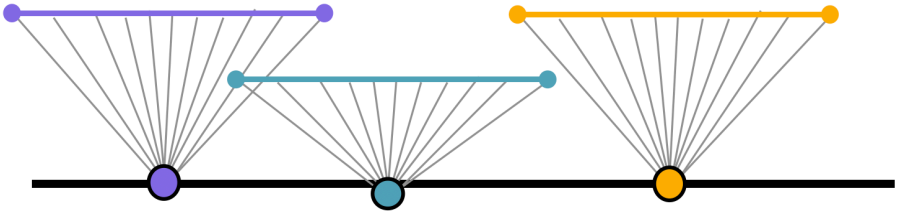


Figure 6.4.: There is a *separate* line of infinitesimal numbers attached to every single point.

Definition 6.2 (Tangent Space to the Plane)**.** To every point $p \in \mathbb{R}^2$, there is attached a separate plane of infinitesimal vectors, denoted $T_p\mathbb{R}^2$ and called the *tangent space to $\mathbb{R}^2$ at p. Inside a fixed tangent space the rules of linear algebra apply, but points of $T_p\mathbb{R}^2$ cannot be combined with points of $\mathbb{R}^2$ itself, or $T_q\mathbb{R}$ for any other point q.

Remark 6.1. "Nice enough" here means that when you zoom in, linear algebra begins to apply. The collection of spaces for which this is possible are called *manifolds*.

These definitions look very similar, and invite an immediate generalization. If X is *any* nice enough space, we can define the *tangent space* at every point as a vector space attached to that point, with the same dimension as X. We will make use of this more and more as the book progresses.

To keep things straight, its useful to have some notation for tangent vectors, that will help us remember where they are based at.

Definition 6.3. Let X be a space (the line, the plane, etc) and let p be a point of X. Then we denote the tangent space at p as T_pX, and when we need to be extra-precise, we put p a subscript even on individual vectors, to show they live in T_pX. For example

$$v_p = \langle 1, 2 \rangle_p = \begin{pmatrix} 1 \\ 2 \end{pmatrix}_p$$

A warning - if you don't keep careful track of where a vector is based, its easy to get confused! The vector $\langle 1, 1 \rangle$ may represent an infinitesimal vector based at $p = (1, 1)$ or an infinitesimal vector based at $q = (0, 1)$: but these two infinitesimal vectors are *different*!

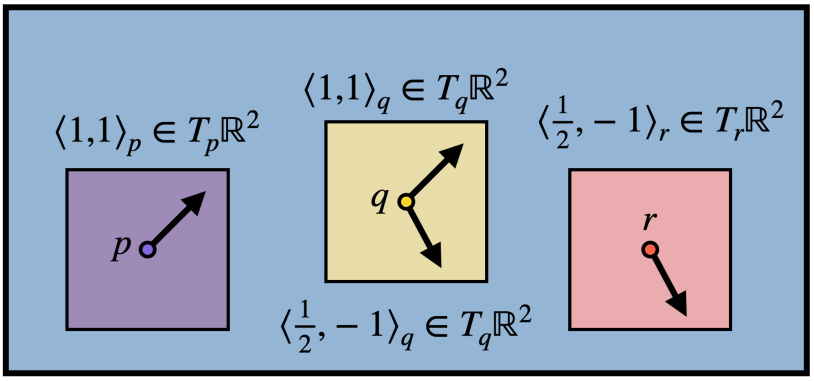


Figure 6.5.: Tangent vectors at different points truly live in different spaces, even if they have the *same description with coordinates*. It's always important to keep track of where a vector is based.

This will feel much more natural once we start actually doing calculus in this way.

6.2. Implementation

In broad strokes, modern applications of calculus follow closely Archimedes template. Given a difficult, nonlinear problem, the first step is to **zoom in**: to look infinitesimally in the tangent space of every point, where the problem simplifies and becomes linear.

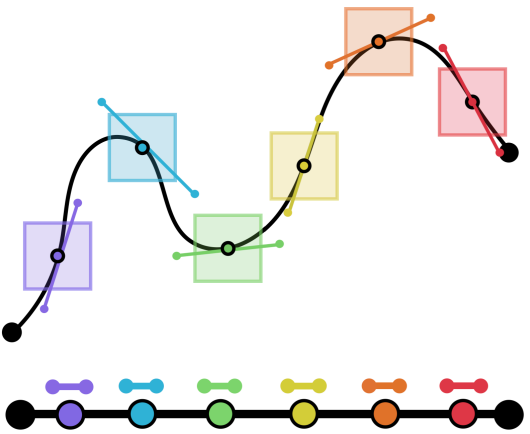


Figure 6.6.: Zooming in to each tangent space replaces the original function with a *linear function*.

While zoomed in, we **work infinitesimally** and simplify the problem as much as possible, taking advantage of the *linear mathematics* we now have to work with.

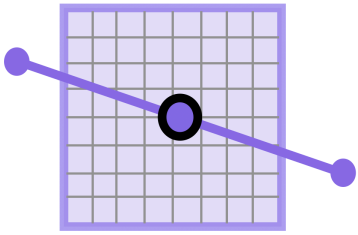


Figure 6.7.: Working infinitesimally often involves linear equations.

Once we have succeeded at this, we **zoom out**: piecing back together the infinitesimal linear information from all the relevant points into finite information answering the original question.

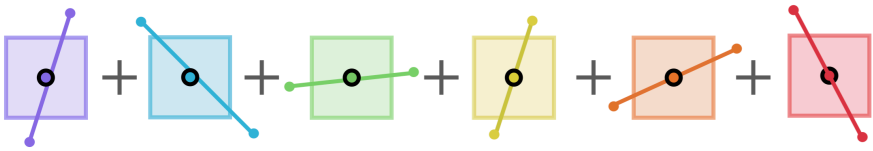


Figure 6.8.: Zooming out: combining infinitely many linear problems to solve one nonlinear problem.

These steps may look different from problem to problem, but the overall strategy remains unchanged. Most of the time the zoom in step involves some form of *differentiation*, or *tangent line approximation*, but the zoom out step can be more varied. The most common means of zooming out is *integration* combining infinitesimal information from an infinite continnuum of points. But infinite summation is also a means of zooming out - combining together an infinite sequence of infinitesimal terms. And sometimes, zooming out requires no extra work at all: if we can solve the problem completely at the infinitesimal level, our *zoom out* is only to come back up to reality and report the answer.

Below are several familiar examples of the fundamental strategy in action, so you can see the variety of methods fitting this general framework.

Example 6.1 (Area Under a Curve)**.** Starting with a continuous function on an interval $[a, b]$ in the real line, the goal is to find the area between the graph of f and the $x-$axis, from $x = a$ to $x = b$:

Zoom In: If we look infinitesimally near a single point x, we can approximate f as *constant* (even better than linear!).

Work Infinitesimally: Thus the infinitesimal area is given by an infinitesimal *rectangle* as the top side is no longer curved.

Zoom Out: To get the total area, we need to combine together (sum up) the areas of these infinitesimal rectangles. At any finite level of zoom this would be a (Riemann) sum, but at the infinite level of zoom of calculus, this becomes an integral!

$$\text{Area} = \int_a^b f(x)dx$$

Example 6.2 (Archimedes' Parabola)**.** To find the area under a parabolic segment, Archimedes followed a similar approach, though without the modern theory of integration. This makes the application of the fundamental strategy even more clear.

Zoom In: Fill the parabolic segment with more and more triangles. As the triangles get smaller, they zoom in more and more on smaller segments of the parabola, which are better and better approximated by the straight edges of the triangles.

Work Infinitesimally: The area of each triangle is easy to calculate, and the relationships between the areas of different triangles (for instance, those in level n vs those in level $n+1$) is deducible using Euclidean geometry. It turns out, the areas of the triangles follow a pattern - at each level the new contribution is $1/4$ the area of the previous.

Zoom Out: To find the area of the entire parabola, we must sum the areas accumulated at every level. Its no longer relevant where these numbers came from as we know the pattern: each number in the list is $1/4^{th}$ the previous. Its a geometric series! So zooming out requires us only to sum this series - the sum gives the total area.

Example 6.3 (Computing Function Values)**.** Consider the problem of evaluating a function line $\sin(x)$: what is the value of $\sin(0.23)$? Unlike polynomials the sine doesn't seem to have a nice *formula* that we can just plug 0.23 into...so, we use calculus to find one!

Zoom In: Some values of $\sin(x)$ we do know how to calculate well: the simple multiples of π that appear on the unit circle. So, we will zoom in on one of these: here choosing $x = 0$ (as its close to 0.23). At this point, we cannot see anything about $\sin(x)$ except its value, and the value of its derivatives at 0.

Work Infinitesimally: Differentiating sin repeatedly we see a pattern: its derivative cycles through the following list: $\sin, \cos, -\sin, -\cos$ and then repeats. We know how to evaluate both sin and cos at $x = 0$, so we can evaluate all the derivatives:

$$f(0) = 0,\ f'(0) = 1,\ f''(0) = 0,\ f'''(0) = -1,\ f''''(0) = 0 \cdots$$

Zoom Out: Now that we know the infinitesimal pattern, we must assemble all this information into a function. Its easy to write down a *linear* function with $f(0) = 0$, $f'(0) = 1$, jsut take $f(x) = x$. To also get $f''(0) = 0$ and $f'''(0) = -1$ we need a cubic function: $f(x) = x - \frac{1}{6}x^3$. And to get *all the derivatives right* we need an infinite series!

$$\sin(x) = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \frac{1}{7!}x^7 + \cdots$$

Summing this series at any point x gives the value of sin at that point - even though we only used the *infinitesimal information at* $x = 0$ to derive it! With this formula, its no trouble to evaluate $\sin(0.23)$, or any other value.

Example 6.4 (Maximizing a Function)**.** Given a smoothly varying function $f(x)$ on an interval $[a, b]$, a difficult problem is to find the point x_0 at which $f(x)$ is largest.

Zoom In: Our main insight is that when a function has reached its maximum value, it changes from *increasing* (before the peak) to *decreasing* (after the peak). Zooming in at some point x inside the interval (a, b), the rate of change of f at that point is captured by the *derivative*. So we just need to study the derivative.

Work Infinitesimally: If a function is increasing the derivative is positive, and when its decreasing it's negative. So, when it switches from increasing to decreasing, we must have $f'(x) = 0$. This has replaced a *calculus problem* (maximization) with an *algebra problem* (solving for the zero of a function).

Zoom Out: To zoom back out, we need to consider the entire interval and make sure we have found the answer. The calculus procedure above let us find all the points where $f'(x) = 0$ *inside* the interval - these are potential places for the maximum to occur. The other potential places are at teh *endpoints* of the interval. So, we need to compare the value of f all these points, and report the largest value we found.

In this Part of the book, we will do a deep dive into these three components of the fundamental strategy, so that we can utilize this powerful tool in the rest of our geometric investigations. First, we will learn about *working infinitesimally* - that is, what linear functions look like in one and two dimensions. Then we will learn how to *zoom in* - starting with a nonlinear function and differentiating it to get something linear. And finally, we will review methods of *zooming out*: integration and power series.

7. Working Infinitesimally

To use the fundamental strategy of calculus, we need to get good at zooming in - replacing a function with its linearization, as well as zooming out - putting these linearizations back together to answer our question. Fundamental to both of these tasks is understanding *linear things* themselves, so this is where we begin.

This class does not assume any previous knowledge of linear algebra, and we will introduce everything we need along the way (which is not that much! We will be using linear algebra as a *tool*, not delving into it deeply as the object of study itself). In this chapter I've collected the essential pieces of linear algebra that will come up throughout the course. For any of you who have taken linear algebra in the past, I would recommend you skim through this chapter to refresh your memory. For those of you who have not - there is no need to read the whole thing right now. Treat this chapter as a reference that you can return to time and again, as our toolkit in class expands. For now, its only necessary to read the section on vectors and the section on matrices.

7.1. Vectors

Vectors are a specific way to describe points in space. To picture vectors, often arrows are drawn based at a fixed point, called *the origin*. The length of the vector is called its *magnitude*, and we interpret this arrow as storing the data of a magnitude and a direction based at this origin. A one dimensional vector is an arrow on the line. If we call its origin *zero*, then We can think of it as ending at some particular real number: the size (or absolute value) of the number gives its *magnitude*, and the sign (positive or negative) is the *direction*.

Remark 7.1. For students familiar with linear algebra, this means we are essentially fixing the basis $\langle 1, 0\rangle$, $\langle 0, 1\rangle$

Vectors do not exist all by their lonesome, but instead come together in a collection called a *vector space*. The subject of linear algebra is really the study of vector spaces, and the power that this level of abstraction can provide. However, we will be much more pragmatic in this course: the only vector spaces we will ever need are the spaces $\mathbb{R}$ (the real line), $\mathbb{R}^2$ (the plane), and $\mathbb{R}^3$ (three dimensional space). Because of this, we will always be able to describe vectors in *cartesian coordinates*, writing them unambiguously as n-tuples of real numbers like this:

$$v = \langle a, b \rangle = \begin{pmatrix} a \\ b \end{pmatrix}$$

Definition 7.1 (Standard Basis)**.** For the vector space $\mathbb{R}^n$, the *standard basis* is the list of vectors all of whose entries are zero except for a single entry, which is equal to 1. For example, the standard basis for $\mathbb{R}^2$ is

$$e_1 = (1, 0) \qquad e_2 = (0, 1)$$

And the standard basis for $\mathbb{R}^3$ is

$$e_1 = (1, 0, 0) \qquad e_2 = (0, 1, 0) \qquad e_3 = (0, 0, 1)$$

7.1.1. Vector Arithmetic

Vectors, much like numbers, can be combined and modified using *operations*: they can be summed up using *vector addition*, and multiplied by numbers using *scalar multiplication*.

Definition 7.2 (Vector Addition)**.** If u, v are two vectors, then their *sum* is the vector whose tip lies at the opposite side of the parallelogram spanned by u and v. In coordinates, this is just the component-wise sum of the two vectors:

$$u = \langle a, b \rangle \qquad v = \langle c, d \rangle$$

$$u + v = \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} a + c \\ b + d \end{pmatrix}$$

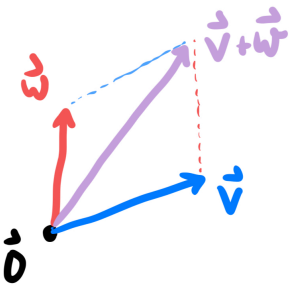


Figure 7.1.: Vector Addition

We will often see vector addition as a means of performing a *translation*: adding a vector $\vec{v}$ shifts a point $\vec{p}$ in the plane to a new point $\vec{p} + \vec{v}$. Doing this simultaneously

to all points in the plane *slides the entire plane by the vector* $\vec{v}$. *For example, if* $v = \langle 1, 2 \rangle$ *then* translation by v$ is the function

$$(x, y) \mapsto \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} x+1 \\ y+2 \end{pmatrix}$$

The second operation we can do to vectors is called *scalar multiplication*: this changes the length of a vector, without changing its direction (though, it flips the vector around *backwards* when the scalar is negative).

Definition 7.3 (Scalar Multiplication)**.** If v is a vector and k is a number (scalar), we can create a new vector kv which points in the direction of v, but is k times as long. In coordinates, if $v = \langle a, b \rangle$ then

$$kv = k \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} ka \\ kb \end{pmatrix}$$

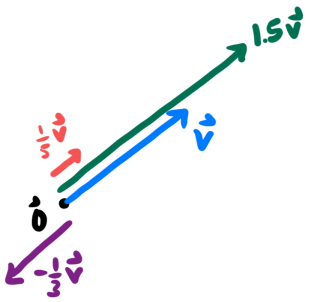


Figure 7.2.: Scalar Multiplication

The collection of all scalar multiples of a nonzero vector $\vec{v}$ trace out the *line* through the origin, containing the vector $\vec{v}$. Combining this with vector addition to allow for translations, we can easily describe lines in space in the language of linear algebra.

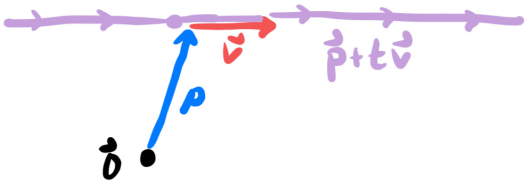


Figure 7.3.: Affine Lines

Definition 7.4 (Affine Lines)**.** An affine line in a vector space is a function of the form

$$\vec{\ell}(t) = \vec{p} + t\vec{v}$$

We can refer to such a line as *the line through* $\vec{p}$ *in direction* $\vec{v}$.

The Youtuber 3Blue1Brown has put together an excellent video series called the "Essence of Linear Algebra". While much if it is beyond what we need for this course - I highly recommend watching the entire series! I'll post throughout this article a few of the installments that are particularly relevant: here's the introductory video on vectors.

https://youtu.be/fNk_zzaMoSs?feature=shared

7.2. Linear Maps

The operations of addition and scalar multiplication are of fundamental importance to vectors. Because of this, functions which play nicely with addition and scalar multiplication will

Definition 7.5 (Linear Maps)**.** A function F between vector spaces is a *linear map* if:

- It preserves addition: $F(u + v) = F(u) + F(v)$ for all vectors u, v.
- It preserves scalar multiplication: $F(cv) = cF(v)$ for all scalars c and all vectors v.

It's easy to find examples of functions which are not linear: all they have to do is violate one of these two properties. For example, $f(x) = x^2$ is not linear since $f(x + y) = (x + y)^2 = x^2 + y^2 + 2xy$ and $f(x) + f(y) = x^2 + y^2$, so $f(x + y) \neq f(x) + f(y)$. In fact, most functions are *nonlinear*.

Example 7.1 (1 Dimensional Linear Map)**.** The single variable function $f(x) = 2x$ is a linear map. To see this, we check both addition and scalar multiplication:

$$f(x + y) = 2(x + y) = 2x + 2y = f(x) + f(y)$$

$$f(cx) = 2cx = c2x = cf(x)$$

Of course, nothing about the 2 above is special the functions $f(x) = mx$ - which we know from algebra classes to describe lines through the origin - are all examples of linear maps. Examples get more interesting in two dimensions:

Example 7.2 (2 Dimensional Linear Map)**.** The function $F(x, y) = (2x, x + y)$ is a linear map. Again, we just need to check addition and scalar multiplication. Let $u = \langle u_1, u_2 \rangle$, $v = \langle v_1, v_2$, and c be any constant. Then check, using the rules we learned above, that

$$F(u + v) = F(u) + F(v)$$
$$F(cu) = cF(u)$$

Below is one relatively straightforward warm-up proposition using the definition of linearity, which nonetheless proves very useful: linear transformations send lines to lines.

Proposition 7.1 (Linear Maps Preserve Lines)**.** *If $\ell(t) = p + tv$ is an affine line and F is a linear map, then $F(\ell(t))$ is also an affine line.*

Proof. This is just a computation, together with the definition of linear map and affine line. Plugging in $\ell(t)$, we use that F preserves addition, so$F(p + tv) = F(p) + F(tv)$. Next we use that F preserves scalar multiplcation, so $F(tv) = tF(v)$. Putting it all together,

$$F(p + tv) = F(p) + tF(v)$$

Since $F(p)$ and $F(v)$ are constant vectors, this result is of the form

$$\text{vector} + t \cdot \text{vector}$$

which is the same form we started with: so its also an affine line. □

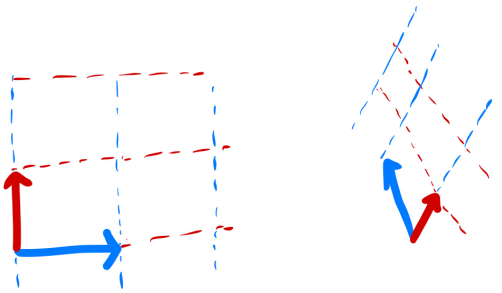


Figure 7.4.: 2D Linear Maps send lines to lines, so we can visualize them as squeezing or stretching a grid

Here's 3Blue1Brown's video on Linear Transformations and Matrices: it does an absolutely excellent job of displaying the geometric meaning of linear maps we just discovered above, as well as motivating the definition of matrices (which we define below).

https://youtu.be/kYB8IZa5AuE?feature=shared

7.3. Matrices

Linear maps are very constrained objects: the fact that they preserve addition and scalar multiplication tells us that its possible to reconstruct *exactly what they do to any point whatsoever* from very little data. We will mostly be concerned with linear maps from $\mathbb{R}^2 \to \mathbb{R}^2$, so I'll use this as an example.

Say we know that L is a linear map, and we also know what happens when we plug in the vectors $(1, 0)$ and $(0, 1)$.

$$L(1, 0) = (2, 3) \qquad L(0, 1) = (-1, 1)$$

How can we figure out what happens to (x, y) after applying L? Well, first we use addition and scalar multiplication to break down the vector (x, y) into simpler pieces.

$$(x, y) = (x, 0) + (0, y) = x(1, 0) + y(0, 1)$$

Then we can feed this *linear combination* into the function L, and use the fact that it preserves these operations to our advantage:

$$\begin{aligned} L(x, y) &= L(x(1, 0) + y(0, 1)) \\ &= L(x(1, 0)) + L(y(0, 1)) \\ &= xL(1, 0) + yL(0, 1) \\ &= x(2, 3) + y(-1, 1) \end{aligned}$$

We can further simplify this answer by using addition and scalar multiplication (again!):

$$\begin{aligned} x(2, 3) + y(-1, 1) &= (2x, 3x) + (-y, y) \\ &= (2x - y, 3x + y) \end{aligned}$$

Thus, from knowing *only* what L does to the vectors $(1, 0)$ and $(0, 1)$, we can deduce the entire formula for L

$$L(x, y) = (2x - y, 3x + y)$$

The takeaway from this computation is that remembering what a linear map does to the *standard basis vectors* is of fundamental importance. In fact, this is exactly what the notation of a matrix is all about!

Definition 7.6 (Matrix)**.** A matrix is an array of numbers. The following are all examples of matrices

$$\begin{pmatrix} 1 & 2 \end{pmatrix} \quad \begin{pmatrix} 3 \\ 7 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$$

Definition 7.7 (Matrix of a Linear Map)**.** If L is a linear map, the *matrix for* L has its first column equal to the *image of the first basis vector*, the second column equal to the *image of the second basis vector* etc. In symbols, for a map from $\mathbb{R}^2$:

$$\begin{pmatrix} | & | \\ L(1,0) & L(0,1) \\ | & | \end{pmatrix}$$

Example 7.3 (Matrix of a Linear Map)**.** Consider the linear transformation $L(x, y) = (2x - y, x + y)$. To find the matrix representation of L, we just need to compute L on the basis vectors $(1, 0)$ and $(0, 1)$:

$$L(1,0) = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \qquad L(0,1) = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

The first of these is the first column of the matrix, and the second is the second column: that's all there is to it!

$$L = \begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix}$$

Exercise 7.1 (Matrix of a Linear Map)**.** Find a matrix for the following linear maps:

- $L : \mathbb{R}^2 \to \mathbb{R}^2$ which has the equation $\$L(x, y) = (4x - 3y, 2x + 2y)$
- $M : \mathbb{R}^2 \to \mathbb{R}$ which has the equation $L(x, y) = 2x - 6y$.
- $N : \mathbb{R}^2 \to \mathbb{R}^3$ with $L(x, y) = (x - y, x + z, y - z)$.

One of the best ways to understand linear maps is to visualize *by hand* how the transform the plane. Below is a picture drawn on the Euclidean plane.

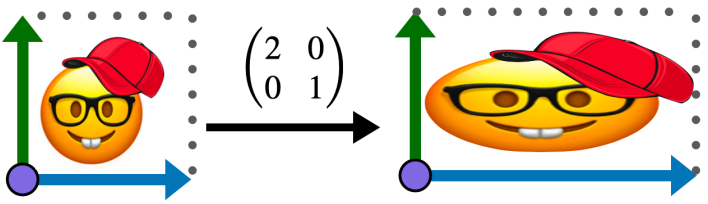


Figure 7.5.: Stretching an image via a linear transformation.

Applying the linear transformation with matrix $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ to this image turns the unit square of rectangles with sides $\langle 2, 0\rangle$ and $\langle 0, 1\rangle$. This transforms our image as below

Similarly, the transformation $\left(\begin{smallmatrix} -1/2 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ reflects the x axis, while leaving the y direction unchanged.

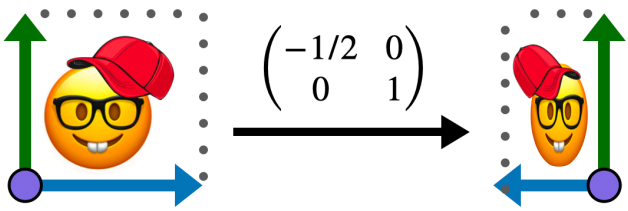


Figure 7.6.: Compressing and reflecting an image via a linear transformation.

But all sorts of changes can happen! Linear maps can rotate, stretch, and squish our original square / image into any sort of parallelogram!

Exercise 7.2. Choose your own image on the plane (hand-drawn is great!), and draw a reference image of it undistorted, inside the unit square. Then draw its image under each of the following linear transformations:

$$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \qquad \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \qquad \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \qquad \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

7.3.1. Composition & Multiplication

Now we have at our disposal an easy-to-remember, easy-to-write notation for linear maps. All we do is store the results of the map on the standard basis! But how do we *use* this? How can we actually apply this linear maps to points? Looking back to our explicit example where $\left(\begin{smallmatrix} 2 & -1 \\ 1 & 1 \end{smallmatrix}\right)$ corresponds to $L(x, y) = (2x - y, x + y)$, its clear: the first row stores the x and y coefficients of the first component, and the second row the coefficients of the second component.

Definition 7.8 (Applying a Matrix to a Vector)**.** Given the matrix $L = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, the linear transformation associated to this is

$$L(x, y) = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix}$$

This formula is called the **multiplication of a matrix by a vector**.

Now we know how to *apply* a linear transformation, but how do we *compose them*? If I have two linear transformations, which are each a function $\mathbb{R}^2 \to \mathbb{R}^2$, I can do one after the other and get a new linear transformation. Abstractly, this is no problem. But if I actually want to *compute things*? Each linear transformation is represented by a *matrix*, how do I combine together two matrices in the right way to make a matrix for the result?

Example 7.4. Its perhaps most instructive to do this directly yourself. Start with two linear transformations, say $L(x, y) = (x - y, 2x + y)$ and $M(x, y) = (3x + y, 2x - 5y)$, and compose them, simplifying the result as much as you can. What are the matrices for the three transformations L, M and $M \circ L$?

If you keep track of what you are doing during your simplification process, you'll notice a pattern: you can deduce the matrix for the composition directly from the matrices of the transformations themselves!

Definition 7.9 (Matrix Multiplication)**.** If L and M are linear transformations with the following two matrix representations

$$L = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad M = \begin{pmatrix} e & f \\ g & h \end{pmatrix}$$

Then the linear transformation $L \circ M$ has the following matrix:

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

The ij-entry of this matrix are formed by multiplying the i^{th} row of the first by the j^{th} column of the second element-wise, and adding up the results.

Early on in the course we will not have too much use for composing linear transformations explicitly, but once we reach the chapter on hyperbolic geometry - we will find this operation extremely useful to help explore spaces we struggle to visualize.

7.3.2. Inversion

How can we *undo* the behavior of a linear map?

Exercise 7.3. Given linear transformation $L(x, y) = (x + y, x - 2y)$, what vector does L send to $\langle 3, 4 \rangle$?

In the exercise above, we attempted to undo the behavior of L for a single vector. If we tried to do this for *all vectors* we would have a function that undoes the action of L. We call this an *inverse function*

Definition 7.10 (Inverses)**.** If $L : X \to Y$ is a function, an *inverse* to L is a function $M : Y \to X$ which undoes the behavior of L. That is, for every $x \in X$, if we apply L to get $y = L(x) \in Y$, the inverse M takes y back to x. Similarly if we start with M and then apply L they undo each other, so nothing changes. In symbols

$$M(L(x)) = x \,\forall x \in X \qquad L(M(y)) = y \,\forall y \in Y$$

Exercise 7.4. Try to invert the linear map from above: $L(x, y) = (x + y, x - 2y)$. Find a function $M(x, y) = (px + qy, rx + sy)$ such that $M(L(x, y)) = (x, y)$ and vice versea.

If you do the above exercise carefully, you'll find that the fact that the original linear map was $(x, y) \mapsto (x + y, x - 2y)$ did not matter: you could have used any constants at all, and ran the same sort of argument for any linear transformation $(x, y) \mapsto (ax + by, cx + dy)$ at all! We will never have need to invert anything besides a 2×2 matrix, so the important takeaway from this section is the following general formula.

Proposition 7.2 (Inverse of a 2×2)**.** *If $L = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ is a linear transformation, it is invertible if $ad - bc \neq 0$, and the inverse has matrix*

$$L^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

7.4. Determinants

In the formula for inverting a linear transformation above, a strange looking linear factor showed up in front of the matrix: the reciprocal of $ad - bc$. What does this quantity measure?

A linear transformation L of the plane takes a square (spanned by the unit basis vectors e_1, e_2) to a parallelogram (spanned by the images of the basis vectors $L(e_1)$ and $L(e_2)$). So, ratio by which L scales areas in the plane is captured by the *area of the parallelogram spanned by $L(e_1)$ and $L(e_2)$*. How can we find this area? It helps to draw a picture of the parallelogram we want. If $L = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, then L sends the first basis vector to $\langle a, c \rangle$ and the second to $\langle b, d \rangle$:

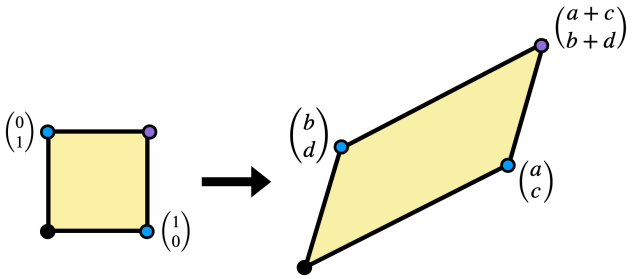


Figure 7.7.: The determinant measures the change in area under a linear map.

We can actually find this area in a pretty satisfying way using just what we've proven about Euclidean geometry so far. We know the areas of squares, rectangles, and right triangles, so let's try to write the area we are after as a difference of things we know:

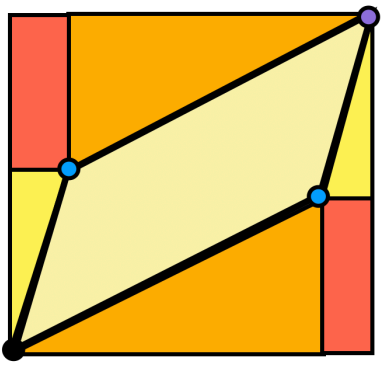


Figure 7.8.: A formula for the determinant can be found knowing only the area of squares, rectangles, and right triangles. (I learned this awesome diagram from Prof Daniel O'Connor!)

Exercise 7.5. Show the area of the parallelogram spanned by $\langle a, c\rangle$ and $\langle b, d\rangle$ is $ad-bc$, using the Euclidean geometry we have done, and the diagram above.

Definition 7.11 (Determinant)**.** The determinant of a linear transformation $M = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ is

$$\det M = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

Thus, the quantity we saw in the definition of the 2×2 matrix inverse was just $1/\det$. This makes sense: if L scales up the area by a certain factor, then its inverse must undo that scaling, it must scale by the reciprocal!

Theorem 7.1 (Invertibility & The Determinant)**.** *A linear transformation is invertible if and only if its determinant is nonzero.*

This theorem lets us think of the determinant as a tool to detect invertibility. If the determinant is zero, then the linear transformation takes a square to something of zero area: a point, or a line segment! And then information has been lost - the square has been crushed onto a smaller dimensional space - and there's no undoing that.

So far we've figured out the meaning of the determinant when it is a *positive number*. But it can also be negative: what does it mean to scale area by a negative number? It's easiest to see via an example - the matrix $\left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ has determinant -1, and it flips an image upside down across the x-axis. This is the meaning of a negative determinant - a reflection!

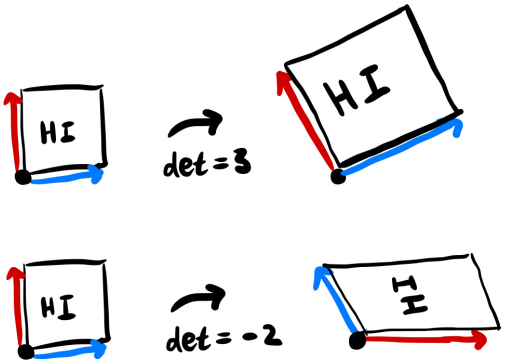


Figure 7.9.: Determinants: the first map expands the area by a factor of three, and the second map expands by a factor of two but also *reverses orientation*, reflecting the image.

We often refer to this concept formally with the term *orientation*. We say a function is *orientation preserving* if it does not reflect, or flip an image, and *orientation reversing* if it does. Thus, the determinant is not only an invertibility detector, but an orientation detector as well.

Definition 7.12 (Orientation Preserving)**.** A linear transformation is *orientation preserving* if its determinant is a positive number.

8. Zooming In

8.1. Single-Variable Calculus

Given this new picture of *where* the infinitesimals of calculus live, its helpful to briefly turn our gaze backwards and consider the calculus we already know in a new light. Instead of drawing a function $f(x)$ as a *graph* on x and y axes, we will start by thinking of it as a *rule*, telling us how to move around points on the line. Here's a depiction of $y = x^2$ from this perspective.

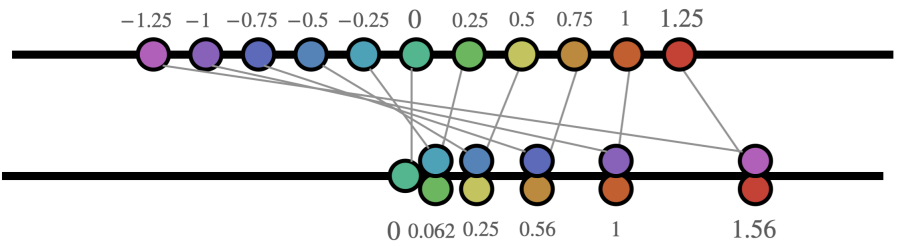


Figure 8.1.: The function $f(x) = x^2$ as a rule taking points on the line to other points on the line, visualized for points between -1.25 and 1.25.

This may be a bit hard to interpret at first, mostly because of all the crossing lines: the *squaring* operation folds the line in half, sending all the negative numbers to positive numbers, which clutters our view. The same point can be made more clearly with a function that does not do this, such as $y = x^3$:

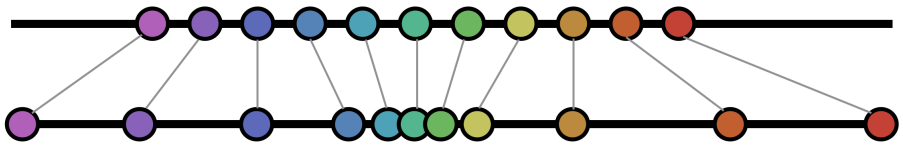


Figure 8.2.: The function $f(x) = x^3$ as a mapping from the line to itself.

It's quite easy to see from this map that our function $f(x) = x^3$ is *stretching* the line at some points, and *compressing it* at others. The gray lines connecting inputs to

outputs guide our eyes in this qualitative judgement, where we see that points very near the origin are getting pulled closer and closer together, whereas points further out are getting pulled apart.

But this is all an analysis of *finite points* along the line: what we are really interested in of course, is the infinite level of zoom required by calculus. To see this, we need to imagine the infinitesimal tangent spaces at each point. Below, I've illustrated this near a point undergoing infinitesimal stretch, as well as a point undergoing infinitesimal compression.

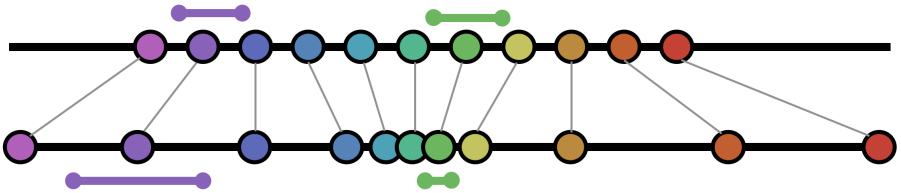


Figure 8.3.: The effect of a function on tangent spaces at each point is a linear *stretch* or *compression*

After passing to the tangent space, we expect (via the *Fundamental Strategy*) that our function becomes a linear function. But the tangent spaces are just lines, and whats a linear map from a line to a line? It's just multiplication by a constant (a 1×1 matrix...). Which constant? The derivative, of course! For the example at hand we have

$$f'(x) = 3x^2$$

Here we interpret the derivative not as a *slope*, but as the infinitesimal stretch factor: the fact that $f'(1) = 3 \cdot 1^2 = 3$ means that near the point 1, infinitesimal lengths are being expanded by a factor of 3. The fact that $f'(0.1) = 3 \cdot (0.1)^2 = 0.03$ means that near the point 1, distances are stretched by 0.03 - that is, compressed by a factor of 33!

It would be great to have a good mental picture of this before we go too far into the weeds. And we are incredibly fortunate that 3Blue1Brown has anticipated our needs, and produced a beautiful video on this topic! This is his final installment in the series "Essence of Calculus", and while it is the one most relevant to our course (the series focuses on the concepts of Calculus 1 and 2) I wholeheartedly recommend taking some time to refresh your knowledge by watching the entire thing!

https://youtu.be/CfW845LNObM?feature=shared

8.2. Linearizing Curves

Now we know how to linearize space, how do we find the linearizations of functions at the infinite level of zoom we desire? Its perhaps easiest to start with *curves*. Curves are functions from some interval $I \subset \mathbb{R}$ into $\mathbb{R}^2$ or $\mathbb{R}^3$. Thus we can write them with *components*, like $\gamma(t) = (x(t), y(t))$.

Proposition 8.1 (Differentiating Curves)**.** *Let $\gamma : \mathbb{R}^2$ be a curve. Then at a fixed time t (and thus a specified point $\gamma(t)$), the linearization of the curve is given by the vector*

$$\gamma'(t) = \lim_{\epsilon \to 0} \frac{\gamma(t+\epsilon) - \gamma(t)}{\epsilon}$$

in the tangent space $T_{\gamma(t)}\mathbb{R}^2$. If this limit does not exist, the curve is said to be not differentiable at that point.

First, we check that this definition makes sense. If ϵ is some small (but finite) number, the points $\gamma(t+\epsilon)$ and $\gamma(t)$ are two points along the curve, very near to each other. Their difference is a *vector* based at $\gamma(t)$! Taking the limit as $\epsilon \to 0$ makes this vector shrink to zero length, but rescaling by $1/\epsilon$ lets us *zoom in*, and the result is a *tangent vector based at $\gamma(t)$!

Exercise 8.1. Show that if we write the curve $\gamma(t) = (x(t), y(t))$ in coordinates, that we can use the rules of vector addition and scalar multiplication to simplify this calculation. Indeed, the tangent vector at $\gamma(t)$ is just given by the derivatives of the coordinate functions

$$\gamma'(t) = \langle x'(t), y'(t) \rangle$$

Geometrically, we should interpret this derivative as being a way of taking an infinitesimal piece of the t line based at the point t, and placing it into the tangent space at $\gamma(t)$:

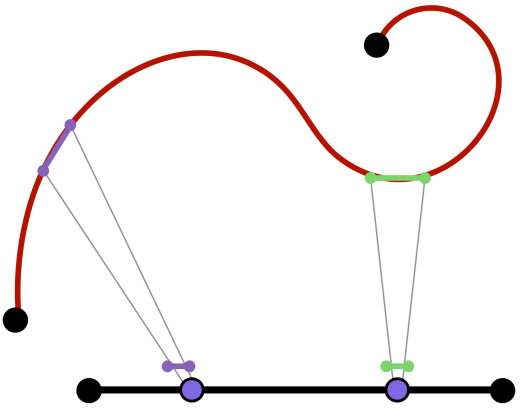


Figure 8.4.: The derivative to a curve is a linear map taking infinitesimal pieces of the line onto infinitesimal pieces of the curve.

Example 8.1. The curve $f(t) = (t, t^2)$ traces out the standard parabola in the plane. This passes through the point $(2, 4)$ when $t = 2$. The derivative of f is the function $f'(t) = \langle 1, 2t \rangle$, so the tangent vector to f when $t = 2$ is the vector

$$f'(2) = \langle 1, 4 \rangle$$

in the tangent space $T_{(2,4)}\mathbb{R}^2$.

Exercise 8.2. Differentiate the following curves:

- Tangent vector to $(t, \sin(t))$ in $T_{(\pi/2,1)}\mathbb{R}^2$.
- Tangent vector to $f(t) = (\frac{1}{e^t - 1}, \sqrt{t+1})$ when $t = 2$. Which tangent space is it in?

8.3. Linearizing Multivariable Functions

Besides curves, the other main type of function we will be interested in are functions from a 2-dimensional space *back to itself.* These are things like rotations of the plane, translations of the plane, but also include even weirder things, that move points about the plane in strange ways.

Definition 8.1 (Multivariable Function)**.** A function ϕ from the plane to itself is a function whose input is a point $p \in \mathbb{R}^2$ and its output is another point in the same space: $\phi(p) \in \mathbb{R}^2$.

We can make this more concrete by writing both the domain and the range in coordinates: since $p \in \mathbb{R}^2$ we can write $p = (x, y)$ for two numbers x, y. Thus, we can write $\phi(p) = \phi(x, y)$. But since $\phi(x, y)$ is *also* in the plane, we can write it in components as well, say $\phi(x, y) = (a, b)$. Since the output *depends on x, y we see that the coordinate a is a function of both x and y, as is b. Thus its more helpful to write them as $a(x, y)$ and $b(x, y)$ to remember this.

Definition 8.2. If ϕ is a function from the plane to itself, we can write it in components as two separate real-valued functions of x and y. Often to aid in readability, we name the *component functions* with the same letter a the overall function:

$$\phi(x, y) = (\phi_1(x, y), \phi_2(x, y))$$

What should the linearization of such functions look like upon zooming in? Well, we already know how *curves* work, so a good place to start is by looking for curves. If we hold x constant in the domain, we get a line parallel to the y axis through p. Similarly, holding y constant we get a line parallel to the x axis through P. Plugging these into ϕ, we get two curves passing through $\phi(p)$.

$$\text{curve}_1(x) = (\phi_1(x, b), \phi_2(x, b))$$

$$\text{curve}_2(y) = (\phi_1(a, y), \phi_2(a, y))$$

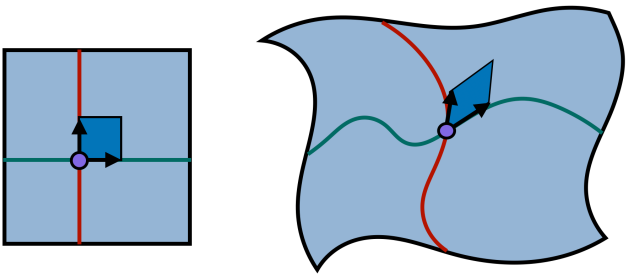


Figure 8.5.: Understanding a multivariate function by looking at curves through a point, and their linearizations.

Zooming in, the linearization of these curves are two vectors in $T_{\phi(p)}\mathbb{R}^2$, which we can compute explicitly in coordinates:

$$v_x = \text{curve}_1{}'(x) = \begin{pmatrix} \frac{\partial}{\partial x}\phi_1(x, b) \\ \frac{\partial}{\partial x}\phi_2(x, b) \end{pmatrix}$$

$$v_y = \text{curve}_2'(y) = \begin{pmatrix} \frac{\partial}{\partial y}\phi_1(a,y) \\ \frac{\partial}{\partial y}\phi_2(a,y) \end{pmatrix}$$

These are each vectors that lie in the tangent space $T_{\phi(p)}\mathbb{R}^2$, and so they span a parallelogram there. Indeed, we see that upon zooming in, the map ϕ seems to take an infinitesimal square with sides $\langle 1,0\rangle$, $\langle 0,1\rangle$ based at $T_p\mathbb{R}^2$ to an infinitesimal parallelogram in the range's tangent space.

ANIMATION

This is the behavior of a *linear map*! And even better, we know exactly how to write down a linear map as a matrix if we are given what it does to the standard basis!

Remark 8.1. Here the symbol ∂ denotes a *partial derivative*: that means, we treat all the other variables as constants, and only differentiate the specified one. It's easiest to see via example, $\frac{\partial}{\partial x}(3x^2y + ax + by) = 6xy + a + 0$ where here we have treated y, a and b as constants, and taken only the derivative with respect to x

Definition 8.3. Let $\phi : \mathbb{E}^2 \to \mathbb{E}^2$ be a multivariable function, written in components as

$$\phi(x,y) = (\phi_1(x,y), \phi_2(x,y))$$

Then at a point $p \in \mathbb{E}^2$, the *derivative* of ϕ at a point p is a 2×2 matrix given by the x and y derivatives of its two component functions:

$$D\phi_p = \begin{pmatrix} \frac{\partial\phi_1}{\partial x} & \frac{\partial\phi_1}{\partial y} \\ \frac{\partial\phi_2}{\partial x} & \frac{\partial\phi_2}{\partial y} \end{pmatrix}$$

Where after taking the derivatives, we plug in the point p to each entry of the matrix, to get a matrix of numbers. This is a linear map from the tangent space $T_p\mathbb{R}^2$ to the tangent space $T_{\phi(p)}\mathbb{R}^2$.

To lighten notation, sometimes we will just write ∂_x for $\frac{\partial}{\partial x}$, and we will write a vertical bar for evaluation, much as in calculus:

Example 8.2. The derivative of $\phi(x,y) = (x+y, \frac{x}{y})$ at the point $(4,7)$ is given by

$$D\phi = \begin{pmatrix} \partial_x(x+y) & \partial_y(x+y) \\ \partial_x(x/y) & \partial_y(x/y) \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1/y & -x/y^2 \end{pmatrix}$$

Plugging in the point,

$$D\phi_{(4,7)} = \begin{pmatrix} 1 & 1 \\ 1/y & -x/y^2 \end{pmatrix}\Bigg|_{(4,7)} = \begin{pmatrix} 1 & 1 \\ 1/7 & -4/49 \end{pmatrix}$$

Because ϕ takes the point $(4, 7)$ to the point

$$\phi(4, 7) = (4 + 7, 4/7) = (11, 4/7)$$

this is a linear map from $T_{(4,7)}\mathbb{R}^2$ to $T_{(11,4/7)}\mathbb{R}^2$

The usual calculus rules hold: differentiation of a sum of functions is a sum of their derivative matrices, and you can pull scalars out from the derivative

Exercise 8.3. Find the derivatives of the following functions, at the specified points.

- The function $f(x, y) = (xy, x + y)$ at the point $p = (1, 2)$.
- The function $\phi(x, y) = \left(xy^2 - 3x, \frac{x}{y^2+1}\right)$ at the point $q = (3, 0)$.

8.3.1. Compositions

In single variable calculus, we often made use of the *chain rule* to take derivatives. This let us remember less things, as we were able to *construct* derivatives of complicated functions from simpler pieces. It's instructive to take a look back at the formula:

$$(f \circ g(x))' = f'(g(x))g'(x)$$

What is this saying in our new language of linearizations? Recall that the number line itself has tangent spaces, just like the plane, and we should interpret something like $g'(x)$ as saying *the linearization of g at x*. From this perspective, in words this says

> The linearization of $f \circ g$ at x is the result of linearizing g at x, and multiplying by the linearization of f at $g(x)$.

This makes perfect sense geometrically, where we start with an infinitesimal piece of the line based at x, apply g so it gets stretched by a factor of $g'(x)$, and moved to be located at $g(x)$. Then we apply f: this further stretches by a factor of $f'(g(x))$! So the multiplication we see in the formula is really a *composition*: its saying *first* stretch by g, and *then* stretch by f.

This has a direct analog in higher dimensions, if we remember that the way to compose linear transformations is by matrix multiplication.

Proposition 8.2 (Differentiating Compositions)**.** *If F, G are both transformations of the plane, the derivative of $F \circ G$ at the point p is the composition of the derivative of G at p with the derivative of F at $G(p)$:*

$$D(F \circ G)_p = DF_{G(p)} DG_p$$

Example 8.3. If $G(x, y) = (xy, x + y)$ and $F(x, y) = (2x - y, xy)$ then we compute the derivative of $F \circ G$ at $(1, 2)$ as follows: First, we find the derivative of G at $(1, 2)$:

$$DG_{(1,2)} = \begin{pmatrix} y & x \\ 1 + y & 1 + x \end{pmatrix}\Bigg|_{(1,2)} = \begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix}$$

Then, since G takes $(1, 2)$ to the point $G(1, 2) = (1 \cdot 2, 1 + 2) = (2, 3)$, we need the derivative of F at $(2, 3)$:

$$DF_{(2,3)} = \begin{pmatrix} 2 - y & -1 \\ y & x \end{pmatrix}\Bigg|_{(2,3)} = \begin{pmatrix} -1 & -1 \\ 3 & 2 \end{pmatrix}$$

Finally, we compose these linear maps with matrix multiplication (making sure to be careful about the order!)

$$DF_{(2,3)} DG_{(1,2)} = \begin{pmatrix} -1 & -1 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix} = \begin{pmatrix} -5 & -3 \\ 12 & 7 \end{pmatrix}$$

Exercise 8.4. If F, G, H are the following multivariate functions

$$F(x, y) = (x - y, xy)$$

$$G(x, y) = (-y, x)$$

$$H(x, y) = (x^3, y^3)$$

Differentiate the following compositions:

- $F \circ G$ at $(1, 1)$
- $G \circ G$ at $(0, 2)$
- $F \circ G \circ H$ at $(-1, 3)$.

8.3.2. Inverses

Inverse of F is a function H with $H(F(x)) = x$, which 'undoes' the action of F at each point. Familiar one dimensional examples include squaring and the square root, exponentials and logarithms, as well as trigonometric functions and their *arc*-versions (sin and arcsin for example). One nice consequence of the chain rules is that it's possible to differentiate an inverse function, even if you don't have an explicit formula for it!

The same reasoning applies directly in higher dimensions: if F is a multivariable function with inverse H, then the composition $HF = I$ is the identity function, sending every point p to itself. This is straightforward to differentiate: if $I(x, y) = (x, y)$ then

$$DI = \begin{pmatrix} \partial_x x & \partial_y x \\ \partial_x y & \partial_y y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Note that the matrix we got by differentiating is *constant* - it has no x's or y's in it: thus this matrix represents the derivative of the identity function at every point in the plane. Now, using the multivariate chain rule we can differentiate the equation $HF = I$ to get

$$D(HF)_a = DH_{F(a)}DF_a = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

But this says that the two matrices, DH at $F(a)$ and DF at a multiply to give the *identity matrix*! This is the definition of being *inverse matrices*, so we have

Theorem 8.1. *If F is an invertible multivarible function, its inverse function H has the following derivative:*

$$DH_{F(a)} = (DF_a)^{-1}$$

Note that this theorem only tells us how to find the derivative at the point $F(a)$: to find it at a point p we want, we need to do some more work, and figure out which point F sends to p.

Example 8.4. Let $F(x, y) = (x^3, x^2y)$, and let H be its inverse where defined. To find the derivative of H at $(8, 4)$, we first find the derivative matrix of F

$$DF = \begin{pmatrix} \partial_x(x^3) & \partial_y(x^3) \\ \partial_x(x^2y) & \partial_y(x^2y) \end{pmatrix} = \begin{pmatrix} 3x^2 & 0 \\ 2xy & x^2 \end{pmatrix}$$

Then, we invert this, using the formula for 2×2 matrix inverses

$$(DF)^{-1} = \frac{1}{3x^2 \cdot x^2 - 0 \cdot 2xy} \begin{pmatrix} x^2 & 0 \\ -2xy & 3x^2 \end{pmatrix}$$
$$= \frac{1}{3x^4} \begin{pmatrix} x^2 & 0 \\ -2xy & 3x^2 \end{pmatrix}$$

By Theorem 8.1, this is the derivative of the inverse H at the point $F(x, y)$. We want to find the derivative at $(8, 4)$, so we need to know which values of (x, y) to plug in. That is, we need to solve for which (x, y) satisfies $F(x, y) = (8, 4)$. This is a system of equations:

$$F(x, y) = \begin{pmatrix} x^3 \\ x^2 y \end{pmatrix} = \begin{pmatrix} 8 \\ 4 \end{pmatrix}$$

The first equation tells us that $x = 2$, as that's the only real number that cubes to 8. Now we can plug this into the second equation, which says $2^2 y = 4$, so $y = 1$. Plugging in $(2, 1)$ gives the result

$$DH_{(8,4)} = (DF_{(2,1)})^{-1} = \frac{1}{3 \cdot 2^4} \begin{pmatrix} 2^2 & 0 \\ -2 \cdot 2 \cdot 1 & 3 \cdot 2^2 \end{pmatrix}$$
$$= \frac{1}{48} \begin{pmatrix} 4 & 0 \\ -4 & 12 \end{pmatrix}$$

As a review of Calculus I, try this out for a function of a single variable yourself:

Exercise 8.5. Consider the function $f(x) = \arccos(x)$ what is its derivative at $x = \frac{1}{2}$?

8.3.3. Differentiating Linear Maps

We won't actually have that many opportunities during the course where we will need to find the explicit derivatives of an inverse function as we did above. But during the proof of Theorem 8.1, we noticed an interesting fact: the derivative of many maps we have calculated depends on which point (x, y) we were differentiating them at. But not the map $I(x, y) = (x, y)$: its' derivative was a constant! If you look at our computation you'll notice this can certainly be generalized: for instance the derivative of $f(x, y) = (2x + y, x - y)$ is a constant matrix for the same reason.

Proposition 8.3 (Derivative of a Linear Map)**.** *If ϕ is a linear map, then $D\phi$ is constant, and equal to ϕ.*

Exercise 8.6. Prove Proposition 8.3.

While the symbolic proof of this is relatively straightforward, its good to pause for a minute and contemplate what it *means*. The derivative at a point is supposed to be *the best linear approximation to the function at that point.* But what happens if the function is already linear? Well - then the best linear approximation at that point is *itself*! And this is true at every point - so the derivative is the same as the map at every point!

In symbols, if M is a matrix and our linear function is $f(p) = Mp$, then the derivative is $Df_p = M$. We ar already very familiar with this from single-variable calculus, although perhaps we did not think through the meaning carefully at the time. After all, what is the derivative of the linear function $y = mx$? Its the *constant* $y = m$: which is just saying that infinitesimally near every x, the function $y = mx$ is scaling things up by a factor of m.

Can we characterize which maps have this property? If $\phi = (\phi_1(x, y), \phi_2(x, y))$, when is it the case that $D\phi_{(x,y)}$ is a constant matrix?

Exercise 8.7 (When the derivative is constant)**.** Prove that a function $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ has a constant derivative if and only if the function is *affine*: that is, a linear map plus constants.

9. Zooming Out

9.1. One Variable Integration

The quintessential 'zoom out' technique in mathematics is integration. It allows us to add up, or *integrate together* a continuum of infinitesimally small changes into a single finite change. While its definition is in terms of a *limit* (a Riemann sum, as we reviewed in the chapter on the Fundamental Strategy) the true power of calculus is that we do not need to compute this limit, but instead can antidifferentiate!

Remark 9.1. In calculus classes we often write integration over an interval $[a, b]$ by putting the bounds on the top and bottom of the integration sign, like $\int_a^b$. You are welcome to continue using this notation, however I will sometimes opt to put the *entire interval* in the subscript as $\int_{[a,b]}$ This fits better with notation for double integrals like $\iint_R$ and other generalizations, where the domain usually appears as a subscript.

Theorem 9.1 (The Fundamental Theorem of Calculus)**.** *Let f be a function defined on $[a, b]$, and F be an* antiderivative *of f - that is, a function such that $F'(x) = f(x)$. Then we may integrate f using this antiderivative:*

$$\int_{[a,b]} f(x)dx = F(b) - F(a)$$

Because of this, we will use the *indefinite integral* $\int f dx$ as a notation for the collection of antiderivatives of f. In this class we'll assume familiarity with the 1-dimensional integral as seen in a Calculus I and II course. That means, we'll be free to use antidifferentiation, u-substitution, integration by parts, etc where helpful.

Exercise 9.1. Compute the following integrals, as a refresher of your calculus skills:

$$\int \sin(2q - 3)dq \qquad \int \frac{x}{x+1}dx$$
$$\int y^2 e^{y^3} dy \qquad \int t^2 e^t dt$$

Besides calculation, theoretical properties of the integral will also be useful in helping us prove things. Two of fundamental properties of the integral are below.

Proposition 9.1 (Subdividing Intervals)**.** *If f is an integrable function on the interval $[a,b]$ and c is some point inside the interval (that is, $a < c < b$), then*

$$\int_{[a,b]} f dx = \int_{[a,c]} f dx + \int_{[c,b]} f dx$$

When we interpret the integral as area, this theorem is is one of the greek *area axioms* - but is now not an assumption but rather something we can prove! There's one other property of the integral that is rather straightforward from its interpretation as area: an integral of a function that has some positive area, but no negative area to cancel it out must be positive!

Proposition 9.2 (Integrating Positive Functions)**.** *Let f be a continuous function, and $[a,b]$ an interval.*

- *If $f(x) \geq 0$ for all x, then $\int_a^b f(x)dx \geq 0$.*
- *If $f(x) > 0$ for all x, then $\int_a^b f(x)dx > 0$.*

As a consequence of this, if we have a continuous function f which is nonnegative on an interval, and we know to be positive at some point, then we know the *integral of that function must be positive.* This will prove useful to us, so we'll separate it off as a corollary:

Corollary 9.1. *If f is continuous and nonnegative on $[a,b]$, and f is nonzero at* some point*, then*

$$\int_a^b f(x)dx > 0$$

Proof. Say f is nonnegative on an interval $[a,b]$, and is nonzero (so, necessarily positive) at some point c. Then since f is continuous there is some small interval $[l,r]$ around c where f is positive, and we can break our original interval into three pieces:

$$[a,b] = [a,l] \cup [l,r] \cup [r,b]$$

By Proposition 9.1, we can break the integral over $[a,b]$ into a sum of integrals over each of these three intervals:

$$\int_{[a,b]} f dx = \int_{[a,l]} f dx + \int_{[l,r]} f dx + \int_{[r,b]} f dx$$

The first and last of these are nonnegative by Proposition 9.2, since f is nonnegative on the whole interval. But the middle one is *strictly positive* as f is positive on the entire interval $[l,r]$. Thus the overall integral is a sum of a positive number and two others which are either positive or zero: the result is positive! And hence,

$$\int_{[a,b]} f dx > 0$$

□

9.2. Multi~Variable integration

If integrals are a means of 'zooming out' along a line, how do we zoom out in the plane? We need a higher dimensional analog of the integral, a *double integral*

Definition 9.1 (Double Integral Riemann Sum)**.**

Are we going to need a whole new theory of calculus for this? Two dimensional Riemann sums, two dimensional integrals, and a two dimensional fundamental theorem? Happily no! It turns out much of two-dimensional integration can be summed up by saying "do one dimensional integration, but twice".

Proposition 9.3 (Fubini's Theorem)**.** *An integral over the plane can be computed as two one dimensional integrals, one for the x variable and one for the y:*

$$\int_{I\times J} f(x,y)dA = \int_I \left(\int_J f(x,y)dy\right)dx$$

Thus, there is nothing more to the theory of double integrals than doing a single-variable integral twice! It's easiest to see via example:

Example 9.1 (Iterated Integrals)**.** Let $R = [0,2]\times[0,3]$ be a rectangle in the x, y plane. To compute the integral $\iint_R xy + 1\, dA$, we write this as an integral for x from 1 to 2 and an integral of y from 0 to 3:

$$\int_{[0,2]} \left(\int_{[0,3]} xy + 1 dy\right) dx$$

We now compute the inside integral (with respect to y) first:

$$\int_{[0,3]} xy + 1 dy = x\frac{y^2}{2} + y\Big|_{y=0}^{y=3} = \frac{9}{2}x + 3$$

Then, we integrate this with respect to x:

$$\int_{[0,2]} \frac{9}{2}x + 3 dx = \frac{9}{4}x^2 + 3x\Big|_{x=0}^{x=2} = 15$$

Its even possible to have the bounds of the first integral contain the variables of the second integral:

Example 9.2. Compute the iterated integral below:

$$\int_0^1 \int_{x-3}^{x^2} x(2y+1)dydx$$

We begin with the inner integral, which is dy, so the x is (temporarily) a constant:

$$\int_{x-3}^{x^2} x(2y+1)dy = x\left(y^2+y\right)\Big|_{x-3}^{x^2}$$
$$= x\left((x^2)^2+(x^2)\right) - x\left((x-3)^2+(x-3)\right)$$
$$= x^5 - 5x^2 - 6x$$

Now we've finished the inner integral, and we need to proceed to the next one:

$$\int_0^1 x^5 - 5x^2 - 6xdx = \frac{x^6}{6} - \frac{5}{3}x^3 - 3x^2\Big|_0^1$$
$$= \frac{1}{6} - \frac{5}{3} - 3 = -\frac{9}{2}$$

Exercise 9.2 (Iterated Integrals)**.** For practice, compute the following iterated integrals.

9.3. Power Series

Besides integration, the other *zoom-out* type technique we saw time and again in introductory calculus was the construction of a power series from the derivatives of a function. Power series constructed this way are often called *Taylor Series*.

Remark 9.2. Named after Brook Taylor, who introduced them in 1715. However many such series were known earlier, used in the works of Issac Newton in the 1600s, and Madhava in the 1300s

Definition 9.2 (Power Series: Taylor's Version)**.** A power series is an infinite series of the form $\sum_{n=0}^{\infty} a_n x^n$ for some constants a_n. If $f(x)$ is a function, the *Taylor series for* f is a power series that represents the function $f(x)$ in terms of its derivatives at $x = 0$:

$$f(x) = f(0) + f'(0)x + f''(0)\frac{x^2}{2} + f'''(0)\frac{x^3}{3!} + \cdots$$
$$= \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n} x^n$$

Example 9.3 (Power Series for e^x)**.** Because the derivative of e^x is itself, and $e^0 = 1$, every derivative of e^x at $x = 0$ is equal to 1, and its power series is

$$e^x = \sum_{n=0}^{\infty} \frac{1}{n!} x^n$$

One of the reasons that power series are such a powerful tool in calculus is the ability to *do math with them*: we can treat them like any other function; composing them with other functions, differentiate them and integrate them!

Example 9.4. Given that the power series for $\frac{1}{1-x}$ is $\sum x^n$, we can find the power series for $1/(1-2x^2)$ by substituting $2x^2$ for x:

$$\frac{1}{1-2x^2} = \sum_{n=0}^{\infty} (2x^2)^n = \sum_{n=0}^{\infty} 4^n x^{2n}$$

Proposition 9.4 (Calculus With Power Series)**.** *Given a power series $f(x) = \sum_n a_n x^n$ we can differentiate and integrate the series term-by-term:*

$$f'(x) = \sum_{n=0}^{\infty} a_n (x^n)' = \sum_{n=0}^{\infty} n a_n x^{n-1}$$

$$\int f dx = \sum_{n=0}^{\infty} a_n \left(\int x^n dx \right) = \sum_{n=0}^{\infty} \frac{a_n}{n+1} x^{n+1}$$

Example 9.5 (The power series for $\arctan(x)$)**.** Given the power series $\frac{1}{1-x} = \sum x^n$, we can create the power series for $\frac{1}{1+x^2}$ by substituting $-x^2$ for x:

$$\frac{1}{1+x^2} = \sum_{n=0}^{\infty} (-x^2)^n = \sum_{n=0}^{\infty} (-1)^n x^{2n}$$

Now, since $\frac{1}{1+x^2}$ is the derivative of $\arctan(x)$, we need only antidifferentiate this

series term by term to find the Taylor series for the arctangent:

$$\begin{aligned}\arctan(x) &= \int \frac{1}{1+x^2}dx \\ &= \int \sum_{n=0}^{\infty}(-1)^n x^{2n}\,dx \\ &= \sum_{n=0}^{\infty}(-1)^n \int x^{2n}dx \\ &= \sum_{n=0}^{\infty}(-1)^n \frac{x^{2n+1}}{2n+1} \\ &= x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7}\cdots\end{aligned}$$

Exercise 9.3. Find Power series for the following functions

All of these techniques make power series a very useful tool indeed. But of course those of you who remember Calculus 2 well know that we have so far left out an important and subtle piece of the story: when do power series work at all? Series don't always converge, and to tell when they do we have a variety of different *convergence tests* to help us out. Happily, all the series we will come across are power series, where checking convergence is straightforward.

Theorem 9.2 (Radius of Convergence)**.** *If $f(x) = \sum a_n x^n$ is a power series, let*

$$\alpha = \lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right|$$

Then f converges by the ratio test at x if $|\alpha x| < 1$, or $|x| < \frac{1}{\alpha}$.

Remark 9.3. Warning: not all functions have power series, and those that do are called *analytic*. Happily all functions we will encounter in this course are analytic, so we can push this concern to the back of our minds

This value $R = \frac{1}{\alpha}$ is called the *radius of convergence*. Many of the series that will be of use to us in this class (sine, cosine, and their hyperbolic counterparts) converge on the entire real line, and we will not have to worry about such things.

Part III.

The Plane

10. Foundations

Definition 10.1 (Points of $\mathbb{E}^2$)**.** The points of the Euclidean plane are pairs $p = (x, y)$ of real numbers: that is

$$\mathbb{E}^2 = \{(x, y) \mid x, y \in \mathbb{R}\} \cong \mathbb{R}^2$$

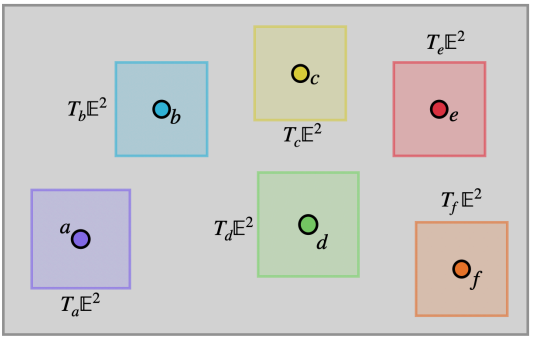


Figure 10.1.: The Plane is built out of the points of $\mathbb{R}^2$, together with a tangent space at each point.

We use the notation $\mathbb{E}^2$ for the *geometry* even though the underlying point set is just the plane $\mathbb{R}^2$ This is because ordered pairs of real numbers can represent many different things (see **?@sec-maps**) and we wish to make it clear here that right now we mean their original usage, to describe the geometry of Euclid.

Definition 10.2 (Vectors of $\mathbb{E}^2$)**.** At each point p of the Euclidean plane, the set of tangent vectors is *another copy* of $\mathbb{R}^2$.

$$T_p\mathbb{E}^2 = \{(v_1, v_2) \mid v_i \in \mathbb{R}\} \cong \mathbb{R}^2$$

Vectors are just pairs of real numbers as we are used to, but we do need to be careful about keeping track of where they are based. Hence, we will often write a subscript on a vector to denote where it lives: $\langle 1, 2\rangle_{(1,5)}$ is the vector $\langle 1, 2\rangle$ based at $(1, 5) \in \mathbb{E}^2$, whereas $\langle 1, 2\rangle_{(-2,3)}$ is the vector with teh same coordinates, but based at $(-2, 3)$.

The **origin** is the point with coordinates $(0, 0)$ in the plane. Because these zeroes will make some calculations easier, we will often find ourselves doing things at the origin,

and so its useful to have a shorthand notation. We will write $O = (0, 0)$ for this point, and denote vectors based at O with the subscript v_o, consistent with the above.

Now we have a precise definition of what the Euclidean plane is made out of (points) and its infinitesimal pieces (vectors), so we can precisely define things like *curves* and their *tangents*.

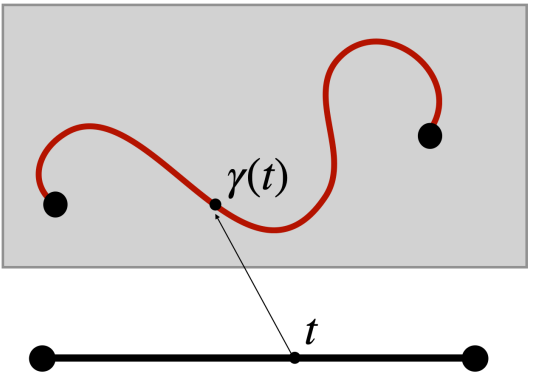


Figure 10.2.: A curve is a function from an interval in $\mathbb{R}$ into the Plane.

Definition 10.3 (Curves)**.** A *curve* in the Euclidean plane is a function $\gamma : I \to \mathbb{E}^2$ for I an interval in the real line (or possibly all of $\mathbb{R}$). The *tangent* to $\gamma(t) = (x(t), y(t))$ at time t_0 is its coordinate-wise derivative

$$\gamma'(t_0) = \langle x'(t_0), y'(t_0) \rangle \in T_{\gamma(t_0)}\mathbb{E}^2$$

A curve is *regular* if its derivative is never equal to the zero vector for ant $t \in I$.

But to make real progress, we need the tools to be able to *measure length*.

10.1. Length of Curves

Our new formulation of geometry puts all curves on an equal footing - an allows us to measure their lengths using the ideas of calculus. This was the dream of Archimedes, realized only nearly two millennia after his death.

Idea: infinitesimally, geometry looks like what was studied by the greeks, as if you zoom in on any curve it appears as a line. To impose this fact on our new geometry we will measure *infinitesimal distances* via the pythagorean theorem. This will be our only geometric axiom - from this alone (together with the tools of calculus) we will rebuild all of geometry.

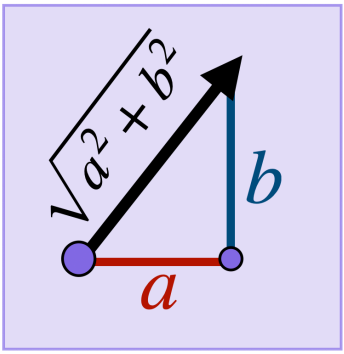


Figure 10.3.: The fundamental axiom of the plane - the Pythagorean Theorem is true infinitesimally, in each tangent space.

Definition 10.4 (Infinitesimal Length in $\mathbb{E}^2$)**.** If v is a tangent vector based at $p \in \mathbb{E}^2$ then its infinitesimal length is given by the pythagorean theorem on the tangent space $T_p\mathbb{E}^2$.

$$\|v\| = \sqrt{v_1^2 + v_2^2}$$

To measure a curve we take inspiration from Archimedes' *measurement of the circle* and approximate it with small line segments. A curve is called *rectifiable* if these approximate lengths converge as their number tends to infinity. It's a consequence of calculus that **regular curves are rectifiable.**

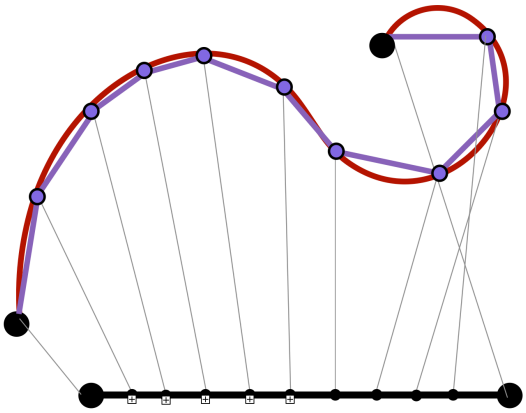


Figure 10.4.: The spirit of Archimedes: we can figure out how to define the length of a curve by thinking about approximations to it.

Then we define the length of real curves by *zooming out* (integrating) their zoomed-in

(differential) lengths. Unlike in Euclid's formulation, now *all curves* are on an equal footing: all lengths are determined by infinitesimal integration!

Definition 10.5 (Length in $\mathbb{E}^2$)**.** If γ is a curve which is differentiable, then we can measure the length of $\gamma(t)$ from $t = a$ to $t = b$ by integrating the infinitesimal lengths of its tangent vectors:

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

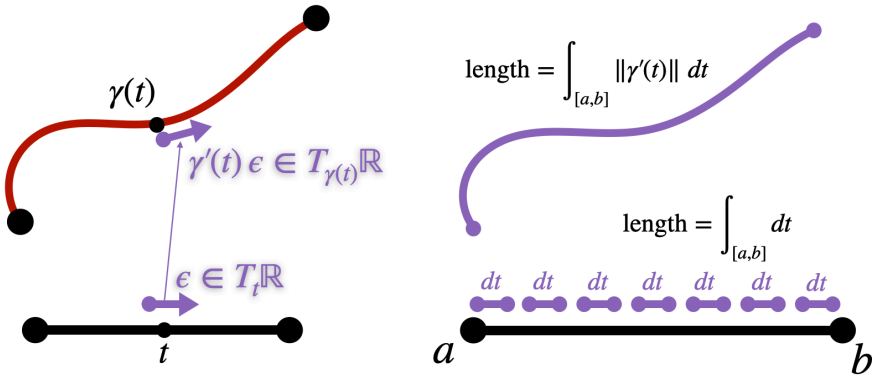


Figure 10.5.: The derivative $\gamma'(t)$ is a linear map taking an infinitesimal vector at t to an infinitesimal piece of arc at $\gamma(t)$. Integrating the lengths of these infinitesimal vectors is how we *define* the length of a curve.

Its helpful to write this definition out in full: if $\gamma(t) = (x(t), y(t))$ then $\gamma'(t) = \langle x'(t), y'(t) \rangle$ and so $\|\gamma'(t)\| = \sqrt{x'(t)^2 + y'(t)^2}$. Thus

$$\text{length}(\gamma) = \int_a^b \sqrt{x'(t)^2 + y'(t)^2} dt$$

Such integrals can be difficult to do in practice because of that nasty square root that shows up in their definition. And when they are possible, these often need several calculus tricks to succeed:

Exercise 10.1 (The Length of a Parabola)**.** Find the length of the parabola $y = x^2$ between from $x = 0$ to $x = a$, following the steps below.

- Parameterize the curve as $c(t) = (t, t^2)$, show the arclength integral is $L(a) = \int_{[0,a]} \sqrt{1 + 4t^2}$

- Perform the trigonometric substitution $x = \frac{1}{2}\tan\theta$ to convert this to some multiple of the integral of $\sec^3(\theta)$.
- Let $I = \int \sec^3(\theta)d\theta$ and do integration by parts with $u = \sec\theta$ and $dv = \sec^2\theta$.
- After parts, use the trigonometric identity $\tan^2\theta = \sec^2\theta - 1$ in the resulting integral to get another copy of $I = \int \sec^3\theta d\theta$ to appear.
- Get both copies of I to the same side of the equation and solve for it! To check your work at this stage, you should have found that

$$\int \sec^3\theta d\theta = \frac{1}{2}\sec\theta\tan\theta + \frac{1}{2}\ln|\sec\theta + \tan\theta|$$

- Relate this back to your original integral, and undo the substitution $x = \frac{1}{2}\tan\theta$: can you use some trigonometry to figure out what $\sec\theta$ is?
- Finally, you have the antiderivative in terms of x! Now evaluate from 0 to a.

Our *main* use isn't to compute the lengths of a bunch of random curves. Instead, its more theoretical - the integral gives a *precise* definition for the length of any differentiable curve, and a simple definition at that! This will be extremely useful in building geometry back from our small foundations.

10.1.1. Parameterization Invariance

All seems well and good with this definition, but the mathematician in us should be a little worried: we defined the length of a curve in terms of a parameterization, but the curve itself doesn't care how we parameterize it!

To get a sense of this its easiest to look at an explicit example: below are four different curves which all trace out the same set of points in the plane: the segment of the x axis between 0 and 4.

$$\begin{aligned} \alpha(t) &= (t, 0) & t &\in [0, 4] \\ \beta(t) &= (2t, 0) & t &\in [0, 2] \\ \gamma(t) &= (t^2, 0) & t &\in [0, 2] \end{aligned}$$

Because these all describe the same set of points, we of course want them to have the same length! But our *definition* of the length function involves integrating infinitesimal arclengths (derivatives), and these curves don't all have the same derivative! Thus, to *really* make sure our definition makes sense, we need to check that it doesn't matter which parameterization we use, we will always get the same length.

Exercise 10.2. Check these three parameterizations of the segment of the x-axis from 0 to 4 all have the same length.

Two curves are said to have the same **image** if the set of points they trace out in the plane are the same. So, all three of the curves above have the same image, and the same length. This requires a bit more calculus to check in general, but remains true.

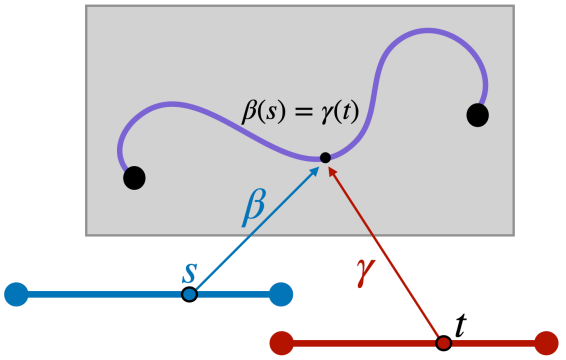


Figure 10.6.: Two curves β and γ with the same image.

Theorem 10.1 (Length & Parameterization Invariance)**.** *If β and γ are two curves with the same image, then*

$$\text{length}(\beta) = \text{length}(\gamma)$$

Proof. Let $\beta(s) : [a,b] \to \mathbb{E}^2$ and $\gamma(t) : [c,d] \to \mathbb{E}^2$ be two curves with the same image. That means they trace out the same set of points in the plane, so for every value of the parameter t for γ there is some value of s for β where $\beta(s) = \gamma(t)$. Write $s(t)$ for the function that does this - chooses the matching s parameter for each t.

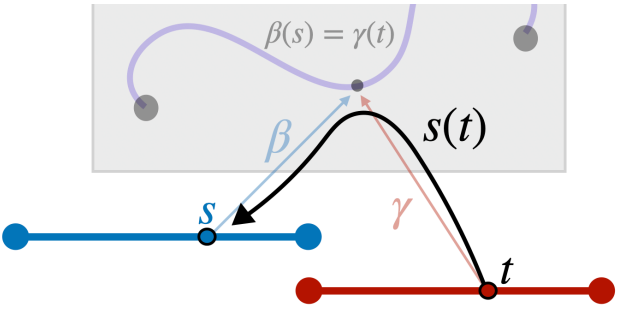


Figure 10.7.: The function $s(t)$ which takes the t parameter for the curve γ, and returns the s parameter for the curve β which maps to the same point in $\mathbb{E}^2$.

We can use this to wite the curve γ *in terms of β, as $\gamma(t) = \beta(s(t))$. We now calculate the length of γ using the definition:

$$\begin{aligned}\text{length}(\gamma) &= \int_{[c,d]} \|\gamma'(t)\| dt \\ &= \int_{[c,d]} \|\beta(s(t))'\| dt \\ &= \int_{[c,d]} \|\beta'(s(t))s'(t)\| dt\end{aligned}$$

Where we used the chain rule in the last step to differentiate the composition. Now, β' is a *vector*, but $s(t)$ was a scalar function so $s'(t)$ is a *scalar*, and we can pull it out of the norm, and do a u-substitution! Picking up from where we left off,

$$\begin{aligned}&= \int_{[c,d]} \|\beta'(s(t))\| s'(t) dt \\ &= \int_{[s(c),s(d)]} \|\beta'(s)\| ds \\ &= \int_{[a,b]} \|\beta'(s)\| ds \\ &= \text{length}(\beta)\end{aligned}$$

□

:::{#rem-curves} T here are some things we need to be careful on here: the curves β and γ are regular - their derivatives are never zero - which implies that they trace the curve from start to finish without stopping or doubling back. This, together with the fact that the curves are traced in the same direction implies that $s'(t)$ is always positive, which justifies the use of u-substitution. :::

11. Isometries

Besides measuring distance, one of the other most fundamental notions to geometry is that of an *isometry*, or a rigid motion of space. This comes from greek meaning *same-measure*, as an isometry is a function that does not change lengths.

Definition 11.1 (Isometries in $\mathbb{E}^2$)**.** An isometry of $\mathbb{E}^2$ is a function $\phi : \mathbb{E}^2 \to \mathbb{E}^2$ which preserves all infinitesimal lengths of $\mathbb{E}^2$.

What does it mean to preserve infinitesimal lengths? If $v \in T_p\mathbb{E}^2$ is a vector (an infinitesimal segment of a curve), then while ϕ takes p to a new point $\phi(p)$, infinitesimally it acts as a *linear transformation* from $T_p\mathbb{E}^2$ to $T_{\phi(p)}\mathbb{E}^2$. That infinitesimal linear transformation is the derivative matrix $D\phi_p$, which takes the original vector v to $D\phi_p(v)$. What we are interested in is whether or not $D\phi_p$ changed the length of v.

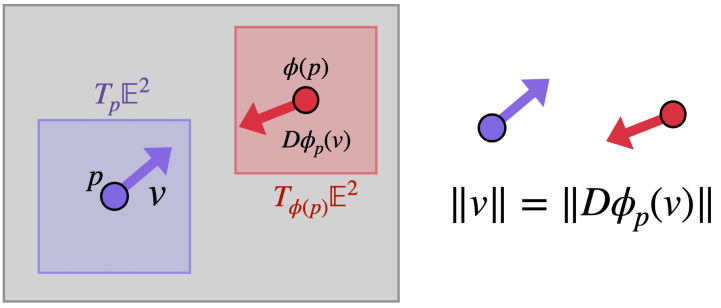


Figure 11.1.: An isometry does not change the length of any infinitesimal vector.

Definition 11.2. A function $\phi : \mathbb{E}^2 \to \mathbb{E}^2$ preserves infinitesimal lengths if for every $p \in \mathbb{E}^2$ and every $v \in T_p\mathbb{E}^2$, we have

$$\|v\| = \|D\phi_p(v)\|$$

Using this condition, one can show with some calculus that every isometry is actually an *invertible function*: that means, if ϕ is an isometry there is a function ϕ^{-1} which *undoes the action of* ϕ. *We will not prove this theorem here (as it is purely a result of advanced calculus, and doesn't help us learn geometry). If you like, you can think of this as an extra condition we are* assuming* about isometries in this course.

Theorem 11.1 (Isometries are Invertible Functions)**.**

Just as one can apply an isometry to points, one can apply it to an entire curve by *composition*: if γ is a curve, the curve $\phi \circ \gamma$ can be thought of as drawing γ, and then performing whatever action ϕ specifies.

Theorem 11.2 (Isometries Preserve Lengths of Curves)**.** *Let* $\phi : \mathbb{E}^2 \to \mathbb{E}^2$ *be an isometry, and* $\gamma : I \to \mathbb{E}^2$ *a curve. Then*

$$\text{length}(\gamma) = \text{length}(\phi \circ \gamma)$$

Proof. Let ϕ be an isometry, and $\gamma : [a,b] \to \mathbb{E}^2$ be a curve. Then we know the length of γ itself is defined as $\text{length}(\gamma) = \int_{[a,b]} \|\gamma'(t)\| dt$, and we wish to compare this with the length of $\phi \circ \gamma$

$$\text{length}(\phi \circ \gamma) = \int_{[a,b]} \|(\phi \circ \gamma)'(t)\| dt$$

To compute this integral we need to first differentiate $\phi \circ \gamma$ using the chain rule:

$$(\phi \circ \gamma)'(t) = D\phi_{\gamma(t)}\gamma'(t)$$

Where here recall that $D\phi_{\gamma(t)}$ is a *matrix* - the linear transformation recording the infinitesimal behavior of ϕ at a point - and $\gamma'(t)$ is a *tangent vector* - an infinitesimal piece of arc. Since we have assumed that ϕ is an isometry, it preserves infinitesimal lengths by definition so

$$\|D\phi_{\gamma(t)}\gamma'(t)\| = \|\gamma'(t)\|$$

Using this, we can simplify our integral:

$$\begin{aligned}
\text{length}(\phi \circ \gamma) &= \int_{[a,b]} \|(\phi \circ \gamma)'(t)\| dt \\
&= \int_{[a,b]} \|D\phi_{\gamma(t)}\gamma'(t)\| dt \\
&= \int_{[a,b]} \|\gamma'(t)\| dt \\
&= \text{length}(\gamma)
\end{aligned}$$

□

In fact, the converse of this is true as well: if a differentiable function preserves the lengths of all curves, then it preserves infinitesimal lengths, and is an isometry.

11.1. Translations & Some Rotations

We will go more in-depth in our discussion of isometries later on, but for now it's good practice with the definition to find a couple examples that we can use.

Theorem 11.3 (Translations are Isometries). *If $v = \langle a, b\rangle$ is a fixed vector, a* translation by v *of $\mathbb{E}^2$ is given by the function $T(p) = p + v$, or, in coordinates,*

$$T(x, y) = (x + a, y + b)$$

. This is an isometry of $\mathbb{E}^2$.

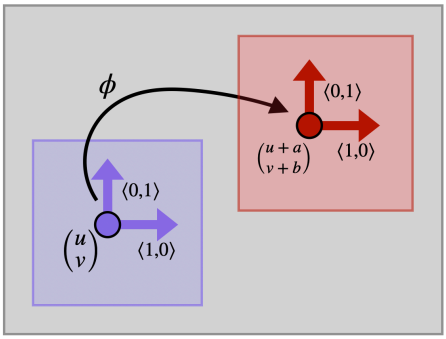


Figure 11.2.: A translation of the plane does not change the coordinates of *any* infinitesimal vectors: thus it does not change their lengths. Translations are isometries.

Proof. Here we need to compute the derivative of T: Since $T(x, y) = (x + a, y + b)$ we get the matrix

$$\begin{aligned} DT &= \begin{pmatrix} \partial_x T_1 & \partial_y T_1 \\ \partial_x T_2 & \partial_y T_2 \end{pmatrix} \\ &= \begin{pmatrix} \partial_x (x + a) & \partial_y (x + a) \\ \partial_x (y + b) & \partial_y (y + b) \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{aligned}$$

This is the *identity matrix* which means it does nothing to vectors: if $v = \langle v_1, v_2\rangle$ is any vector based at $p \in \mathbb{E}^2$ then

$$DT_p(v) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

Thus, since DT did not change anything at all about v it did not change its length and so T is an isometry.

$$\|DT_p(v)\| = \|v\|$$

□

A particularly nice collection of functions to work with are the *linear maps* $\mathbb{E}^2 \to \mathbb{E}^2$. One of the nicest properties of these they are easy to differentiate: recall that if A is a matrix representing the linear map $\phi(v) = Av$ then $D\phi = A$ is the same matrix! So, if we are looking for *linear isometries* we can save ourselves the work of differentiation.

Example 11.1. The following linear map is an isometry of $\mathbb{E}^2$

$$\phi(x, y) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

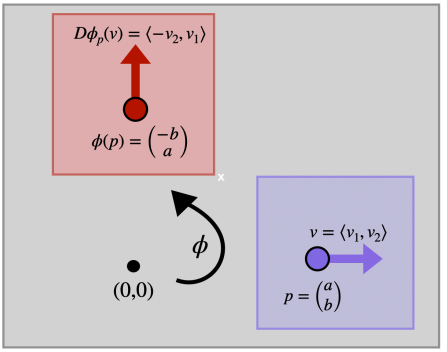


Figure 11.3.: This linear map affects both points and tangent vectors, but it does not change the length of any tangent vector. Thus, it is an isometry (this is a rotation by 90 degrees)

Proof. We check that this preserves all infinitesimal lengths. Denote by A the matrix $A = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, then ϕ is the linear map $\phi(x) = Ax$, so its derivative is given by the same linear map, $D\phi_p(v) = Av$ at every point p.

Thus, to see that ϕ is an isometry, all we need to do is check whether or not the length of Av is the same as the length of v for an arbitrary vector $v = \langle v_1, v_2 \rangle$.

$$Av = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} -v_2 \\ v_1 \end{pmatrix}$$

$$\begin{aligned} \|Av\| &= \|\langle -v_2, v_1 \rangle\| \\ &= \sqrt{(-v_2)^2 + (v_1)^2} \\ &= \sqrt{v_1^2 + v_2^2} \\ &= \|v\| \end{aligned}$$

As these lengths are the same, ϕ is an isometry. □

Of course, not all linear maps are isometries: its easy to cook up something that doesn't preserve infinitesimal lengths.

Example 11.2. The following linear map is not an isometry of $\mathbb{E}^2$:

$$\phi(x, y) = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

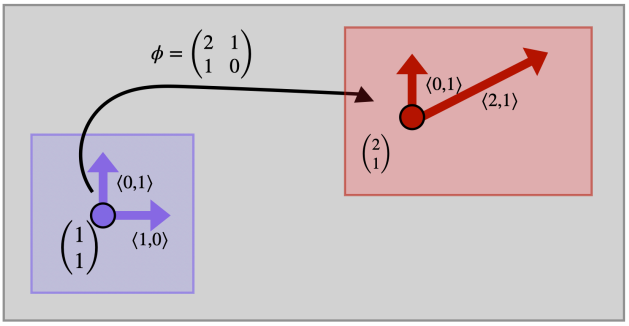


Figure 11.4.: This map does not change the length of the infinitesimal vector $\langle 0, 1 \rangle$ at any point, but it *does* change the length of $\langle 1, 0 \rangle$. Thus it is **not** an isometry.

Proof. Since ϕ is linear, $D\phi$ is equal to the same linear map $\begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}$ at each point of $\mathbb{E}^2$. To prove ϕ is *not* an isometry, all we need to do is find one vector which has its length changed by $D\phi$. Consider the vector $v = \langle 1, 0 \rangle$ based at $p = O \in \mathbb{E}^2$. Then

$$D\phi_p(v) = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$$

While v had unit length $D\phi_p(v)$ has length 2, so ϕ does not preserve all infinitesimal lengths, and therefore is not an isometry. □

What are the conditions on a linear map being an isometry? Well, if it needs to preserve all infinitesimal lengths, it needs to send the unit vector $\langle 1, 0\rangle$ to some other unit vector, and same for $\langle 0, 1\rangle$. Since the image of these vectors are the first and second columns of the matrix representing them, this means that every linear isometry has a matrix whose rows are unit vectors. Is every such matrix an isometry?

Exercise 11.1. Write down a linear map that sends both $\langle 1, 0\rangle$ and $\langle 0, 1\rangle$ to unit vectors, but is not an isometry.

However, if we choose unit vectors *correctly*, we do get an linear isometry! Intuitively from our previous experience with the plane we know what to should happen, we are looking for a rotation! The theorem below confirms that rotations about O in the plane exist: you can fix that point, and perform an isometry that moves $\langle 1, 0\rangle$ to *any* other unit tangent vector in $T_O\mathbb{E}^2$.

Theorem 11.4. *Let v be an arbitrary unit vector based at O in $\mathbb{E}^2$. Then there exists an isometry ϕ of $\mathbb{E}^2$ which takes fixes O and takes $\langle 1, 0\rangle$ to v. Such an isometry is called a* rotation about O.

Proof. Let $v = \langle v_1, v_2\rangle$ be a unit vector. Then the vector $v^\perp = \langle -v_2, v_1\rangle$ is a rotated copy of v by 90 degrees. From these, we can build a linear map which sends $\langle 1, 0\rangle$ to v (and also $\langle 0, 1\rangle$ to $v^\perp$):

$$R(x, y) = \begin{pmatrix} v_1 & -v_2 \\ v_2 & v_1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

Now we check this is an isometry. Let p be an arbitrary point in $\mathbb{E}^2$ and $u = \langle a, b\rangle$ be an arbitrary tangent vector based at p. We need to see that $\|u\| = \|DR_p u\|$. Since R is a linear transformation, we know that it is its own derivative, so

$$DR_p = \begin{pmatrix} v_1 & -v_2 \\ v_2 & v_1 \end{pmatrix}$$

And so we can apply without much trouble to u:

$$DR_p u = \begin{pmatrix} v_1 & -v_2 \\ v_2 & v_1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} v_1 a - v_2 b \\ v_2 a + v_1 b \end{pmatrix}$$

Calculating the length is now just a matter of algebra, using the fact that v is a unit vector so $v_1^2 + v_2^2 = 1$. After simplifying, we see

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

Thus the infinitesimal length of u was not changed by the transformation R, and as p, u were arbitrary this is true for all infinitesimal lengths - R is an isometry. □

Exercise 11.2. Check the calculation that is skipped in the proof above actually works out as claimed.

11.2. Creating Isometries: Conjugation

Some additional exercises to explore deeper the idea of isometries, and practice the chain rule!

Exercise 11.3 (Composition of Isometries)**.** If ϕ and ψ are two isometries of $\mathbb{E}^2$, then the composition $\phi \circ \psi$ is also an isometry.

Exercise 11.4 (Inversion of Isometries)**.** If ϕ is an isometry of $\mathbb{E}^2$, then its inverse function ϕ^{-1} is also an isometry.

Together these say that the isometries of a space form a *group*. Being able to compose and invert isometries is quite useful when you need to create an isometry that does a specific task out of a limited set of pieces.

As a first example, suppose you wanted to show there is a rotation about O that takes some unit vector $v \in T_O\mathbb{E}^2$ to another unit vector $w \in T_O\mathbb{E}^2$. So far we only have one theorem about rotations - Theorem 11.4, which tells us that we can find one taking $\langle 1, 0\rangle_0$ to any vector. We will need to create two of these, and combine them via composition and inversion:

Proposition 11.1. *For any two unit vectors $v, w \in T_O\mathbb{E}^2$, there is a Euclidean isometry which fixes O and sends v to w.*

Proof. Let ϕ_v be a rotation taking $\langle 1, 0\rangle$ to v, and ϕ_w be an rotation taking $\langle 1, 0\rangle$ to w: both of these are linear, and exist by Theorem 11.4. Now, consider the inverse function ϕ_v^{-1}. This is an isometry (by Exercise 11.4) which undoes the action of ϕ_v, so it fixes O and takes v to $\langle 1, 0\rangle$.

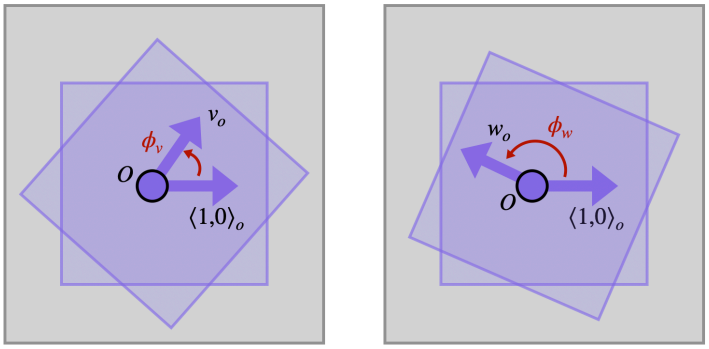


Figure 11.5.: Rotations taking $\langle 1, 0\rangle_0$ to v_o and w_o respectively.

Now consider the composition $\phi_w \circ \phi_v^{-1}$. This is a composition of isometries, and hence an isometry (Exercise 32.27). It fixes O since ϕ_v^{-1} does and ϕ_w does, so all we need to see is that it takes v to w. So, just follow the vector v! We first feed it into ϕ_v^{-1}, which takes it to $\langle 1, 0\rangle$, and then we feed the result into ϕ_w, which takes $\langle 1, 0\rangle$ to w!

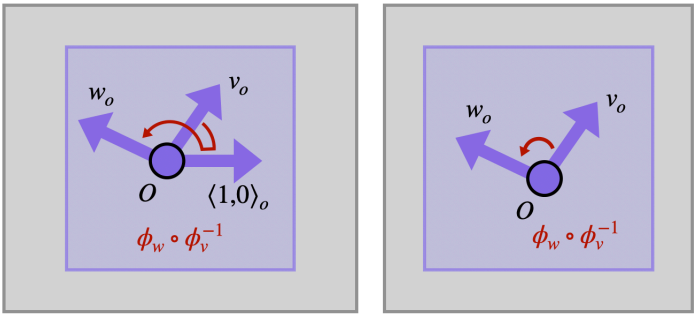


Figure 11.6.: The combination $\phi_w\phi_v^{-1}$ takes v_o to w_o directly. On the left, we see its motion step by step, passing through the intermediate vector $\langle 1, 0\rangle_0$. On the right, we see the net result.

If you wanted to write this in symbols instead of pictures or words, it looks like this:

$$\begin{aligned} D(\phi_w \circ \phi_v^{-1})_O(v) &= (D\phi_w)_O(D\phi_v^{-1})_O(v) \\ &= (D\phi_w)_O(\langle 1, 0\rangle) \\ &= w \end{aligned}$$

□

Next, we will look at trying to build an isometry that rotates around an arbitrary point p in the plane. We already found the isometries that rotate around 0: they are the nice *linear maps* of Theorem 11.4. But tracking down isometries that rotate around other points of the plane sounds more difficult. First of all - they cannot be linear maps! A linear map fixes the point O, but a rotation about the point p fixes....p! However, combining a translation *taking p to zero* with a rotation about zero in the right way, we can succeed!

Theorem 11.5. *Let p be a point in the Euclidean plane and $v = \langle v_1, v_2 \rangle$ a tangent vector based at p. Then there is an isometry of $\mathbb{E}^2$ which fixes p, and takes $\langle 1, 0 \rangle$ to v.*

Proof. Let T be the translation $T(q) = q + p$: this is an isometry by Theorem 11.3, and it takes O to p. Also, let R be the rotation about O which takes the vector $\langle 1, 0 \rangle$ based at O to the vector $v_0 = \langle v_1, v_2 \rangle$ based at O. (Recall v_o means a vector with the same coordinates as $v = v_p \in T_p\mathbb{E}^2$, but based at 0 instead of p.

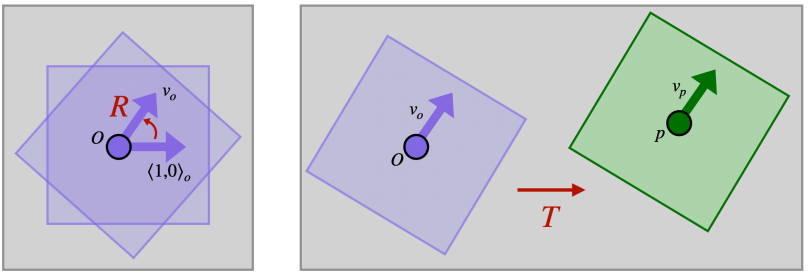


Figure 11.7.: We require a rotation about O and a translation from O to p in order to build a rotation about p.

From these, we construct the map $\phi = T \circ R \circ T^{-1}$. This is an isometry because its a composition of isometries and their inverses (Exercise 32.27,Exercise 11.4), so we just need to check that it does what is claimed.

This fixes the point p: since T takes O to p, its inverse takes p to O. Then R fixes O, and finally, T takes O back to p:

$$\begin{aligned} \phi(p) &= TRT^{-1}(p) \\ &= TR(O) \\ &= T(O) \\ &= p \end{aligned}$$

Next, we nee to check it does what we claim to the tangent vectors. To do so, we need to take the derivative of ϕ at p, and see that it takes $\langle 1,0\rangle_p$ to v_p. In symbols, we want to show $D\phi_p(\langle 1,0\rangle_p) = v_p$.

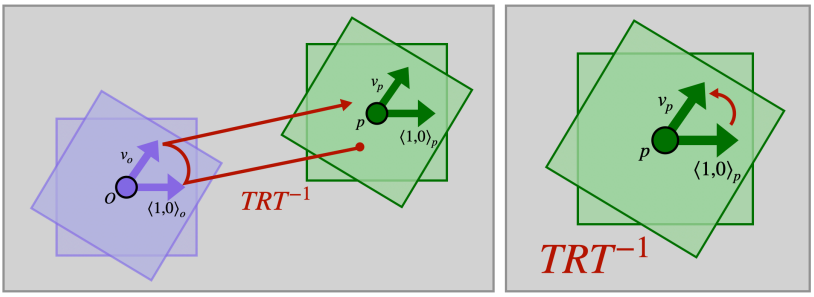


Figure 11.8.: The composition TRT^1 fixes p, and takes $\langle 1,0\rangle_p$ to v_p. On the left, we see the step-by-step action of this combination: starting from $\langle 1,0\rangle_p$ we translate to O, rotate to v_o, and then translate back to v_p. The right shows the net result: just a rotation at p.

We know by the proof of Theorem 11.3 that the derivative of T is the identity matrix. Thus, the derivative of T^{-1} is also the identity matrix (we differentiate an inverse by using the inverse of the derivative matrix, by Theorem 8.1). So applying DT^{-1} at p to $\langle 1,0\rangle_p$ leaves it unchanged, except it moves the basepoint to O (since $T^{-1}(p) = O$).

Next, we apply R. This fixes O, and by Theorem 11.4 we know DR_o takes $\langle 1,0\rangle_o$ to v_o. Finally, we apply T: since its derivative is the identity matrix it does not affect the coordinates of any vector just the basepoint, so it takes v_o to v_p.

In symbols:

$$\begin{aligned} D\phi_p(\langle 1,0\rangle_o) &= D(TRT^{-1})_p(\langle 1,0\rangle_o) \\ &= DT_o DR_o DT_p^{-1}(\langle 1,0\rangle_p) \\ &= DT_o DR_o(\langle 1,0\rangle_o) \\ &= DT_o(v_o) \\ &= v_p \end{aligned}$$

□

Exercise 11.5. Can you modify the argument of Theorem 11.5 above to prove that in fact for any point p and any two unit tangent vectors v_p, w_p in $T_p\mathbb{E}^2$, there is an isometry which fixes p and takes v_p to w_p?

Hint: look at the proof of Proposition 11.1 for inspiration.

This operation - *move, then do your next trick, then undo the original movement* is an extremely common manuever in mathematics to build new things from known things. Its essential not only in geometry, but also at the heart of abstract algebra and other fields, and is called *conjugation*.

Definition 11.3 (Conjugation)**.** If a and b are two mathematical objects that can be multiplied or composed, then the object

$$bab^{-1}$$

is called the *conjugate of a by b*.

Often, we will interpret this as *doing the action determined by a, at the location determined by b*. Thus we can describe the previous theorem much more succinctly with our new terminology: to rotate about the point p, we conjugate a rotation about 0 by a translation from 0 to p. Or - we perform a *rotation* at the *location we translate to*.

11.3. Homogenity and Isotropy

The fundamental property of Euclidean geometry that allowed the greeks and ancients to make so much progress was the incredible amount of symmetry that the plane has. It doesn't matter where you draw a triangle, a circle or another figure: all locations of the plane *look and act the same*. This concept that space looks the same at every point and also behaves the same in every direction is fundamental to modern geometry

Definition 11.4 (Homogeneous Space)**.** A space is *homogeneous* if for every pair of points in the space, there is an isometry taking one to the other.

Definition 11.5 (Isotropic Space)**.** A space is *isotropic* if for any point p and any two directions leaving p, there is a rotation of the space taking one direction to the other.

The existence of translations shows us that the Euclidean plane is homogeneous, while the ability to rotate about any point shows us that it is isotropic.

Theorem 11.6 (Euclidean space is Homogeneous and Isotropic)**.**

In practice, we will use the homogenity and isotropy of Euclidean space to simplify a lot of arguments. Once we prove something is true at one location (like the origin, where calculation is simple) we will immediately be able to deduce that the analogous theorem is true at all other points of the plane! To make such arguments, its useful to repackage homogenity and isotropy into a useful *tool*.

Proposition 11.2 (Moving from p to q.). *Given any two pairs p, v_p and q, w_q of points p, q in Euclidean space and unit tangent vectors $v_p \in T_p\mathbb{E}^2$, $w_q \in T_q\mathbb{E}^2$ based at them, there exists an isometry taking v_p to w_q.*

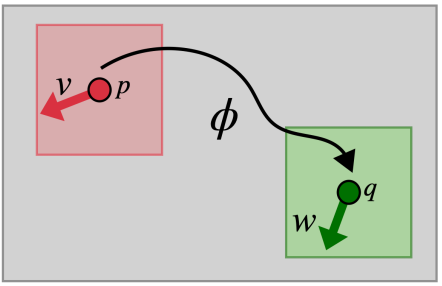


Figure 11.9.: The homogenity and isotropy of $\mathbb{E}^2$ lets us take any unit vector at any point, to any other via an isometry.

Exercise 11.6. Prove Proposition 11.2 above.

Hint: use Theorem 11.3 to construct isometries taking O to both p and q, and Proposition 11.1 to build the right sort of rotation around O that you need. Compose these (or their inverses) to get a map taking v_p to v_o, then to w_o, and finally to w_p.

11.4. Similarities

Isometries - maps that preserve all infinitesimal lengths - are very special among the collection of all possible maps of the plane. Most mappings $F : \mathbb{E}^2 \to \mathbb{E}^2$ don't do anything understandable to lengths!

However, there is one important intermediate ground of maps: they don't preserve distances - but they don't change them arbitrarily either. We will call a map a *similarity* if it scales all infinitesimal lengths by the same factor:

Definition 11.6. An map $\sigma : \mathbb{E}^2 \to \mathbb{E}^2$ is called a *similarity* if there is a positive real number k such that

$$\|D\sigma_p(v)\| = k\|v\|$$

for all tangent vectors v. This constant k is called the *scaling factor* or *dilation* of the map σ.

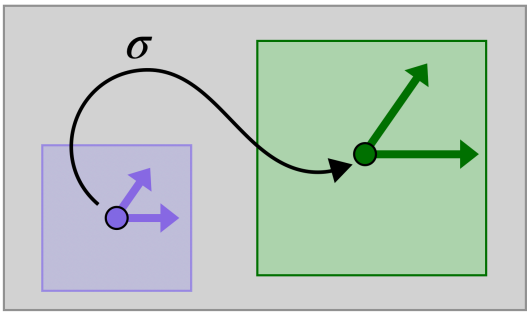


Figure 11.10.: A similarity uniformly scales all tangent vectors.

Perhaps the simplest similarities of the plane are given by vector scalar multiplication: just take the map $\sigma(x, y) = (kx, ky)$.

Example 11.3. The map $\sigma(x, y) = (2x, 2y)$ is a similarity with scaling factor 2. Computing its derivative we see

$$D\sigma = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

and so for any point p and any vector $v \in T_p$ applying $D\sigma_p$ just multiplies all its coordinates by 2. Thus if $v = \langle v_1, v_2 \rangle_p$,

$$\|D\sigma_p(v)\| = \|\langle 2v_1, 2v_2 \rangle\| = 2\|\langle v_1, v_2 \rangle\| = 2\|v\|$$

Since this is the same constant for *every* vector v, this implies that σ is a similarity!K

Because similarities do exactly the same thing to every tangent vector in the plane, we can compute exactly how they scale the lengths of curves.

Proposition 11.3 (Similarities Scale Lengths)**.** *Let $\gamma : [a, b] \to \mathbb{E}^2$ be a curve, and σ a similarity with scaling factor k. Then*

$$\text{length}(\sigma \circ \gamma) = k \, \text{length}(\gamma)$$

Proof. We compute the length of $\sigma \circ \gamma$ via an integral:

$$\begin{aligned}
\text{length}(\sigma \circ \gamma) &= \int_a^b \|(\sigma \circ \gamma)'(t)\| dt \\
&= \int_a^b \|D\sigma_{\gamma(t)}\gamma'(t)\| dt \\
&= \int_a^b k\|\gamma'(t)\| dt \\
&= k \int_a^b \|\gamma'(t)\| dt \\
&= k \,\text{length}(\gamma)
\end{aligned}$$

Where in the middle we used the fact that σ was a similarity so $\|D\sigma(v)\| = k\|v\|$ for any vector v. □

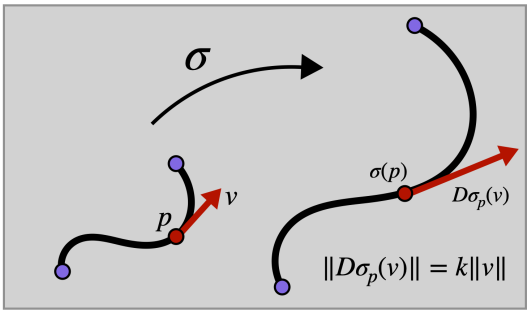


Figure 11.11.: Since all infinitesimal lengths are scaled up, so is their integral. Thus similarities linearly expand the length of all curves by their scaling constant.

Just like for isometries, we have as a theorem of calculus that this condition actually implies that our map is invertible! We will not prove this theorem here, and if you like you can instead treat this as an extra condition we require of a function to be a similarity.

Theorem 11.7 (Every Similarity is Invertible)**.**

For isometries, you proved the inverse of an isometry is an isometry (Exercise 11.4) by showing that if ϕ didn't change the length of any vectors, than neither could ϕ^{-1}. Here we investigate the analogous question for similarities.

Proposition 11.4. *If σ is a similarity with scaling factor k, then σ^{-1} is also a similarity, this time with scaling factor $1/k$.*

Proof. Let σ be such a similarity, and σ^{-1} be its inverse. Then by definition we know their composition is the identity function $

$$\sigma \circ \sigma^{-1} = \mathrm{id}$$

The identity function $(x, y) \mapsto (x, y)$ has the identity matrix as its derivative. On the other side, we can use the multivariable chain rule to get

$$D(\sigma\sigma^{-1})_p = D\sigma\sigma^{-1}(p)D\sigma_p^{-1} = I$$

Now start with any vector v based at a point p. We first feed this vector into $D\sigma_p^{-1}$, which returns a new vector - let's call it w. We don't know anything about w at the moment, but we do know that when we feed it into $D\sigma$, its length will multiply by k, since σ is a similarity. But we know more than this! The end result must be literally the vector v: since we started with v and the composition of $D\sigma$ with $D\sigma^{-1}$ is the identity matrix.

Thus we know that whatever w is, when you multiply its length by k you get the length of v, so

$$k\|w\| = \|v\|$$

But - remember w is just the vector $D_p\sigma^{-1}(v)$: so we've found

$$\|D\sigma_p^{-1}(v)\| = \frac{1}{k}\|v\|$$

And this holds for all vectors v - so the inverse is indeed a similarity, and the scaling constant is $1/k$. □

More generally, we can use the same sort of reasoning to understand compositions of any similarities.

Exercise 11.7. Prove that the composition of a similarity and isometry is another similarity, with the same scaling factor.

Exercise 11.8. If σ and ψ are two similarities with scaling constants c and k respectively, the composition $\sigma \circ \psi$ is also a similarity, with scaling constant ck.

From this, we can build many more similarities from the simple ones we know.

Exercise 11.9. The similarities $(x, y) \mapsto (kx, ky)$ fix O in the plane: can you use translations to build a similarity with scaling constant k which instead fixes the point $p = (a, b)$?

12. Lines

Laying down the foundations at a *deeper* level than the Greeks, we have some work to do before we can hope to recover the axioms of Euclid. Indeed - no where in our foundations does the term line even appear: we are in the awkward position of being able to work with any curve we like, but we do not know which among them is a straight line!

To find the lines among the sea of curves, we need a good and precise definition. Definitions single out an important property characterizing the object being defined, and for that definition to be good, we would like that condition to be *checkable* within the framework we are building. So - Euclid's definition of a line as a *breadthless length* is not going to do much for us here.

However, looking to history, we find several good candidate definitions among the properties of lines the ancients took as essential.

Definition 12.1 (Essential Properties of Lines)**.**

- Archimedes used as an *axiom of length* that the line segment between two points is the shortest among all curves connecting them. This could be turned upside down and directly used as the *defining feature* of lines: whichever curve is shortest, we call a line.
- As a followup to the infamously unhelpful *breadthless length* Euclid states the important feature of a line being that *the points lie evenly with themselves.* This also requires a bit of translation, but if we can define what it means for a curve to *turn*, we could then specify straight lines as *curves that do not turn.*
- The term line also shows up in phrases such as *line of symmetry* - for instance in discussing that the human form is left-right-symmetric. The fact that reflections fix a line is foundational to geometric arts like Origami, which is what allows the use of Euclidean geometry to describe the collection of creases made: they arise as lines of symmetry, so they are the lines of Euclid!

In fact all three of these things can be made into precise statements in our new geometry, and we can compute exactly what sort of curves satisfy each of them. The main purpose of this section is to do so, and to show that all three of them end up specifying exactly the same class of curves! This is one reason that lines are so important to geometry: they are the single objects sharing *all three* of these very natural properties!

12.1. Shortest

We start first with the insight of archimedes, and attempt to make precise the notion of *shortest curve between two points*. In doing so, we will actually first define *line segment*, and then use this to define lines more generally.

Definition 12.2 (Line Segment)**.** Given two points $p, q \in \mathbb{E}^2$, a curve γ starting at p and ending at q is called a line segment if it is distance minimizing. That is, for all other curves α from p to q, we have

$$\text{length}(\gamma) \leq \text{length}(\alpha)$$

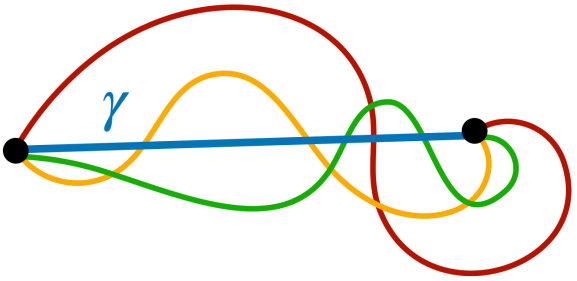


Figure 12.1.: Defining a line segment γ as the shortest curve joining its endpoints.

This definition seems very powerful: if you know something is a line segment you know *a lot about it*: you know how its length relates to the length of every single other curve!

Theorem 12.1 (Segments of x-axis are Minimizers)**.** *Finite segments of the x axis, that is, curves of the form*

$$\gamma(t) = (t, 0) \qquad a \leq t \leq b$$

are length minimizers.

Proof. First, we compute the length of the x-axis between 0 and L by integrating the infinitesimal lengths of γ:

$$\gamma(t) = (t, 0) \implies \gamma'(t) = (1, 0) \implies \|\gamma'(t)\| = 1$$

$$\text{length}(\gamma) = \int_0^L \|\gamma'(t)\| dt = \int_0^L dt = L$$

This of course is unsurprising! But its good to know explicitly that we have found a curve of length *exactly* L. Now, let $\alpha(t) = (x(t), y(t))$ be any arbitrary (regular)curve connecting $(0, 0)$ to $(L, 0)$.

Our goal is now to show that $\text{length}(\alpha) \geq L$, as this would mean no curve can have a length less than L, and our segment of the x-axis above is indeed the shortest curve! The difficulty in doing so is that we know *very little* about α, and hence very little about its coordinate functions $x(t), y(t)$. If α is defined on the interval $[a, b]$ knowing that it starts and ends at $(0, 0)$ and $(L, 0)$ implies

$$\begin{aligned} x(a) &= 0 \quad x(b) = L \\ y(a) &= 0 \quad y(b) = 0 \end{aligned}$$

but this is essentially all we know. Nonetheless, let's push onwards and see what we can learn about $\text{length}(\alpha)$ by writing out its definition.

$$\text{length}(\alpha) = \int_a^b \|\alpha'(t)\| dt = \int_a^b \sqrt{x'(t)^2 + y'(t)^2}\, dt$$

Now we do some estimation: we know that whatever y is, $y'(t)^2$ is nonnegative - because its squared, after all! So

$$y'(t)^2 \geq 0 \implies x'(t)^2 + y'(t)^2 \geq x'(t)^2$$

We can then take the square root of both sides of this equation (which preserves inequalities) to get

$$\sqrt{x'(t)^2 + y'(t)^2} \geq \sqrt{x'(t)^2} = |x'(t)| \geq x'(t)$$

Igoring all the middle terms in this string of inequalities, (and recalling the left hand side is the norm of α') we see that

$$\|\alpha'(t)\| \geq x'(t) \qquad \text{for all } t$$

Thus, as functions of t, we see that the curve x' lies below the curve $\|\alpha'\|$: since the area under the lower curve must be less than or equal to the upper, this inequality is still preserved after we integrate.

$$\int_a^b \|\alpha'(t)\|\, dt \geq \int_a^b x'(t)\, dt$$

But now we have really made some progress: on the right side here we are integrating a derivative, so we can use the fundamental theorem of calculus! The antiderivative of $x'(t)$ is just $x(t)$ of course, so we evaluate at the endpoints:

$$\int_a^b x'(t)\,dt = x(t)\Big|_a^b = x(b) - x(a) = L - 0 = L$$

And with that, we've done it! The integral on the left side was precisely the length of α, so

$$\text{length}(\alpha) \geq L$$

□

Now that we have a firm understanding of *segments*, how can we properly bootstrap this idea to a definition of lines? A line itself has no endpoints, and so is not a distance minimizing curve! However, it has the property that if you cut out any segment from it, that segment *is* distance minimizing. To say this formally, we need a word for "cut out a segment of a curve"

Definition 12.3 (Finite Segment of a Curve)**.** Given a curve $\gamma : \mathbb{R} \to \mathbb{E}^2$, a finite segment of γ is the restriction of γ to some finite interval $[a, b] \subset \mathbb{R}$.

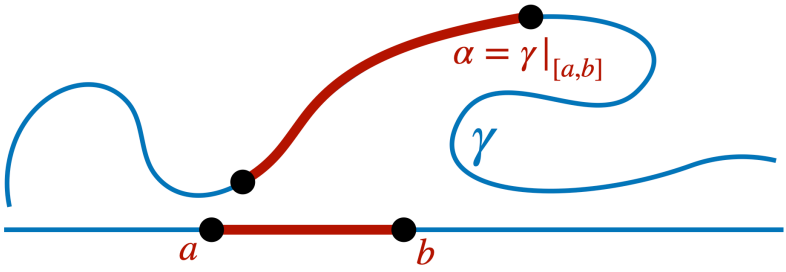


Figure 12.2.: A segment of a curve is a restriction of that curve to a sub-interval of its domain.

This makes the definition for a line completely precise:

Definition 12.4 (Line)**.** A curve γ is a line if all of its finite segments are line segments.

This sounds pretty useless until we unpack it: since line segments are distance minimizers, this is saying that to be a line, a curve must have the property that it is *distance minimizing between any two points it passes through!* A strong condition indeed.

However, given the work we did above on segments of the x axis, we can now immediately apply this to the entire axis itself.

Corollary 12.1 (The x-axis is a line)**.** *Every finite segment of the x-axis is a distance minimizing line segment, so the x-axis is a line.*

Well, after all this theory we have finally managed to track down one line in the plane! How can we find more? One option of course is to mimic the argument given here: with trivial modifications we can similarly prove that the y axis is a line, and that curves of the form $x = a$ or $y = b$ are all lines as well. But it would take a little more work (in the form of a clever u-substitution) to apply this further: we took big advantage of the fact that one of the coordinate derivatives was *zero* in our proof!

Instead, we take this as our first opportunity to use one of the most powerful ideas in modern geometry: *symmetry*. We proved that isometries preserve the length of all curves, and this has an important consequence: isometries send lines to lines!

Theorem 12.2 (Isometries Send Lines to Lines)**.** *Let $\gamma : \mathbb{R} \to \mathbb{E}^2$ be a line, and $\phi : \mathbb{E}^2 \to \mathbb{E}^2$ be an isometry. Then $\phi \circ \gamma$ is also a line.*

Proof. To argue that $\phi \circ \gamma$ is a line, we need to show that all of its finite segments are length-minimizing. So, pick some arbitrary interval $[a, b] \subset \mathbb{R}$ and look at the restriction of our curve to that segment, which goes from $\phi(\gamma(a)) = p$ to $\phi(\gamma(b)) = q$.

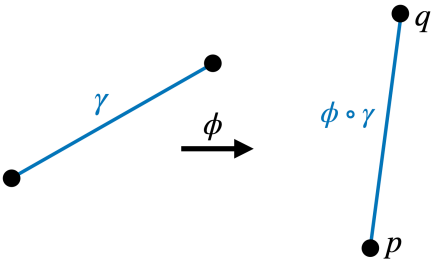


Figure 12.3.: A line segment γ and its image under an isometry ϕ.

Assume for the sake of contradiction that this is *not* length minimizing: then there is some other curve α connecting p to q which is of shorter length.

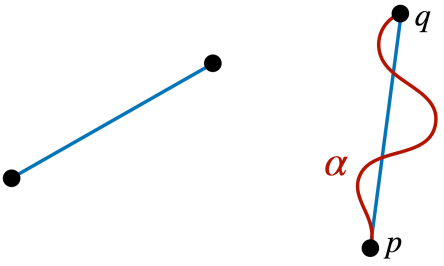


Figure 12.4.: A mysterious curve α which is assumed to be shorter than $\phi \circ \gamma$.

Now, apply the inverse function ϕ^{-1} to everything: this takes the segment $\phi \circ \gamma$ back to γ, and takes α to a new curve $\phi^{-1} \circ \alpha$, starting and ending at the same points as the corresponding segment of γ:

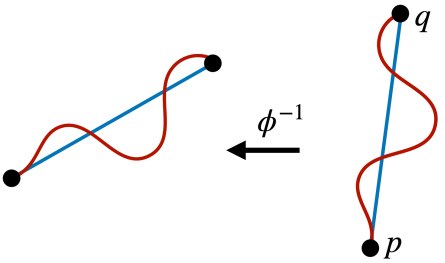


Figure 12.5.: Applying ϕ^{-1} moves α back to share endpoints with the original γ, which we know to be length-minimizing

Since isometries preserve length, we know that since α was shorter than $\phi \circ \gamma$, we must now have that $\phi^{-1} \circ \alpha$ is shorter than γ! But this is impossible: we assumed that γ itself was a *line*, so all of its segments are length-minimizing: there are no shorter curves!

Thus, its impossible that α exists, so $\phi \circ \gamma$ must have been the shortest segment between p and q after all. As all segments of this curve are distance minimizers, its a line! □

This gives us an easy prescription to track down lines: we already know $\gamma(t) = (t, 0)$ is a line - and if we apply *any isometry at all* to this, we will get another line!

Corollary 12.2 (Affine Equations are Lines)**.** *Every linear equation $f(t) = (at, bt)$ describes a line that passes through the origin. Every affine equation of the form*

$$\ell(t) = \begin{pmatrix} at + c \\ bt + d \end{pmatrix}$$

is also a Euclidean line.

Here we concentrate on the main case where $\langle a, b \rangle$ is a unit vector. We comment below the proof on the small change needed when it is not.

Proof. Then, the rotation $\phi = \left(\begin{smallmatrix} a & -b \\ b & a \end{smallmatrix}\right)$ taking $\langle 1, 0 \rangle \in T_{(0,0)}\mathbb{E}^2$ to $\langle a, b \rangle$ is an isometry, so it sends lines to lines. Applying it to the x-axis $\gamma(t) = (t, 0)$, we see

$$\phi \circ \gamma(t) = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \begin{pmatrix} t \\ 0 \end{pmatrix} = \begin{pmatrix} at \\ bt \end{pmatrix}$$

Thus, $t \mapsto (at, bt)$ is a line! Next, we can use the fact that for fixed c, d the translation $\psi(x, y) = (x + c, y + d)$ is an isometry of $\mathbb{E}^2$, so

$$\psi(at, bt) = (at + c, bt + d)$$

is also a line! □

If $v = \langle a, b \rangle$ is not a unit vector, then we can run the argument above using the unit vector $\frac{v}{\|v\|}$. This gives us that the curve below is a line

$$\beta(t) = \left(\frac{a}{\sqrt{a^2 + b^2}} t, \frac{b}{\sqrt{a^2 + b^2}} t \right)$$

Since we know that the length of curves does not depend on their parameterization, we can speed up or slow down β by pre-composing it with another function, and not change the fact that it is distance minimizing! Speeding it up by $t \mapsto \sqrt{a^2 + b^2} t$ gives

$$\beta\left(\sqrt{a^2 + b^2} t\right) = (at, bt)$$

Thus $t \mapsto (at, bt)$ is a line for any $a, b \in \mathbb{R}$!

We saw in Theorem 12.2 that any isometry will carry a line to another line. The same is true more generally of similarities:

Exercise 12.1 (Similarities Send Lines to Lines)**.** Let $\gamma : \mathbb{R} \to \mathbb{E}^2$ be a line, and $\sigma : \mathbb{E}^2 \to \mathbb{E}^2$ be a similarity. Prove that $\sigma \circ \gamma$ is also a line.

*Hint-replicate the proof of Theorem 12.2 as closely as possible, replacing the isometry ϕ with the similarity σ, and keeping track of the scaling factors of σ versus σ^{-1} (Proposition 11.4).

Using these tools, we can already start our process of rebuilding the Elements from below!

Theorem 12.3 (Proving Euclid's Axiom I)**.** *Given any two points $p, q \in \mathbb{E}^2$, there is a line segment connecting p to q.*

Proof. Knowing that line segments are given by affine equations, we need just fine an affine equation $\gamma(t)$ where $\gamma(0) = p$ and $\gamma(1) = q$. Perhaps the simplest such is

$$\gamma(t) = p + (q - p)t$$

□

Theorem 12.4 (Proving Euclid's Axiom II)**.** *Given a line segment between two points of $\mathbb{E}^2$, it can be extended indefinitely in either direction.*

Proof. Let $p, q \in \mathbb{E}^2$ and define the line segment $\gamma : [0,1] \to \mathbb{E}^2$ by $\gamma(t) = p + (q-p)t$ as in the previous theorem. To extend this line segment indefinitely, we need only extend the domain from $[0,1]$ to an arbitrary interval $[a,b]$ containing $[0,1]$.

The result is still an affine equation on a closed interval, and so still is a line segment by Corollary 12.2. And, as $[a,b]$ contains $[0,1]$ this new segment contains the original segment from p to q, so it represents an *extension* of the segment. □

12.1.1. Uniqueness of Lines

Above we proved the *existence* of lines, and found that all affine equations describe lines in the plane. But are these all the lines there are to be found? In fact they are - and we can confirm this with very little extra work: we had all the ideas in place already during the proof of Theorem 12.1.

Proposition 12.1. *Segments of the x-axis are the* unique *distance minimizers between their endpoints.*

Proof. Let $\gamma(t) = (t, 0)$ between $t = a$ and $t = b$, and $\alpha(t) = (x(t), y(t))$ be a different curve with the same endpoints. Then since α does not just trace the x axis, we must have $y(t) \neq 0$ at some point. But $y(a) = 0$ and $y(b) = 0$ at the endpoints, so for y to go from zero to nonzero, it must have *nonzero derivative* on some interval.

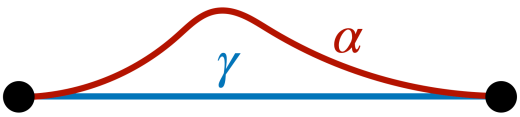


Figure 12.6.: Proving that the line segments we already know of are the *unique* minimizers.

But, this means that $y'(t)^2$ is *strictly greater than zero* on some interval, so $\|\alpha'(t)\| > \|\gamma'(t)\|$, and $\|\alpha'(t)\| - \|\gamma'(t)\| > 0$ on some interval inside of $[a,b]$. Furthermore, since we already knew $\|\alpha'\| \geq \|\gamma'$, this quantity is never negative. Thus

$$\int_a^b \|\alpha'(t)\| - \|\gamma'(t)\| \, dt > 0$$

And, re-arranging the integral, this immediately implies

$$\text{length}(\alpha) = \int_a^b \|\alpha'\| dt > \int_a^b \|\gamma'\| dt = \text{length}(\gamma)$$

□

Applying isometries to this, we can extend this result to any of the segments we already know:

Exercise 12.2. Prove that all Euclidean lines are given by affine equations. *Hint: we already know that the affine equation $\gamma(t) = (at + b, ct + d)$ defines a line. Can you show there is no curve of an equally short length, by using steps similar to the proofs Theorem 12.2 and* ***?@cor-cor-affline-eqns-are-lines*** *to reach a contradiction given that we just proved Proposition 12.1?*

12.2. Straightest

Another notion of line is "curve that doesn't turn". How do we make this precise? The unit tangent vector to a curve gives its direction, so we say a curve "turns" if the tangent changes direction.

The derivative of the tangent vector is acceleration, a "straight curve" would have acceleration zero.

Definition 12.5 (Straight)**.** A curve γ is called *straight* if its tangent vector does not change. That is, if its acceleration is zero.

Remark 12.1. You might worry what it *means* to say that tangent vectors are constant, since each one of them technically lives in a different tangent space! This difficulty will be absolutely crucial to deal with later on, when space itself is curved. But here in $\mathbb{E}^2$, we can take advantage of the fact that we can make sense of the basis vectors $\langle 1, 0\rangle$ and $\langle 0, 1\rangle$ in each tangent space $T_p\mathbb{E}^2$: then constant just means that the components of the vectors are constant in time.

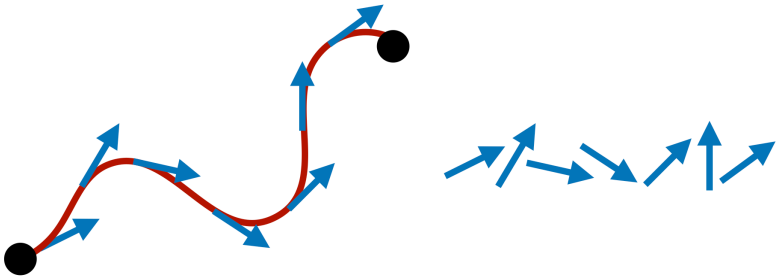


Figure 12.7.: This curve is not straight: which we can measure by seeing that its tangent vectors are not constant: they change in direction as we move along the curve.

We've already done all of the hard work above, and we can now quickly confirm that this alternative definition picks out *exactly the same class of curves.*

Theorem 12.5 (Distance Minimizers are Straight)**.** *A curve γ is distance minimizing if and only if it is straight.*

Proof. This is a direct computation, now that we've proven that every distance minimizer is given by a linear equation $\ell(t) = p + tv$. Differentiating once leaves $\ell'(t) = v$, and differentiating twice gives

$$\ell''(t) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Thus, ℓ is straight.

To prove the other direction, we now assume we start with a straight curve γ, and we wish to prove its distance minimizing. If γ is straight, then $\gamma'' = 0$ and integrating twice we see that

$$\gamma(t) = (at + c, bt + d)$$

for some constants a, b, c, d. Thus, γ is an affine curve, and we know affine curves are distance minimizers (Corollary 12.2). So, we are done! □

This will turn out to be true in general: while we will have to be a little more careful when moving onwards to other geometries, curves that are straight will coincide with curves that minimize distance.

12.3. Folding

Finally we come to the third possible definition of line, and show that it also picks out the same collection of curves!

Definition 12.6 (Line of Symmetry)**.** A *fixed point* of an isometry $\phi : \mathbb{E}^2 \to \mathbb{E}^2$ is a point p with $\phi(p) = p$.

A curve γ is called a *line of symmetry* of $\mathbb{E}^2$ if there exists an isometry which fixes $\gamma(t)$ for all t.

This captures the intuitive notion of a crease from folding paper, or reflecting across a line: this swaps the two sides of the plane but leaves What are the fixed sets of reflections?

Proposition 12.2 (Reflecting in the x-Axis is an Isometry)**.** *The map $\phi(x, y) = (x, -y)$ is an isometry of $\mathbb{E}^2$.*

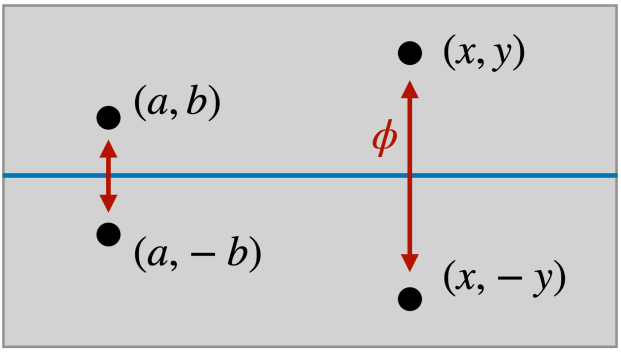


Figure 12.8.: Reflection across the x axis is an isometry of $\mathbb{E}^2$.

Proof. First, notice that ϕ is actually a *linear map*, so we can write it as a matrix:

$$\phi(x, y) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

Since ϕ is linear, its derivative is constant and also equal to ϕ at every point. Thus to check that it is an isometry, we only need to see that it does not change the length of any vectors.

Let $v = \langle v_1, v_2 \rangle_p \in T_p\mathbb{E}^2$ be a tangent vector based at some arbitrary point p. Then

$$D\phi_p(v) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} v_1 \\ -v_2 \end{pmatrix}$$

And measuring lengths with the vector norm,

$$\|D\phi_p(v)\| = \sqrt{v_1^2 + (-v_2)^2} = \sqrt{v_1^2 + v_2^2} = \|v\|$$

Thus, ϕ is an isometry. □

This map fixes the line of points $(x, 0)$ as it only negates the y component.Thus, the x axis is a line of symmetry! Similar to before, we can use isometries to prove that every affine curve is the fixed point of some reflection.

Exercise 12.3 (Reflections in Any Line)**.** Prove that every affine curve is a line of symmetry. *Hint: given an isometry that reflects in the x axis, can you build an isometry that reflects in any other line? Consider moving the line to the x axis, reflecting, and then moving back.*

The converse is also true: that every line of symmetry is an affine equation: so this characterization of lines exactly agrees with the two previous. To prove this, we will need a bit better understanding of the isometries of Euclidean space, and so will postpone until that chapter,

12.4. Distance

So far in our development of Euclidean geometry, we have defined the *length* of a curve, but we have not defined any notion of *distance* between two points. This makes some sense, as the distance between two locations depends on how you get from one to the other, and that's exactly what our definition captures!

However, now that we *know* there is a unique shortest curve between any two points, there's a natural candidate for distance: the shortest possible path.

Definition 12.7 (Distance)**.** The distance between two points $p, q \in \mathbb{E}^2$ is the length of the shortest possible curve starting at p and ending at q.

Because of all of our hard work above, we can turn this rather abstract definition into something concrete and practical!

Theorem 12.6 (The Euclidean Distance)**.** *Let p and q be any two points in the plane. Then the Euclidean distance between them is given by*

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

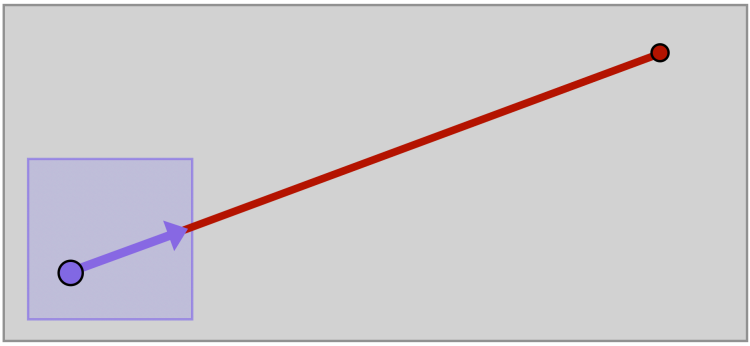


Figure 12.9.: Because lines are affine equations (and thus have *constant* derivative), the infinitesimal pythagorean theorem scales up to the distance function.

Proof. We can write down a distance minimizing curve from p to q as an affine equation:

$$\gamma(t) = p + t(q - p)$$

This is equal to p when $t = 0$ and q when $t = 1$. Thus, its length is given by the integral of γ' over $[0, 1]$. Computing the derivative is straightforward since γ is affine: $\gamma'(t) = q - p = (q_1 - p_1, q_2 - p_2)$, and so the length is

$$\begin{aligned}
\operatorname{dist}(p, q) &= \operatorname{length}(\gamma) \\
&= \int_0^1 \|\gamma'\| \, dt \\
&= \int_0^1 \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2} \, dt \\
&= \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2} \int_0^1 dt \\
&= \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}
\end{aligned}$$

□

Proposition 12.3 (Distance is preserved by isometries). *If p and q are any two points in the plane and ϕ is an isometry, then*

$$\operatorname{dist}(\phi(p), \phi(q)) = \operatorname{dist}(p, q)$$

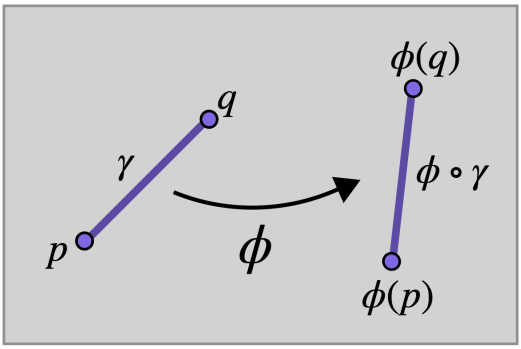


Figure 12.10.: Distance is invariant under isometry since isometries send lines to lines, and preserve the length of all curves.

Proof. First, we start with the isometry case. Given two points p, q we can construct a line segment γ from p to q (Theorem 12.3), and as this segment is the minimizer we know its length accurately measures the distance: \$dist$(p, q) = \text{length}(\gamma)$.

Applying ϕ we recall that isometries carry lines to lines (Theorem 12.2) to note that $\phi \circ \gamma$ is a line segment between $\phi(p)$ and $\phi(q)$, and as line segments are distance minimizers, we know

$$\text{dist}(\phi(p), \phi(q)) = \text{length}(\phi \circ \gamma)$$

Finally, we recall that isometries don't change the length of curves (Theorem 11.2) to see

$$\text{length}(\gamma) = \text{length}(\phi \circ \gamma)$$

and stringing all these equalities together gives

$$\text{dist}(p, q) = \text{length}(\gamma) = \text{length}(\phi \circ \gamma) = \text{dist}(\phi(p), \phi(q))$$

□

Exercise 12.4. If p,q are any two points in the plane and σ is a similarity with scaling factor k, prove

$$\text{dist}(\sigma(p), \sigma(q)) = k\,\text{dist}(p, q)$$

Hint: follow closely the argument for isometries above, replacing the theorems relating isometries, line segments, and lengths with the corresponding results for similarities.

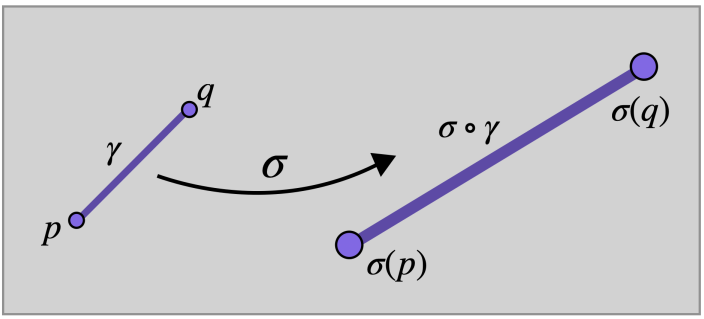


Figure 12.11.: Distance is scaled under similarity since similarities send lines to lines, and linearly scale the length of all curves.

It will often be useful to measure distances not just to points, but to more complicated objects in the plane.

Remark 12.2. We are avoiding a detail here, that anyone who has seen real analysis may be interested in. Sometimes, the *minimum* distance between p and a point in R doesn't exist, but only the *infimum* of such distances does. However, we will never encounter such cases in this text.

Definition 12.8 (Distance to a Set)**.** Let $R \subset \mathbb{E}^2$ be a region in the plane. Then the distance from a point $p \in \mathbb{E}^2$ to R is defined as the *shortest line segment connecting p to any point of R, and is denoted $\text{dist}(p, R)$.

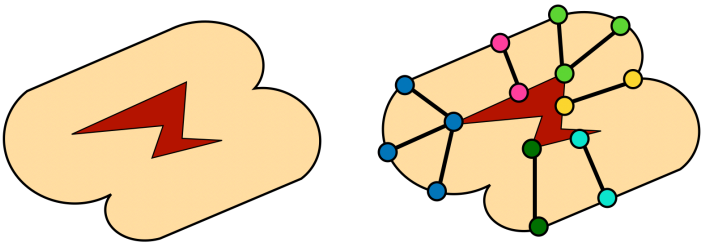


Figure 12.12.: The set of points at constant distance from a set (red region). On the right, a collection of points and the shortest line connecting them to a point of the set.

12.4.1. Useful Computations

Now that we know exactly what lines *are*, we can convert elementary geometric problems - such as when they intersect - into algebraic problems, solvable via systems

of equations. Here's an example.

Exercise 12.5 (Intersecting Lines)**.** Calculate the point of intersection between two lines $ax + by = c$ and the the diagonal line $x = y$. When is there no intersection?

With the ability to *solve equations* (since lines are given by affine equations, which are easy to work with!) we have developed a geometric *superpower.* To demonstrate this, we can use this to prove Playfair's axiom (remember, this is *equivalent* to Euclid's 5th Postulate!)

Proposition 12.4. *Given any line L in $\mathbb{E}^2$, and any point $p \in \mathbb{E}^2$ not lying on L, there exists a unique line Λ through p which does not intersect p.*

Exercise 12.6. Prove Proposition 32.1.

Hint: use isometries to help you out!
First, use an isometry to move L to the x-axis. Then, use another isometry to keep L on the x axis, but to move p to some point along the y axis (and possibly, use a reflection to then insure p has been moved to a point on the positive y axis, if you like!). Then, prove that through any point on the y axis there is a unique line that does not intersect the x-axis.

In addition to algebra, founding our new geometry on calculus makes all of these tools also available to us. As a first example, we will use our knowledge of derivatives to minimize the distance between a point and a line. Minimizing distance turns out to be a pretty common thing one needs to do in applications of geometry, and while straightforward theoretically (take the derivative, set it equal to zero), its annoying in practice because of the square root in the distance formula. But there is a nice trick to get around this:

Exercise 12.7 (Minimizing the Square: A Very Useful Trick!)**.** Let $f(x)$ be a differentiable positive function of one variable, and let $s(x) = f(x)^2$ be its square. Show that the minima of $s(x)$ and $f(x)$ occur at the same points, by following the steps below:

- First, assume $x = a$ is the location of a minimum of f. What does the first and second derivative test tell you about the values $f'(a)$ and $f''(a)$? Use this, together with the fact that $f(a) > 0$ to show that $x = a$ is also the location of a minimum of s (using the second derivative test).
- Conversely, assume $x = a$ is the location of a minimum of $s(x)$. Now, you know information about the derivatives $s'(a)$ and $s''(a)$. Use this to conclude information about $f'(a)$ and $f''(a)$ to show that a is a minimum for f as well.

This tells us anytime we want to minimize a positive function, we could always choose to find where its square is minimized instead, if that turns out to be easier. The main question is below, where without loss of generality we have taken the point to be the origin (as we could always slide it there via an isometry).

Exercise 12.8 (Closest Point on Line)**.** Let L be the line traced by the affine curve $\gamma(t) = \begin{pmatrix} at + c \\ bt + d \end{pmatrix}$, and O be the origin, as usual. Calculate $\text{dist}(O, L)$

Hint: use calculus to find the closest point on L to O. Can you minimize the squared distance from $\gamma(t)$ to $(0, 0)$?

13. Shapes

Defining the notion of lines and distance really helps in getting our new geometry off the ground. In this relatively short chapter, we give precise definitions for familar shapes: including polygons and cirlces, but also other *conic sections* - parabolas, ellipses, and hyperbolas known to the ancient Greeks.

13.1. Polygons

Definition 13.1 (Polygon)**.** A **polygonal chain** is a sequence $L_1, L_2, \dots, L_n$ of line segments, where the ending vertex of L_n coincides with the starting vertex of L_{n+1}.

A polygon is a closed, non-intersecting polygonal chain. The interior of a polygon is called a *polygonal region*.

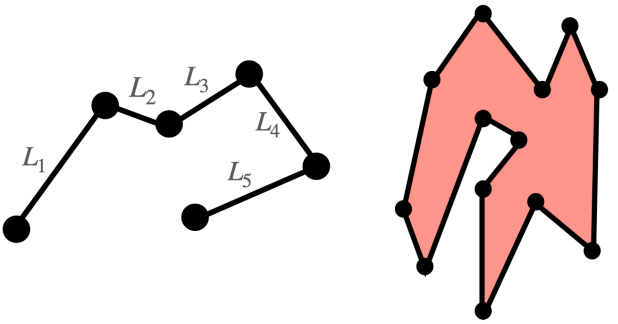


Figure 13.1.: A polygonal chain (left) and a polygon (right). The polygonal region is shaded red.

A triangle is a polygon with three sides, a quadrilateral is a polygon with four sides, and so on. We will study polygons in more detail later on, especially in the curved geometries of the sphere and hyperbolic space. But for now, we content ourselves in getting used to the definitions by re-proving some familiar results of the greeks.

Exercise 13.1 (Constructing an Equilateral Triangle)**.** Beginning with the segment $[0, \ell]$ along the x-axis, construct an equilateral triangle by finding the coordinates of a point $p = (x, y) \in \mathbb{E}^2$ which is equidistant from both endpoints of the segment.

Exercise 13.2 (Equilaterals of Half the Size, Reprise)**.** Re-prove that inside of an equilateral triangle, you can inscribe a smaller one with exactly half the side length. You already did this problem, using Euclid's Axioms, but now we can do it much easier in our new foundations!
Hint: just find where the vertices should be, and then measure the distances between them!

13.2. Circles

Euclid's definition of a circle is as follows: *A circle is a plane figure bounded by one curved line, and such that all straight lines drawn from a certain point within it to the bounding line, are equal.* In modern terminology, we may phrase this as below:

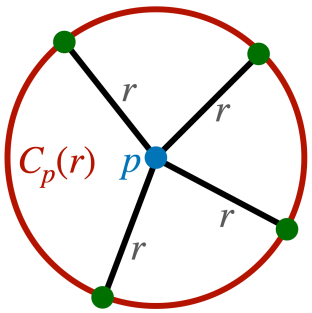


Figure 13.2.: The circle $C_p(r)$ is the set of points at distance r from a fixed point p.

Definition 13.2. A circle centered at a point $c \in \mathbb{E}^2$ is a curve such that the distance between p and any point on the curve is the same. This fixed distance is called the *radius* of the circle. We denote the circle of radius r centered at p as $C_p(r)$.

Now that we know the distance function on $\mathbb{E}^2$ we can formally write down the equation of a circle directly from this definition.

Theorem 13.1. *The circle of radius r centered at $p = (h, k)$ is given by the set*

$$C_p(r) = \{q = (x, y) \in \mathbb{E}^2 \mid (x - h)^2 + (y - k)^2 = r^2\}$$

Proof. This is just a direct computation: if $q = (x, y)$ is an arbitrary point in the plane, its distance from $p = (h, k)$ is

$$\text{dist}(q, p) = \|q - p\| = \sqrt{(x-h)^2 + (y-k)^2}$$

For q to lie on the circle, this distance needs to be set equal to r. The equation is easier to work with after squaring both sides to remove the square root, giving

$$(x-h)^2 + (y-k)^2 = r^2$$

□

Corollary 13.1 (Proving Euclid's Axiom III)**.** *Given any point p, and any radius r, the circle $C_p(r)$ exists.*

Proof. Now that we have the actual *equation* for a circle, Euclid's axiom is rather straightforwardly true. Saying we can draw a circle about any point of any radius is just asserting that the equation

$$(x-h)^2 + (y-k)^2 = r^2$$

has solutions for every h, k, r. And this in turn is just a property of the real numbers, and the existence of square roots! For simplicity considering the case centered at 0, for any $x \in [-r, r]$ we can solve directly for $y = \pm\sqrt{r^2 - x^2}$, which is a real number as $r > |x|$ so $r^2 - x^2$ is positive, and all positive reals have real square roots! □

One intuitive result about circles that we will use a lot in the near future is that any isometry of the plane that fixes the circles center must preserve the circle:

Proposition 13.1 (Isometries Fixing the Center Preserve the Circle)**.** *Let $C_c(r)$ be a circle centered at c, and ϕ be any isometry of $\mathbb{E}^2$ sending c to itself. Then ϕ preserves C: that is, if p is any point on C, $\phi(p)$ is also on C.*

Proof. Let $C_c(r)$ bea circle, and ϕ an isometry fixing c. If $p \in C_c(r)$ is any point, then by definition $\text{dist}(p, c) = r$. Applying the isometry ϕ, since isometries do not affect distances, we see

$$r = \text{dist}(p, c) = \text{dist}(\phi(p), \phi(c)) = \text{dist}(\phi(p), c)$$

Where the second inequality is because $\phi(c) = c$. But this says the distance from $\phi(p)$ to the circles center is also r, so $\phi(p)$ lies on the circle! Thus ϕ sends the circle to itself.

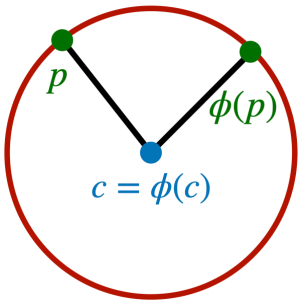


Figure 13.3.: If an isometry fixes the center of a circle, it sends the entire circle to itself.

□

Exercise 13.3. Prove that applying any isometry or similarity to a circle results in another circle.

Computations

Recall that in all of Euclid's axioms, conditions for intersections with circles were never specified! Indeed - Euclid intersected two circles in his construction of the equilateral triangle. Now that we have a precise description of circles in our new foundations, we can fix this gap:

Exercise 13.4. Prove that two circles intersect each other if the distance between their centers is less than or equal to the sum of their radii.

Hint: start by applying an isometry to move one of the circles to have center $(0, 0)$*, and then another isometry to roate everything so the second circle has center* $(x, 0)$ *along the x-axis. This will make computations easier!*

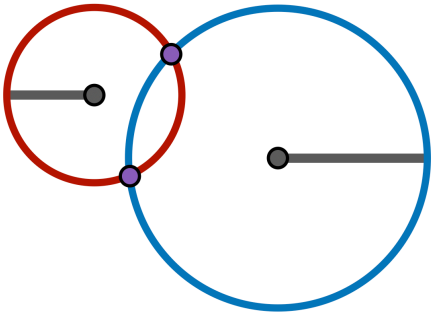


Figure 13.4.: Two circles intersect if the distance between their centers is less than or equal to the sum of their radii.

The other case that is interesting but does not appear in Euclid is the intersection of a circle an a line.

Exercise 13.5 (Circles Intersecting Lines)**.** Prove that a circle intersects a line whenever the shortest distance from that line to the circles center is less than the circles' radius.

Hint: start by applying an isometry that either (1) moves the line to the x axis, or (2): moves the circle's center to the origin - whichever one makes the computation easier for you!

13.3. Application: Classifying Isometries

The above exercises allowed us to compute exactly when two circles intersect - and crucially: how many times they do so. While our motivating reason to compute these things may have been to fill a gap in Euclid, this information can take us quite far when used correctly. Indeed, here we show that it is the key which allows us to classify *all possible isometries of the plane*!

We have discovered many isometries so far - translations, rotations (both about O and other points), reflections across lines, and all possible combinations thereof by composition and inversions. However, we have conducted no methodological search for isometries and so there is no good reason to think we are done. In fact, it seems dauntingly hard to ever prove we have found them all: who is to say that theres not some absurdly complicated function we have never thought of, that still preserves all infinitesimal lengths?

One way to begin making progress on this question is to ask how much information do you need to completely determine an isometry? Is it possible that there could be

two isometries that act like exactly the same function inside some region, but differ elsewhere? If so, that would mean it will be very hard to track down all isometries! But if not, we could ask ourselves what is the smallest region we need to understand an isometry on to specify it uniquely. And the answer is rather suprisingly minimal!

Proposition 13.2 (Isometries Fixing 3 Points)**.** *Let p, q, r be a triple of non-collinear points in the plane. If an isometry ϕ fixes all three points, then ϕ is the identity.*

Proof. Let p, q, r be three noncollinear points, and ϕ an isometry with $\phi(p) = p, \phi(q) = q$ and $\phi(r) = r$. We aim to show that ϕ is the identity: so we will consider an arbitrary point a in the plane, and show that $\phi(a) = a$. To do so, it will prove important to pay attention to the distance between a and the points p, q, r.

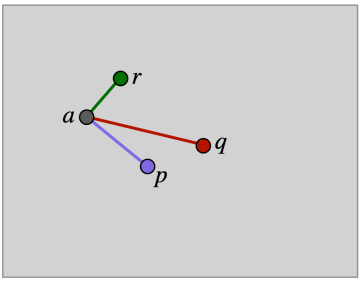


Figure 13.5.: The fixed points p, q, r and their distances to an arbitrary point a.

First, look at p. Since ϕ is an isometry we know $\text{dist}(a, p) = \text{dist}(\phi(a), \phi(p))$: but ϕ fixes p! Thus a and $\phi(a)$ are at the same distance from p, and lie on a circle centered at p.

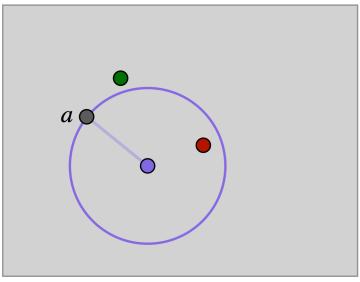


Figure 13.6.: Both $a and $\phi(a)$ must lie on the same circle centered at p.

As q is also fixed by ϕ, we similarly see that both a and $\phi(a)$ must lie on the same circle centered at q.

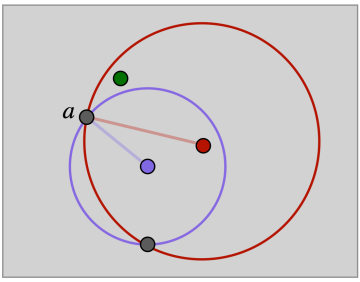


Figure 13.7.: BThe same is true for q. So both a and $\phi(a)$ must lie on both of these circles!.

However by Exercise 32.37, we know that generically circles will intersect in *two points*, so from the information we have so far, its not *guaranteed* that a and $\phi(a)$ are the same point: one could be at each intersection. But this isn't surprising as we haven't considered all the information at hand! We also know that ϕ fixed r and so both a and $\phi(a)$ must lie on the same circle centered at r.

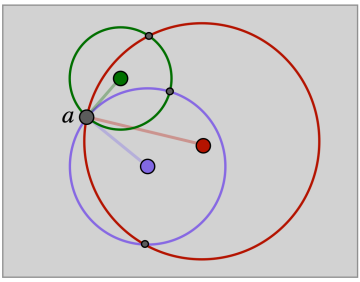


Figure 13.8.: Because p, q and r are noncollinear, these three circles cann only intersect at a single point! This point must be both a and $\phi(a)$, so ϕ fixes a.

Each pair of these circles intersects in two points. And, as the points are noncollinear, all *three* circles intersect in a *single point* (homework exercise, below). But this must be both a and $\phi(a)$: thus $\phi(a) = a$!

And a was an arbirary point, and we showed that ϕ did not move it. This means ϕ must not move any points at all in the plane - so ϕ is the identity! □

Remark 13.1. In special cases (when a lies at the midpoint of the line segment determined by p and q) the circles centered at p and q through a are *tangent* at a, making this their only point of intersection. In this case, we already know $a = \phi(a)$ even without considering r.

Exercise 13.6 (Circle Triple Intersection)**.** Three circles with non-collinear centers can intersect in at most one point.

Hint: use some isometries to simmplify the situation: move everything so that two of the circles center's are on the x axis, and the third is on the y axis.

PICTUREs

Corollary 13.2 (Isometries Agreeing on 3 Points are Equal)**.** *If ϕ and ψ are two isometries of the plane which agree on a set of three non-collinear points, then they are* equal.

Proof. Let p_1, p_2, p_3 be three points for which $\phi(p_i) = \psi(q_i)$. Now consider the isometry $\psi^{-1} \circ \phi$. This isometry actually fixes each of the three p_i (since ϕ sends them somewhere, and ψ^{-1} brings them back). Thus, by Proposition 13.2 this is the identity. But if $\psi^{-1}\phi = \mathrm{id}$ then composing with ψ shows

$$\phi = \psi$$

□

Using this, we can prove that we have actually found all the isometries, by starting from an arbitrary isometry and building it using only translations, rotations, and reflections.

Theorem 13.2 (Classification of Isometries)**.** *Every isometry of $\mathbb{E}^2$ is a composition of reflection, rotation, and translation.*

Proof. Let ϕ be an arbitrary isometry of the plane. and consider the three points O, $p = (1, 0)$ and $q = (\frac{1}{2}, \frac{\sqrt{3}}{2})$ forming the vertices of an equilateral triangle.

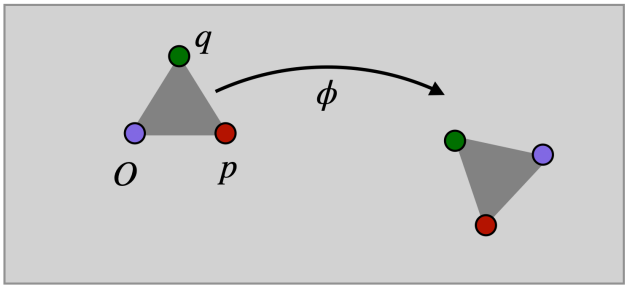


Figure 13.9.: An isometry taking pne triple of points to another.

The isometry ϕ sends O somewhere: let T be a translation of the plane taking $\phi(O)$ to O. Now, the composition $T\phi$ *fixes the origin*! The point p lies at distance 1 from O, so the $T\phi$ takes it to another point unit distance away - call this point r.

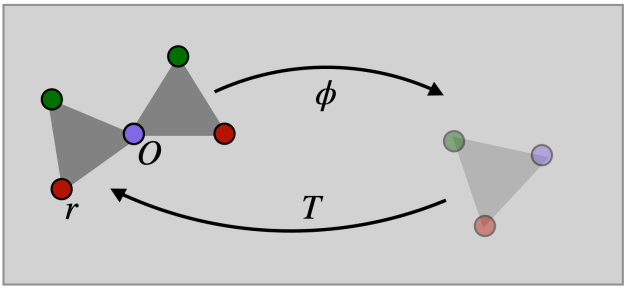


Figure 13.10.: The isometry $T\phi$ fixes the origin.

Let R be a rotation isometry fixing O but taking r back to $p = \langle 1, 0 \rangle$. Now, the composition $RT\phi$ fixes both O and p! All that remains is to think about where q has been sent.

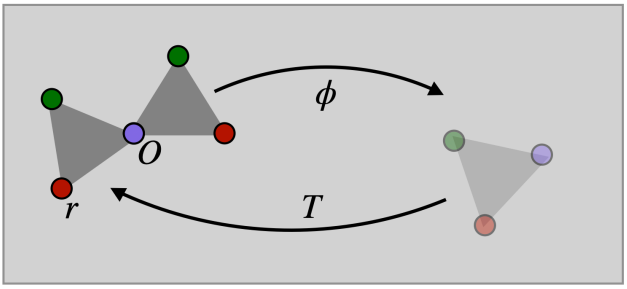


Figure 13.11.: The isometry $T\phi$ fixes the origin.

Since $RT\phi$ is an isometry it preserves distances, so it must send q to a point which is unit distance from O and p (remember - it send O and p to themselves)!

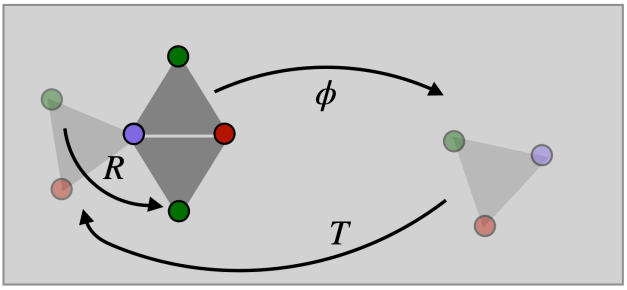


Figure 13.12.: The isometry $RT\phi$ fixes the both O and p, so there are only two options on where it sends q. Here we see the case where it does *not* get sent directly to q.

There are only two such points in the plane, which lie at the intersections of the cirlces

$$x^2 + y^2 = 1 \qquad (x-1)^2 + y^2 = 1$$

The two options are either q itself - $(\frac{1}{2}, \frac{\sqrt{3}}{2})$ or the same point with the y-coordinate negated. So, let F be the following isometry: its the identity if $RF\phi(q)$ was already equal to q, and is the reflection in the x axis otherwise.

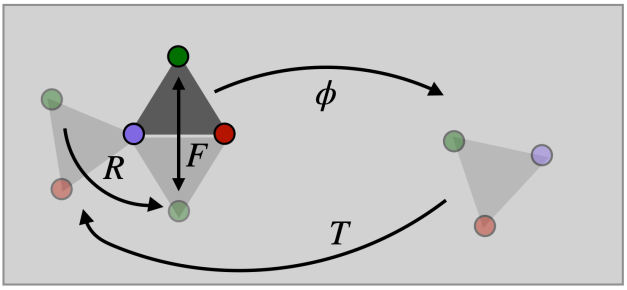


Figure 13.13.: Composing $RT\phi$ with a flip F across the x axis creates an isometry that fixes all of O, p and q.

Now, we have the map $FRT\phi : \mathbb{E}^2 \to \mathbb{E}^2$, which fixes all three points of the equilateral triangle! This is the same thing the identity map does on these three points, so by Corollary 13.2 this must actually be the identity!

$$FRT\phi = \mathrm{id}$$

One by one composing with the inverses of the maps we've added on, we can now solve for ϕ:

$$\phi = T^{-1}R^{-1}F^{-1}$$

Since the inverse of a translation is just another translation, the inverse of a rotation is another rotation, and the inverse of a reflection is another reflection (its the same reflection, actually!) we see ϕ is a composition of translation, rotation and reflection as claimed. □

Thus - every isometry is built from the building blocks we already know of: there are no new mystery isometries out there to be discovered! This is a powerful fact as it lets us claim things about *all isometries* by just checking them with rotaitons, translations, and reflections. For example:

Corollary 13.3 (Isometries are Affine Maps)**.** *All Euclidean isometries are affine maps of the plane.*

Proof. Let ϕ be an arbitrary Euclidean isometry. By Theorem 13.2, we may write $\phi = TRF$ for T a translation, R a rotation about the origin, and $F(x, y) = (x, \pm y)$ either a flip across the x axis or the identity. Thus, the proof is just a direct computation: a rotation about the origin is given by a linear map (Theorem 11.4) so

$$RF(x, y) = \begin{pmatrix} u & -v \\ v & u \end{pmatrix}\begin{pmatrix} x \\ \pm y \end{pmatrix} = \begin{pmatrix} ux \pm vy \\ -vx \pm uy \end{pmatrix}$$

Next, any translation is an affine map of the form $T(x, y) = (x + a, y + b)$ So

$$\phi(x, y) = TRF(x, y) = T\begin{pmatrix} ux \pm vy \\ -vx \pm uy \end{pmatrix} = \begin{pmatrix} ux \pm vy + a \\ -vx \pm uy + b \end{pmatrix}$$

Each of these components is an affine function, so the entire isometry ϕ is affine. □

13.4. Conic Sections

Though greek geometry lacked the ability to deal with general curves, they did know quite a lot about a specific family of curves called *conic sections* These include the circles and lines we have already discussed, but also parabolas, ellipses, and hyperbolas. We will not spend much time on them here, except to show how our new formalism lets us come up with *precise equations* for these curves much as it did for lines and circles already.

Remark 13.2. These curves are called conic sections because we can alterantively define them as the possible curves one can get by slicing a 3-dimensional cone by a plane at different angles.

13.4.1. Parabolas

A geometric definition of the parabola dating back to ancient greece is that it is the set of points which lie at an equal distance from a given point and line in the plane. More precisely

Definition 13.3. Let L be a line (called the directrix), and f a point not on that line (called the focus). The *parabola* P with directrix L and focus f is the set of points (x, y) which lie at the same distance from ℓ as they do from f:

$$\text{dist}((x, y), L) = \text{dist}((x, y), p)$$

Recall that the distance to a line L is defined as the *shortest distance between* (x, y) *and any point on* L *(Definition 12.8).*

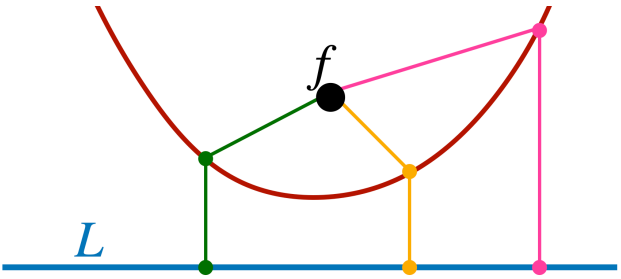


Figure 13.14.: A parabola is the set of points which are the same distance from a point (the focus) and a line (the directrix). In this figure, line segments of the same color are supposed to be the same length.

Exercise 13.7. Let L be the x-axis, and f the point $(0, 2)$ along the y-axis. Find an equation that points (x, y) on the parabola determined by L and f must satisfy.

Exercise 13.8. In this problem we confirm that $y = x^2$ is indeed a parabola! Let L be a horizontal line intersecting the $y-$axis at some point $(0, -\ell)$, and $f = (0, h)$ be a point along the y-axis for $\ell, h > 0$.

- Write down an algebraic equation for the coorddinates of a point (x, y) determining when it is on the parabola with focus f and directrix L.
- Find which point f and line L make this parabola have the algebraic equation $y = x^2$.

13.4.2. Ellipses & Hyperbolas

Ellipses and hyperbolas are also defined by a condition involving distance. Instead of distance from a single point (circles) or distances between a point and a line (parabolas), these shapes involve the distances from a pair of points.

Definition 13.4. Let f_1 and f_2 be two points in the plane (called focii), and d a number (called the distance sum). The ellipse with focii f_1, f_2 and distance sum d is the set of points $p = (x, y)$ in the plane where

$$\operatorname{dist}(p, f_1) + \operatorname{dist}(p, f_2) = d$$

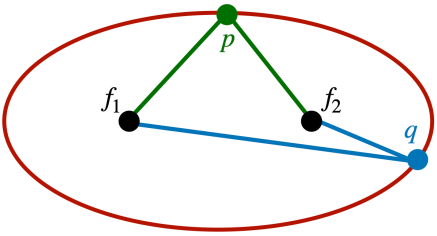


Figure 13.15.: An ellipse is the set of points where the sum of distances to two points is constant. Here, p and q are both on the ellipse as the lengths of the green and blue polygonal segments are equal.

Exercise 13.9. Find an equation of the form $ax^2 + by^2 = c$ determining when a point lies on the ellipse with focii $(1, 0)$ and $(-1, 0)$ with distance sum 4.

A hyperbola is defined similarly, except it is a difference of distances instead of a sum:

Definition 13.5. Let f_1 and f_2 be two points in the plane (called focii), and d a number (called the distance difference). The hyperbola with focii f_1, f_2 and distance sum d is the set of points $p = (x, y)$ in the plane where

$$|\operatorname{dist}(p, f_1) - \operatorname{dist}(p, f_2)| = d$$

This distance difference may be positive or negative (hence the absolute value). Each sign determines one *branch* of the hyperbola - its a disconnected curve with two components!

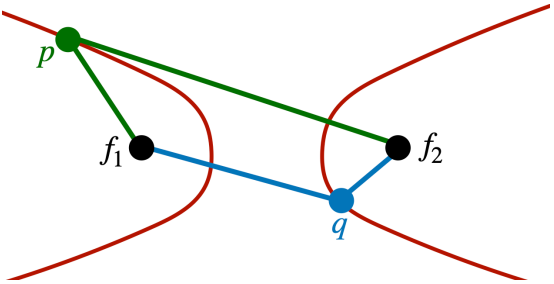


Figure 13.16.: An hyperbola is the set of points where the differece of distances to two points is constant. Here, p and q are both on the hyperbola as the difference between the lengths of the green segments equals the difference between the blue segments.

Exercise 13.10. Prove that the equation $y^2 - x^2 = 1$ determines a hyperbola. What are the two focii? What's the distance difference?

13.4.3. Equidistants to a Line

This is not itself a conic section, but like the previous shapes we've discussed fits into the class of "shapes defined by a distance constraint."" A circle is the set of points which are *equidistant from a point* (its center). One could attempt to generalize this notion by replacing the point at the center with something more general, and measuring

The most reasonable "generalized center" to consider first is a line: it is the only other shape we know so far, after all! What curves are equidistant to a line? In fact, the answer here is not that interesting: its just a pair of two lines.

Proposition 13.3. *Given a line L, the set of points lying at distance d from L are two disjoint lines.*

Proof. Let L be a line, and choose an isometry ϕ that moves L to the x-axis (which we will denote X). Now the distance from a point $p = (a, b)$ to a point $(x, 0)$ on X is $\sqrt{(x-a)^2 + b^2}$, which is minimized when $x = a$: thus

$$\operatorname{dist}((a, b), X) = \sqrt{b^2} = |b|$$

Thus, the set of points at distance r from X contains all pairs (x, r) and $(x, -r)$ for any x: these are two lines

$$L_+(x) = (x, r) \qquad L_-(x) = (x, -r)$$

To solve the original problem, we apply the inverse isometry ϕ taking X back to the line L. Since isometries take lines to lines and preserve distance, this takes $L_\pm$ to two lines, each at distance r from L as claimed. □

However, we include brief mention of this fact for two reasons. One, in a lost work of Archimedes, *On Parallel Lines*, it seems that he tried to work with alternative definitions of parallelism based on this fact, to simplify Euclids theory. Below is a quote from Boris Rosenfeld's *A History of Non-euclidean Geometry*:

> It seems that the first work devoted to this question [the theory of parallels] was Archimedes' lost treatise *On parallel lines* which appeared a few decades after Euclid's Elements. [...] it is very likely that Archimedes used a definition of parallel lines different from Euclid's. [...] it is possible that Archimedes based his definition of parallel lines on distance.

And secondly, while we found here that the equidistant curves to a line are just another pair of lines, this fact (as presciently investigated by Archimedes) is actually *equivalent* to the parallel postulate, and so will be false in the other geometries we study! Thus, we will reference this short section in those future geometries, to contrast our new discoveries with the old.

14. Angles

In the geometry of Euclid, an angle was defined as being delimited by straight lines:

> A plane angle is the inclination to one another of two lines in a plane which meet one another and do not lie in a straight line.

However as Greek mathematics (and beyond) turned to the problem of curves, it became necessary to also speak of *curvilinear angles*: that is, the angle of intersection between two curves. This was a difficult concept, as at no finite level of zoom could this be made into a "true angle", the sides were never going to be straight.

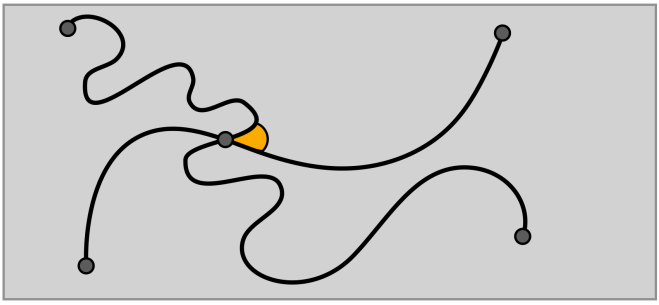


Figure 14.1.: How can we define the angle between two curves?

From the modern perspective this is no issue, as zooming in on the point of intersection we may pass to the tangent space, and replace each of the curves by their linearizations. This allows us to think of angles as *infinitesimal quantities* based at a point.

Definition 14.1. An angle α at a point $p \in \mathbb{E}^2$ is an ordered pair of tangent vectors $\alpha = (v, w)$ based at p.

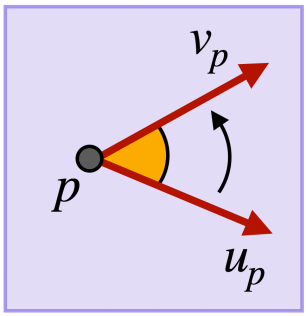


Figure 14.2.: An angle is defined by an ordered pair of two tangent vectors at a point.

The order of the tangent vectors tells us which curve is "first" and which is "second" as we trace out the angle. By convention, we will trace the angle *counterclockwise* from start to finish.

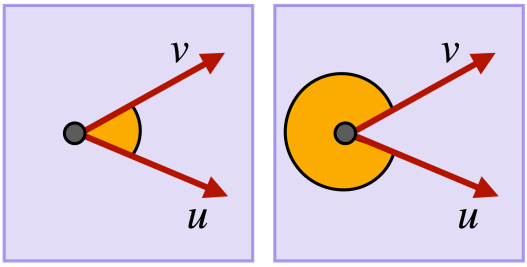


Figure 14.3.: The angle measure of (u, v) on the left, versus the measure of (v, u) on the right. Both measured as positive angles.

(Occasionally, we wish to read an angle clockwise instead: in this case we will say that it is a *negative angle*, whereas counterclockwise default angles are *positive*)

From this, we can define the angle between two curves in terms of their tangents:

Definition 14.2. Given two curves C_1, C_2 which intersect at a point p in the plane, and let v_1 be tangent to C_1, and v_2 be tangent to C_2 at p. Then the angle from C_1 to C_2 is just the pair (v_1, v_2) of tangents.

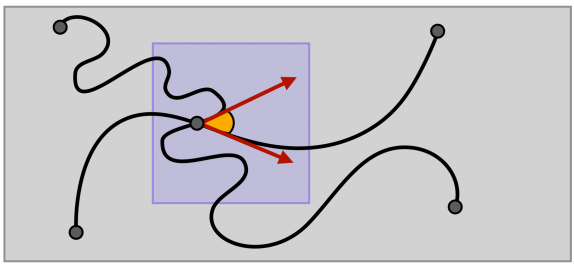


Figure 14.4.: The angle between curves, defined at a point of intersection by their tangent vectors.

14.1. Angle Measure

In all of Euclid's elements, angles were not measured: by numbers. There was a definition of a right angle (half a straight angle), and definitions of acute and obtuse (less than, or greater than a right angle, respectively). Though they never attached a precise number, they did have the *angle axioms* specifying how to work with angles *like they were a kind of number*, however.

In our modern development we find numerical measures extremely convenient: if we can measure angles with a function, we can do calculus with angles! So we want to go further, and an actual number to each angle (which we'll call its *measure*) in a way that's compatible with the original angle axioms.

How do we construct such a number? At the moment we do not have much to work with, as our development of geometry is still in its infancy: we have essentially only constructed lines, circles and the distance function. These strict constraints essentially force a single idea for angle measure upon us:

Definition 14.3 (Angle Measure)**.** If u and v are two unit vectors based at the same point p forming angle $\alpha = (u, v)$, then the measure of α is defined as the arclength of the unit circle centered at p that lies between them. Its denoted

$$\text{Angle}(\alpha) \text{ or } \text{Angle}(u, v)$$

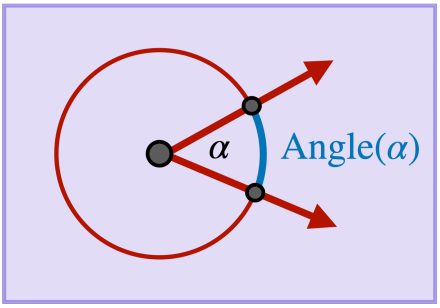


Figure 14.5.: The measure of an angle is defined in terms of the arclength of a circle.

This is a very "basic" way of dealing with angles, as it uses so few concepts from our geometry (its close to the base definitions). Indeed, its geometric simplicity makes it exceedingly useful throughout mathematics, you've all met this definition before under the name *radians*.

Because angles are defined in terms of *unit circle arclength*, it will prove very convenient to have a name for the entire arclength of the unit circle. That way we can express simple angles as fractions of this, instead of as some long (probably irrational) decimal representing their arclength. We will denote the arclength of the unit circle by τ, standing for *turn* (as in, one full turn of the circle)

Definition 14.4 (τ)**.** The arclength of the unit circle is τ.

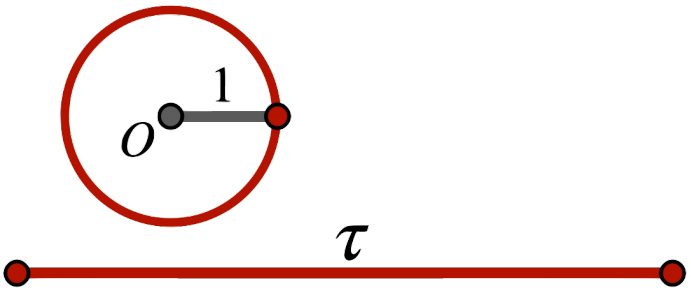


Figure 14.6.: The circumference τ

The first thing we may wish to explore is how this concept interacts with *isometries*.

Proposition 14.1 (Angles Measures are Invariant under Isometries)**.** *If α and β are two angles in $\mathbb{E}^2$, and ϕ is an isometry taking α to β, then the measures of α and β are equal.*

Proof. Let α and β be two angles: so precisely α is a pair of tangent vectors a_1, a_2 based at a point p, and β is a pair of tangent vectors b_1, b_2 at a point q.

Any isometry taking p to q takes the unit circle based at p to the unit circle based at q (since isometries preserve the distance function). And, if the isometry takes a_1 to b_1 as well as a_2 to b_2, it takes the arc of the unit circle defining α to the arc defining β.

Since isometries preserve the length of all curves, the lengths of these arcs must be the same. Thus these angles have the same measure. □

Because angle measures are defined in terms of the *unit circle*, we can also attempt to run the above argument with a similarity σ instead of isometry. If the scaling factor is k, the main change is that σ takes the unit circle to a circle of radius k (as similarities take circles to circles, Exercise 13.3). We know how similarities affect lengths - so the length of this arc is k times the original angle measure. But to correctly compute the new angle, we need to be measuring on the *unit circle*. So, we need to rescale it down by $1/k$ a similarity. This then divides the length by k, and overall we see the new length is identical to the original: so the angle is the same!

Corollary 14.1 (Angle Measures are Invariant Under Similarities)**.** *If α and β are two angles in $\mathbb{E}^2$, and σ is an similarity taking α to β, then the measures of α and β are equal.*

Computing an angle *directly from this definition* is challenging, as it requires us to measure arclength. Much of the later work in this chapter will establish a beautiful means of doing this. But in certain situations, angles can be measured directly by more elementary means.

Example 14.1 (Angle between x- and y-axes)**.** The angle from $(1, 0)$ to $(0, 1)$ is $\tau/4$. To see this, recall that we have a *rotation isometry* fixing the origin and taking $(1, 0)$ to $(0, 1)$ - this was the first rotation we discovered (Example 11.1).

$$R(x, y) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

Now, look what happens when you apply R multiple times in a row. This repeated composition is actually straightforward to compute, as it's just matrix multiplication!

$$R^2 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \quad R^3 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad R^4 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Looking at the last line, we see that applying R four times in a row results in the *identity matrix* - or the transformation that does nothing to the inputs! That is, after four rotations we are back to exactly where we have started.

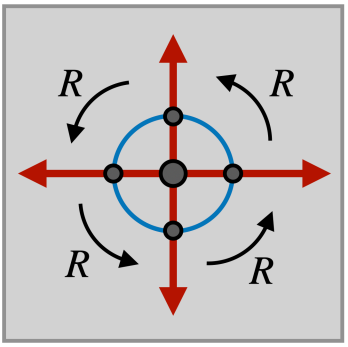


Figure 14.7.: The angle from $\langle 1, 0\rangle$ to $\langle 0, 1\rangle$ is one quarter of the unit circle.

Because isometries preserve angles (Proposition 14.1), we see that our isometry R has covered the entire unit circle in exactly four copies of our original angle. Thus, the angle measure must be $1/4$ of a circle:

$$\theta = \frac{\tau}{4}$$

(This example tells us our first rotation angle: the matrix $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ rotates the plane by a quarter turn or $\tau/4$! This will be a very useful fact, and we will even put it to use shortly, in Proposition 14.3 where we find the derivative of sin and cos.)

Exercise 14.1 (Angle Measure of Equilateral Triangle)**.** Show the angle measure of an equilateral triangle is $\tau/6$, in a similar method to the example above.

To start, draw an equilateral triangle with unit side length, and one side along the x-axis. In Exercise 32.30, you found that the third vertex of this must lie at $p = (1/2, \sqrt{3}/2)$. From this, we can write down a rotation isometry (Theorem 11.4) taking $(1, 0)$ to p. The angle this rotates by is exactly the angle our triangle's sides make at the origin.

Show that if you apply this rotation three times, you get negative the identity matrix. Use this to help you figure out how many times you have to apply it before you get back to the identity! Then use that isometries preserve angles, and the circumference of the unit circle is τ to deduce the angle you are after.

Proposition 14.1 and Example 14.1, Exercise 14.1 showcase two essential properties of the angle measure which stem from the fact that we defined it as a *length*: its invariant under isometries, and easy to subdivide. In fact, these are precisely the *angle axioms* of the greeks!

Example 14.2 (Proving the "Angle Axioms")**.** Now that we have a definiiton of angle measure in terms of more primitive quantities (vectors, lengths, circles), we can *prove* that this measure satisfies the greek axioms.

- **Congruent Angles have Equal Measures** Since two angles are congruent if there is an isometry taking one to the other, and the measure of an angle is invariant under isometries, congruent angles have the same measure.
- **Subdividing an Angle** If we divide an angle θ into two angles θ_1, θ_2 by a line, then $\theta = \theta_1 + \theta_2$. This follows directly from a property of *integrals*! Since an angle is a *length*, and a *length is an integral* we can use the property $\int_a^b f dx = \int_a^c f dx + \int_c^b f dx$ to prove $\theta = \theta_1 + \theta_2$.

14.2. Working With Angles

We now turn to the problem of actually *computing* things with the angle measure. To do so, it's helpful to choose a "basepoint" on the circle to take first measurements from - here we'll pick $(1, 0)$.

Definition 14.5 (Arclength Function)**.** The arclength function takes in a point on the unit circle (x, y), and measures the arclength θ from $(1, 0)$ to this point.

$$\Theta(x, y) = \text{arclength from } (1, 0) \text{ to } (x, y)$$

In this section, we will study in detail expressions for this function, and its inverse.

14.2.1. Arc~ Sine and Cosine

Because the points of the unit circle satisfy $x^2 + y^2 = 1$, if we know the sign of x, y (which half of the circle the point lives in) we can fully reconstruct the point from a single coordinate: either $(x, \pm\sqrt{1 - x^2})$ or $(\sqrt{1 - y^2}, y)$. Thus, so long as we remember the correct sign of the second coordinate of interest, the arclength function is essentially a *function of one variable.* We could take just the x or y coordinate of a point on the circle and define a function like $\Theta(x)$ which would measure the arclength to $(x, \sqrt{1 - x^2})$, or $\Theta(y)$ if we expressed the point $(\sqrt{1 - y^2}, y)$. However, we do not need to invent our own notation for these arclength functions of one variable, they are already well known to us from trigonometry!

Definition 14.6 (Arc Functions: Inverse Trigonometry)**.** The functions arccos and arcsin compute the arclength along the unit circle from $(1, 0)$ to the point (x, y) as

$$\arccos(x) = \text{arclength from } (1, 0) \text{ to } (x, \sqrt{1 - x^2})$$

$$\arcsin(y) = \text{arclength from } (1,0) \text{ to } (\sqrt{1-y^2}, y)$$

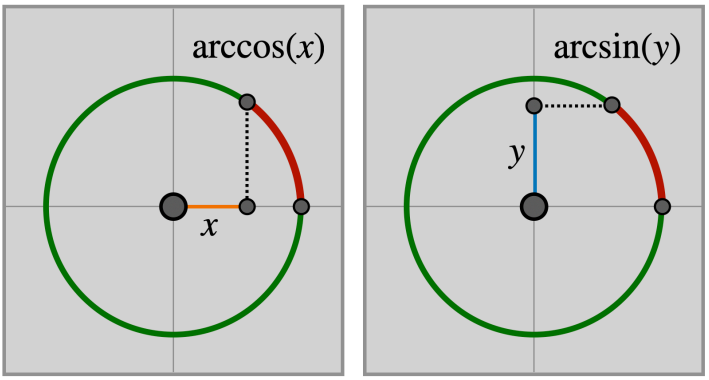


Figure 14.8.: The arclength functions given the x and y coordinates of a point on the circle.

This of course does nothing to help us *compute* these functions: we've just given a name to them. In fact, we can compute only very few values from first principles:

Example 14.3 ("Unit Circle Values" for Arc Functions)**.** -The point $(1,0)$ lies at arclength zero from $(1,0)$...as they are the same point! Thus,

$$\arccos(1) = 0 \qquad \arcsin(0) = 0$$

- The point $(0,1)$ lies a quarter of the way around the circle (Example 14.1), so has arclength $\tau/4$. Thus we see

$$\arccos(0) = \tau/4 \qquad \arcsin(1) = \tau/4$$

- Looking at Exercise 14.1 where we found the angle of a unit equilateral triangle with sides vertices $(0,0), (1,0)$ and $(1/2, \sqrt{3}/2)$ to be $\tau/6$, we see

$$\arccos(1/2) = \tau/6 \qquad \arcsin(\sqrt{3}/2) = \tau/6$$

What we need is some sort of concrete expression telling us how to compute $\arccos(x)$ (or arcsin) for arbitrary values of x. And here we can make essential use of our definition of angle as a length, and length as an integral! Unpacking it all gives directly an *integral formula* to compute $\arccos(x)$:

Proposition 14.2 (Integral Representation of $\arccos(x)$)**.** *For $x \in [-1,1]$, the arccosine function can be computed via an integral*

$$\arccos(x) = \int_x^1 \frac{1}{\sqrt{1-t^2}} dt$$

Proof. By definition, the $\arccos(x)$ is the length of circle between x and 1. Since $x^2 + y^2 = 1$, we may parameterize the top half of the circle using x as the parameter by

$$c(t) = (t, \sqrt{1 - t^2})$$

Thus, as lengths are integrals, we can already give an expression for $\arccos(x)$:

$$\arccos(x) = \int_x^1 \|c'(t)\| dt$$

To be useful, we must expand out the integrand here, and compute $\|c'\|$:

$$c'(t) = \left(1, \frac{-t}{\sqrt{1 - t^2}}\right)$$

And then, we must find its norm:

$$\begin{aligned} \|c'(t)\| &= \sqrt{1 + \left(\frac{-t}{\sqrt{1 - t^2}}\right)^2} \\ &= \sqrt{1 + \frac{t^2}{1 - t^2}} \\ &= \sqrt{\frac{1}{1 - t^2}} \end{aligned}$$

Thus, we have an integral representation of the arccosine!

$$\arccos(x) = \int_x^1 \frac{1}{\sqrt{1 - t^2}} dt$$

□

If we start at $x = -1$ and go to $x = 1$, this curve traces out exactly half of the unit circle (the top half). Thus, twice this value is an integral representing our fundamental constant τ:

Corollary 14.2 (Defining τ:)**.**

$$\tau = \int_{-1}^1 \frac{2}{\sqrt{1 - x^2}} dx$$

Exercise 14.2. Complete an analogous arugment to the above to show

$$\arcsin(y) = \int_0^y \frac{1}{\sqrt{1 - t^2}} dt$$

These formulas, via the fundamental theorem of calculus, tell us the derivative of arcsin and arccos as well!

Corollary 14.3 (Differentiating the Arc Functions)**.** *The derivative of the arc functions are*

$$\frac{d}{dx}\arccos(x) = \frac{-1}{\sqrt{1-x^2}}$$

$$\frac{d}{dy}\arcsin(y) = \frac{1}{\sqrt{1-y^2}}$$

Proof. Each of these is an immediate application of the fundamental theorem of calculus: but there's a small subtlety in how we usually apply this theorem to the first, so we will start with arcsin.

The fundamental theorem says that $\frac{d}{dx}\int_a^x f(t)dt = f(x)$, so we immediately get

$$\frac{d}{dy}\arcsin(y) = \frac{d}{dy}\int_0^y \frac{1}{\sqrt{1-t^2}}dt = \frac{1}{\sqrt{1-y^2}}$$

The difficulty with arccosine is that in the way we have it written, the variable x is the *lower bound* of the integral. To prepare this expression for an application of the fundamental theroem, we must first switch the bounds, which *negates the integral*. Thus

$$\frac{d}{dx}\arccos(x) = \frac{d}{dx}\int_1^x \frac{-1}{\sqrt{1-t^2}}dt = \frac{-1}{\sqrt{1-x^2}}$$

This is where the negative sign in the derivative of arccosine comes from: I have to remind myself of this every time I teach calculus 1. □

14.2.2. Sine and Cosine

Now that we have the functions that measure arclength, its natural to ask about their *inverses*: if we know the arclength from $(1,0)$ to a point, can we recover the coordinates of the point?

Definition 14.7 (Sine and Cosine)**.** Let p be the point on the unit circle centered at O which lies at a distance of θ in arclength from $(1,0)$. Then we define $\cos(\theta)$ as the x-coordinate of p, and $\sin\theta$ as the y-coordinate of p.

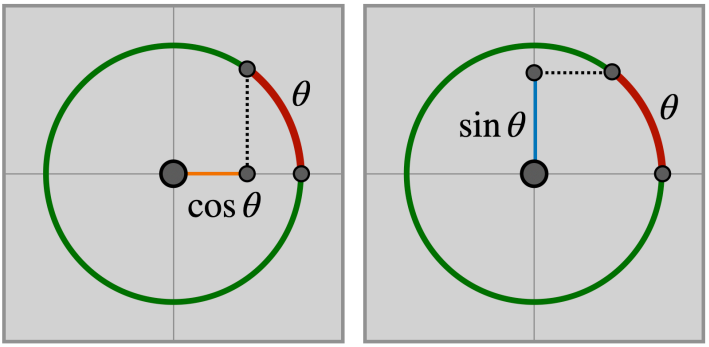


Figure 14.9.: The functions sin and cos return the cartesian coordinates of a point lying at arclength θ along the circle.

Example 14.4. At $\theta = 0$, we have moved no distance along the circle from $(1, 0)$, so we are still at $(1, 0)$. Thus

$$\cos(0) = 1 \qquad \sin(0) = 0$$

(Because sine and cosine are defined as lengths, which are invariant under isometries, we see that we could equally well define these functions from a unit circle centered at any point in $\mathbb{E}^2$. After the chapter on symmetry, we will see that we can further generalize to base them on any circle whatsoever.)

Beyond this, the definition doesn't give us any means at all of calculating the value of cos or sin: we're going to need to do some more work to actually figure out what these functions are! For their inverses, the secret was unlocked by integration, and so it makes sense that here we must do the opposite, and look to differentiation for help!

Proposition 14.3 (Differentiating Sine & Cosine)**.** *The derivatives of* $\sin(\theta)$ *and* $\cos(\theta)$ *are as follows:*

$$\frac{d}{d\theta}\sin\theta = \cos\theta \qquad \frac{d}{d\theta}\cos\theta = -\sin\theta$$

There are many beautiful geometric arguments for computing the derivative of $\sin\theta$ and $\cos\theta$, involving shrinking triangles and side ratios. Below is a different style argument, which fits well with the calculus-first perspective of our course. We use the fact that the derivative gives the *tangent line* to figure out what it must be!

Proof. Let C be the unit circle centered at $(0, 0)$, and $\gamma(t) = (\cos(\theta), \sin(\theta))$ be the arclength parameterization defined above. Because γ traces out the circle with respect to arclength, its derivative *with respect to arclength* is unit length, by definition.

We start by finding the derivative at $(1,0)$. As the radius of the circle is 1 and the x-coordinate at $\gamma(0) = (1,0)$ equal to 1, we are at a *maximum value of the x-coordinate*. Differential calculus (Fermat's Theorem) tells us that at a maximum, the derivative is zero. So, $x' = 0$ at $(1,0)$. But y' cannot also be zero here: as y is increasing as we trace out the circle, differential calculus (a corollary of the Mean Value Theorem) tells us that $y'(0)$ is positive. But now the fact that $\|\gamma'(0)\| = 1$ uniquely singles out a vector:

$$\gamma'(0) = \langle 0, 1 \rangle_{(1,0)}$$

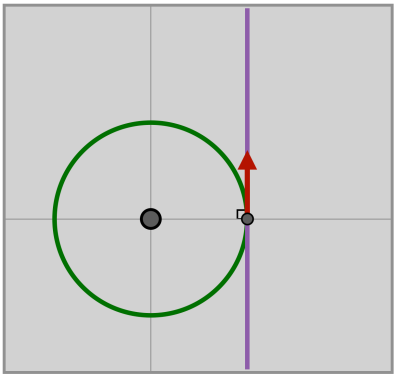


Figure 14.10.: The tangent line to the unit circle at $(1,0)$ is vertical, with unit tangent vector $\langle 0, 1 \rangle$.

This is actually all the differentiation we have to do! The rest of the argument amounts to a clever use of isometries. Choose a point $q = \gamma(\theta) = (\cos(\theta), \sin(\theta))$ on the circle, where we wish to compute the tangent vector. Now create an isometry ϕ taking $(1,0)$ to q. Since rotations about the center preserve circles (Proposition 13.1), the curve $\phi \circ \gamma$ also traces out the unit circle. Thus, if we find the tangent to $\phi \circ \gamma$ at q we'll have found the tangent to the circle at q!

At $(1,0)$, we saw the tangent vector was a $\tau/4$ rotation from the vector connecting its basepoint to the origin. Since isometries preserve angles (Proposition 14.1), it must *also be true* that the tangent vector at q is a $\tau/4$ rotation of $q - O = \langle \cos\theta, \sin\theta \rangle_q$. And we know how to rotate a vector by $\tau/4$: switch its coordinates and negate the first (Example 11.1)!

$$\langle \cos\theta, \sin\theta \rangle_q \mapsto \langle -\sin\theta, \cos\theta \rangle_q$$

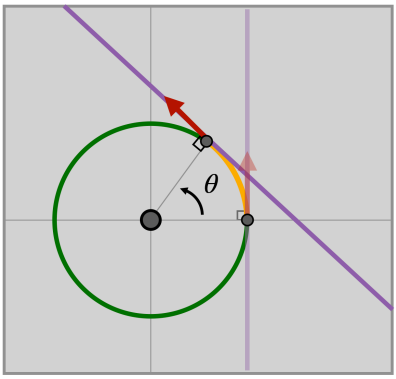


Figure 14.11.: At a point θ along the circle, the tangent can be found by symmetry. Since the tangent at $(1, 0)$ is orthogonal a $\tau/4$-rotation of the position, the same must be true at every point of the circle. Thus, at $(\cos\theta, \sin\theta)$ it is $\langle -\sin\theta, \cos\theta\rangle$.

Since the tangent vector to the the circle at q is the derivative of γ at θ, this tells us exactly what we were after:

$$\begin{aligned}\gamma'(\theta) &= \langle(\cos\theta)', (\sin\theta)'\rangle \\ &= \langle -\sin\theta, \cos\theta\rangle\end{aligned}$$

□

Remark 14.1. An alternative to the first step of this proof is to consider that the circle is sent to itself under the isometry $\phi(x, y) = (x, -y)$. This map fixes the point $(1, 0)$, and so it must send the tangent line at $(1, 0)$ to itself. But as ϕ is linear, its derivative is itself, so it applies to tangent vectors also as $\phi(v_1, v_2) = \langle v_1, -v_2\rangle$. Thus, whatever the tangent vector at $(1, 0)$ is, it must be a vector such that $\langle v_1, v_2\rangle$ is parallel to $\langle v_1, -v_2\rangle$. This forces $v_1 = 0$, and then unit length forces $\langle 0, \pm 1\rangle$.

Exercise 14.3. Because of our hard work with the arc functions already, we have an alternative approach to differentiating sine and cosine, using purely the rules of single variable calculus!

- Explain why from the definition of sin, cos we know that $\sin^2\theta + \cos^2\theta = 1$
- Use the technique for differentiating an inverse () to differentiate sin as the inverse function of arcsin, whose derivative we know.
- Combine these two facts to simplify the result you got, and show $\sin(\theta)' = \cos(\theta)$.
- Repeat similar reasoning to show $\cos(\theta)' = -\sin(\theta)$.

Remark 14.2. An alternative second part to this proof is just to write down the matrix: we know the rotation taking $(1,0)$ to (p_1,p_2) is $\left(\begin{smallmatrix} q_1 & -q_2 \\ q_2 & q_1 \end{smallmatrix}\right)$, so for $q = (\cos t, \sin t)$ its $\left(\begin{smallmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{smallmatrix}\right)$. This is linear so we can apply it to both points and vectors: applying to the tangent vector $\langle 0,1\rangle_{(1,0)}$ just reads off the second column! Thus the tangent vector at q is $\langle -\sin\theta, \cos\theta\rangle_q$.

Believe it or not - we already have enough information to completely understand the sine and cosine functions! Since they are each other's derivatives, and we *know both values at zero*, we can directly write down their series expansions!

Proposition 14.4 (Series Expansions of Cos)**.** *The series expansion of the cosine function is*

$$\begin{aligned}\cos\theta &= \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}\theta^{2n} \\ &= 1 - \frac{\theta^2}{2} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \frac{\theta^8}{8!} - \frac{\theta^{10}}{10!} - \cdots\end{aligned}$$

Proof. We first build what the series *ought to be* (assuming it exists), and then we prove that our candidate actually converges! Assume that $\cos\theta = a_0 + a_1\theta + a_2\theta^2 + \cdots$ for some coefficients a_n. Evaluating this at $\theta = 0$ we see

$$\cos 0 = a_0 + a_1 \cdot 0 + a_2 \cdot 0^2 + \cdots = a_0$$

Since we know $\cos 0 = 1$ this says $a_0 = 1$, and we have determined the first term in the series:

$$\cos\theta = 1 + a_1\theta + a_2\theta^2 + a_3\theta^3 + a_4\theta^4 \cdots$$

So now we move on to try and compute a_1. Taking the derivative of this series gives

$$(\cos\theta)' = a_1 + 2a_2\theta + 3a_3\theta^2 + \cdots$$

And again - every term except the first has a θ in it, so evaluating both sides at zero gives us a_1. Using that cosine's derivative is $-\sin\theta$, we get $-\sin(0) = 0 = a_1$, so a_1 is zero.

$$\cos\theta = 1 + 0\theta + a_2\theta^2 + a_3\theta^3 + a_4\theta^4 + \cdots$$

Moving on to a_2, we must differentiate one more time to get the term with a_2 to have no θs in it:

$$(\cos\theta)'' = 2a_2 + 3 \cdot 2a_3\theta + 4 \cdot 3a_4\theta^2 + \cdots$$

Since $(\cos\theta)' = -\sin\theta$ and $(-\sin\theta)' = -\cos\theta$, this series is equal to $-\cos\theta$! And evaluating at 0 gives -1. Thus $2a_2 = -1$ so $a_2 = -\frac{1}{2}$.

$$\cos\theta = 1 + 0\theta - \frac{1}{2}\theta^2 + a_3\theta^3 + a_4\theta^4 + \cdots$$

Continuing to a_3 differentiating the left side once more gives $\sin\theta$ which evaluates to zero, and the right results in a function with constant term $3 \cdot 2a_3$: thus a_3 is zero.

$$\cos\theta = 1 + 0\theta - \frac{1}{2}\theta^2 + 0\theta^3 + \cdots$$

Differentiating once more, the left side has returned to $\cos\theta$ and the right now has constant term $4 \cdot 3 \cdot 2a_4$: thus $a_4 = 1/4!$.

$$\cos\theta = 1 + 0\theta - \frac{1}{2}\theta^2 + 0\theta^3 + \frac{1}{4!}\theta^4 + \cdots$$

After repeating the process four times, we've cycled back around to the same function $\cos\theta$-that we started with! And so continuing, the same pattern in derivatives, $1, 0, -1, 0 \cdots$ will continue to repeat. This tells us every odd term will be zero in our series, and the even terms will have alternating signs:

$$\cos\theta = 1 - \frac{1}{2!}\theta^2 + \frac{1}{4!}\theta^4 - \frac{1}{6!}\theta^6 + \frac{1}{8!}\theta^8 - \cdots$$

Thus *if $\cos\theta$ can be written as a series at all - it must be this one! We only have left to confirm that this series actually converges (and thus, by Taylor's theorem, equals the cosine). □

Exercise 14.4. Prove that the series for cos converge for all real inputs: that is, that their radius of convergence is ∞. *Hint: review the ratio test!*

Exercise 14.5 (Series Expansion of Sin)**.** Run an analogous argument to the above to show

$$\sin\theta = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!}\theta^{2n+1}$$

14.3. The Dot Product

Having explicitly computable formulas for arccos and arcsin (even though they are via integral expressions, and probably need to be evaluated numerically as Riemann sums) lets us act as though knowing the cosine or sine of an angle is just as good as knowing the angle itself.

Example 14.5. What is the angle between $(1, 0)$ and $(1, 2)$?

We figure out where the second vector intersects the unit circle by dividing by its magnitude: $(1/\sqrt{5}, 2/\sqrt{5})$. Now we know *by definition* that whatever the arclength θ is, its cosine is the first coordinate. And being able to compute arccosines, we immediately get

$$\cos\theta = \frac{1}{\sqrt{5}} \implies \theta = 1.10714\text{rad}$$

Our goal in this section is to generalize the example above into a universal tool, that lets us compute the measure of any angle in the Euclidean plane using a simple tool from linear algebra: the dot product.

Definition 14.8. The dot product of $u = \langle u_1, u_2\rangle$ and $v = \langle v_1, v_2\rangle$ is

$$u \cdot v = u_1 v_1 + u_2 v_2$$

We can already directly compute the angle a unit vector $v = \langle v_1, v_2\rangle$ makes with $\langle 1, 0\rangle$: in this situation $v_1 = \cos\theta$ by definition, or $\langle 1, 0\rangle \cdot v = \cos\theta$. But if u and v are two unit vectors based at 0, how can we analogously compute the angle they form?

Theorem 14.1 (Dot Product Measures Arclength)**.** *If u, v are unit vectors based at $p \in \mathbb{E}^2$ making angle θ, then*

$$u \cdot v = \cos\theta$$

Proof. Let u, v denote an angle based at p.The idea is to use isometries to reduce this to the case we already understand! First, let ϕ be a translation isometry taking p to 0. Since the derivative of a translation is the identity, this takes u and v to the origin without changing their coordinates. And, since isometries preserve angle measures, we know the angle θ between u_p, v_p is the same as the angle between u_o, v_o.

Now let R be a rotation about O taking u_o to $\langle 1, 0\rangle_o$. Since angles are invariant under isometry, the angle measure between u_o and v_o is the same as between Ru_o and Rv_o. But since $Ru_o = \langle 1, 0\rangle_o$, we know

$$\cos\theta = \text{first component of } Rv$$

Thus, all we need to do is compute the matrix R, apply it to v, and read off the first component of the resulting vector! We know from Theorem 11.4 how to create a rotation fixing O that takes $\langle 1, 0\rangle_o$ to u_o is $\left(\begin{smallmatrix} u_1 & -u_2 \\ u_2 & u_1 \end{smallmatrix}\right)$. The transformation R we need is the *inverse* of this:

$$R = \begin{pmatrix} u_1 & -u_2 \\ u_2 & u_1 \end{pmatrix}^{-1} = \begin{pmatrix} u_1 & u_2 \\ -u_2 & u_1 \end{pmatrix}$$

Now we apply this to u and v to get our new vectors: applying to u gives $\langle 1, 0\rangle$ (check this!) and applying to v gives

$$Rv = \begin{pmatrix} u_1 & u_2 \\ -u_2 & u_1 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} u_1v_1 + u_2v_2 \\ -u_2v_1 + u_1v_2 \end{pmatrix}$$

The first component here is exactly $u \cdot v = u_1v_1 + u_2v_2$. Thus we are done! □

The above applies explicitly to unit vectors, as we used the rotation constructed in Theorem 11.4 to requires a unit vector to send $\langle 1, 0\rangle$ to. However, this is easily modified to measure the angle between non-unit vectors: just divide by their magnitudes first!

Corollary 14.4. *The measure θ of the angle between any two vectors u, v based at a point $p \in \mathbb{E}^2$ is related to the dot product via*

$$u \cdot v = \|u\|\|v\| \cos\theta$$

Proof. Let u, v be any two vectors based at p. Then $u/\|u\|$ and $v/\|v\|$ are two unit vectors based at p, and the angle between a pair of vectors is independent of their lengths (as its defined as an arclength along the *unit circle* no matter what). Using Theorem 14.1 we find the angle between these unit vectors:

$$\frac{u}{\|u\|} \cdot \frac{v}{\|v\|} = \cos\theta$$

Multiplying across by the product of the magnitudes gives the claimed result. □

Exercise 14.6. Prove that rectangles exist, using all of our new tools! (Ie write down what you know to be a rectangle, explain why each side is a line segment, parameterize it to find the tangent vectors at the vertices, and use the dot product to confirm that they are all right angles).

14.3.1. Trigonometric Identities

Using very similar reasoning to the above proposition relating angles to dot products, we can leverage our knowledge of rotations to efficiently discover trigonometric identities! We consider here the *angle sum identities* for sine and cosine.

Theorem 14.2 (Angle Sum Identites). *Let α, β be lengths of arc (equivalently, measures of angles) and $\alpha + \beta$ the angle formed by concatenating the two lengths. Then*

$$\sin(\alpha + \beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta$$

$$\cos(\alpha + \beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$$

Proof. Let $v = \langle v_1, v_2\rangle$ be a vector that makes angle α with $\langle 0, 1\rangle$, and $u = \langle u_1, u_2\rangle$ be a vector that makes angle β with $\langle 1, 0\rangle$.

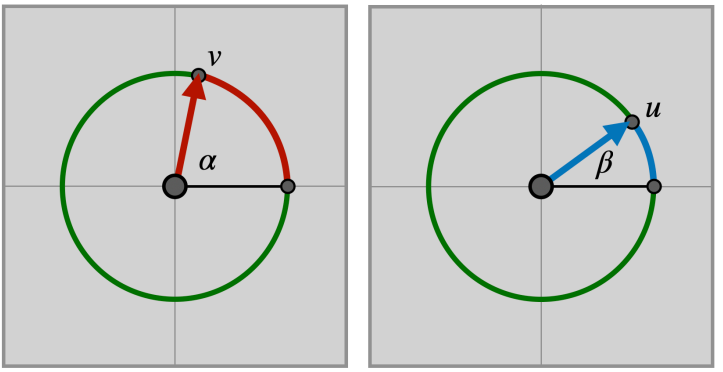


Figure 14.12.: Vectors v and u making angles α and β respectively.

Now let R be an isometry that rotates $\langle 1, 0\rangle$ to u. Since isometries preserve length, this takes the segment of the unit circle between $\langle 1, 0\rangle$ and v to a segment of the same length between u and Rv. So, now the total length of arc from $\langle 1, 0\rangle$ to Rv is $\alpha + \beta$

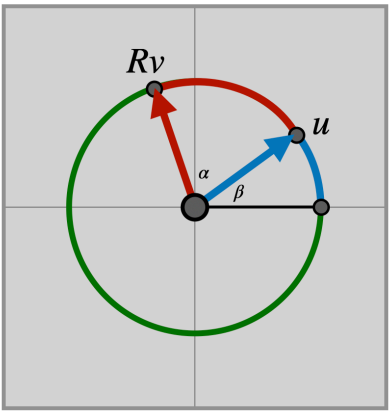


Figure 14.13.: A rotation R taking $\langle 1, 0\rangle$ to u takes v to a vector describing the angle sum $\alpha + \beta$.

Thus, the x and y coordinates of Rv are the cosine and sine of $\alpha + \beta$ respectively. Writing down the rotation R (via Theorem 11.4) we see

$$Rv = \begin{pmatrix} u_1 & -u_2 \\ u_2 & u_1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} u_1 v_1 - u_2 v_2 \\ u_1 v_1 + u_1 v_2 \end{pmatrix} = \begin{pmatrix} \cos(\alpha + \beta) \\ \sin(\alpha + \beta) \end{pmatrix}$$

Finally - since v makes an angle of α wtih $\langle 1, 0\rangle$ and u makes an angle of β, we know *by definition* that their coordinates are $v = (\cos\alpha, \sin\alpha)$ and $u = (\cos\beta, \sin\beta)$. Substituting these in gives the identities we seek. □

Analogously, we have the *angle difference identities*, which differ only in the choice of ± signs.

Theorem 14.3 (Angle Difference Identities)**.**

$$\sin(\alpha - \beta) = \sin\alpha\cos\beta - \cos\alpha\sin\beta$$

$$\cos(\alpha - \beta) = \cos\alpha\cos\beta + \sin\alpha\sin\beta$$

Exercise 14.7. Prove Theorem 14.3 similarly to how we proved Theorem 14.2 (you may need an inverse matrix!).

From these we can deduce the double angle formulas by setting both α and β equal to the same angle θ in the addition formula

Corollary 14.5 (Double Angle Identities)**.**

$$\cos(2\theta) = \cos^2\theta - \sin^2\theta \qquad \sin(2\theta) = 2\sin\theta\cos\theta$$

And the half angle formulas by algebraic manipulation of the above:

Corollary 14.6 (Half Angle Identities)**.**

$$\cos\left(\frac{\theta}{2}\right) = \sqrt{\frac{1+\cos\theta}{2}} \qquad \sin\left(\frac{\theta}{2}\right) = \sqrt{\frac{1-\cos\theta}{2}}$$

Proof. We prove the cosine identity here, and leave the other as an exercise. Starting from the double angle identity for cosine, and the fact that $\sin^2(x) + \cos^2(x) = 1$, we can do the following algebra:

$$\begin{aligned}\cos(2x) &= \cos^2(x) - \sin^2(x) \\ &= \cos^2(x) - (1 - \cos^2(x)) \\ &= 2\cos^2(x) - 1\end{aligned}$$

Now, we just solve for $\cos(x)$ in terms of $\cos(2x)$:

$$2\cos^2(x) = 1 + \cos(2x) \implies \cos(x) = \sqrt{\frac{1+\cos(2x)}{2}}$$

This lets us compute the cosine of an angle in terms of twice that angle! Replace $2x$ with θ to get the form above. □

Exercise 14.8. Prove the half-angle identity for $\sin x$.

These formulas are actually quite useful in practice, to find exact values of the trigonometric functions at different angles, given only the few angles we have computed explicitly ($\tau/4$, for a square, and $\tau/6$ from an equilateral triangle).

Example 14.6 (The exact value of $\sin(\tau/24)$)**.** There are several ways we could approach this: one is to start with $\tau/6$ and bisect twice. Another is to notice that

$$\frac{\tau}{24} - \frac{\tau}{6} - \frac{\tau}{8}$$

and use the angle subtraction identity. We will do the latter here, and below in the discussion of Archimedes cover the repeated bisection approach.

Theorem 14.3 tells us that

$$\sin\frac{\tau}{24} = \sin\frac{\tau}{6}\cos\frac{\tau}{8} - \cos\frac{\tau}{6}\sin\frac{\tau}{8}$$

We've successfully reduced the problem to knowing the sine and cosine of the larger angles $\tau/6$ and $\tau/8$. These are both do-able by hand: for $\tau/8$ we could either note

that this is half a right angle so lies along the line $y = x$, and solve for the point (x, x) on the unit circle getting

$$\sin\frac{\tau}{8} = \sqrt{22} \qquad \cos\frac{\tau}{8} = \frac{\sqrt{2}}{2}$$

Or, we could have started with the right angle directly and applied bisection. For $\tau/6$, we may use Exercise 14.1 to see

$$\sin\frac{\tau}{6} = \frac{\sqrt{3}}{2} \qquad \cos\frac{\tau}{6} = \frac{1}{2}$$

The rest is just algebra:

$$\begin{aligned}\sin\frac{\tau}{24} &= \frac{\sqrt{3}}{2}\frac{\sqrt{2}}{2} - \frac{1}{2}\frac{\sqrt{2}}{2} \\ &= \frac{\sqrt{2}}{4}\left(\sqrt{3}-1\right) \\ &\approx 0.258819\end{aligned}$$

14.3.1.1. The Measurement of the Circle

The half angle identities played a crucial role in Archimedes' ability to compute the perimeter of n-gons in his paper *The Measurement of the Circle*. Indeed, to calculate the circumference of an inscribed n-gon, its enough to be able to find $\sin\tau/(2n)$:

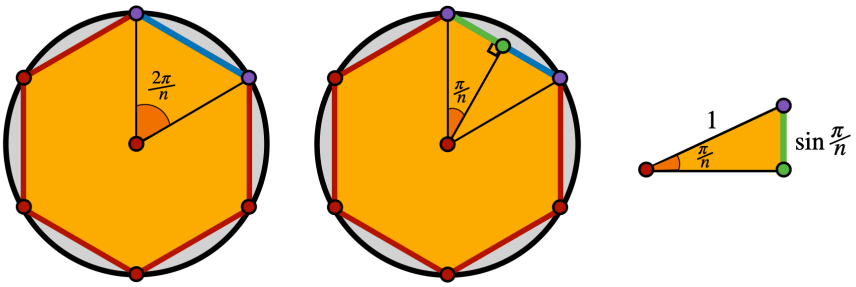


Figure 14.14.: The side-length of an inscribed n-gon is $2\sin\frac{\tau}{2n}$, found via bisecting the side to form a right triangle. The perimeter of the n-gon is just n times this.

By repeatedly bisecting the sides, we can start with something we can directly compute - like a triangle, and repeatedly bisect to compute larger and larger n-gons.

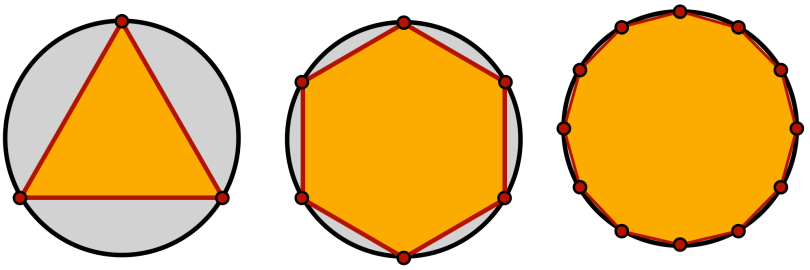


Figure 14.15.: Archimedes' method: repeatedly doubling the number of sides of the n-gon to get polygons approaching the circle.

Example 14.7 (From Triangle to Hexagon to 12-Gon)**.** Start by inscribing an equilateral triangle in the circle. The angle formed by each side at the center is $\tau/3$, and so bisecting a side gives an angle of $\tau/6$ - the same as the angle of the equilateral triangle itself! We know the sine and cosine of this angle from Exercise 14.1:

$$\cos\frac{\tau}{6} = \frac{1}{2} \qquad \sin\frac{\tau}{6} = \frac{\sqrt{3}}{2}$$

Thus, the length of one side is $2 \cdot \frac{\sqrt{3}}{2} = \sqrt{3}$, and the circumference is $3\sqrt{3} \approx 5.1961524$.

Doubling the side number to get to the hexagon requires we compute $\sin\frac{\tau}{12}$, which we do via-half angle:

$$\sin\frac{\tau}{12} = \sqrt{\frac{1-\cos\frac{\tau}{6}}{2}} = \sqrt{\frac{1}{4}} = \frac{1}{2}$$

Thus, the side length here is 1 and the circumference is six times that, or 6. Doubling once more we now need to compute $\sin\frac{\tau}{24}$ via the half-angle identity:

$$\sin\frac{\tau}{24} = \sqrt{\frac{1-\cos\frac{\tau}{12}}{2}}$$

Unfortunately - we do not know $\cos\tau/12$ yet: but we can find it! Since $\cos^2(x) + \sin^2(x) = 1$ we may use the fact that we know $\sin\frac{\tau}{12} = \frac{1}{2}$ to calculate it:

$$\cos\frac{\tau}{12} = \sqrt{1-\left(\frac{1}{2}\right)^2} = \frac{\sqrt{3}}{2}$$

Plugging this back in, we get what we are after:

$$\sin\frac{\tau}{24} = \sqrt{\frac{1-\frac{\sqrt{3}}{2}}{2}} = \frac{\sqrt{2-\sqrt{3}}}{2} \approx 0.258819$$

Thus the length of one side of the 12-gon is $\sqrt{2-\sqrt{3}}$, and its total perimeter is $12\sqrt{2-\sqrt{3}} \approx 6.21165$

(Note: using a different set of identities we get a different looking expression for our final answer here: a square root of a square root! But - its exactly the same value. Can you do some algebra to prove it?)

Exercise 14.9. Continue to bisections until you can compute $\sin(\tau/(2\cdot 96))$. What is the perimeter of the regular 96-gon (use a computer to get a decimal approximation, after your exact answer).

Explain how we know that this is *provably an underestimate* of the true length, using the definition of line segments.

Be brave - and go beyond Archimedes! Compute the circumference of the 192-gon.

Exercise 14.10. In the 400s CE, Chinese mathematician Zu Chongzi continued this process until he reached the $24,576$-gon, and found (in our modern notation) that $3.1415926 < \pi < 3.1415927$. How many times did he bisect the original equilateral triangle?

Exercise 14.11. Can you use trigonometry to find the perimeter of circumscribed n-gons as well? This would give you an *upper bound* to τ, to complement the lower bound found from inscribed ones.

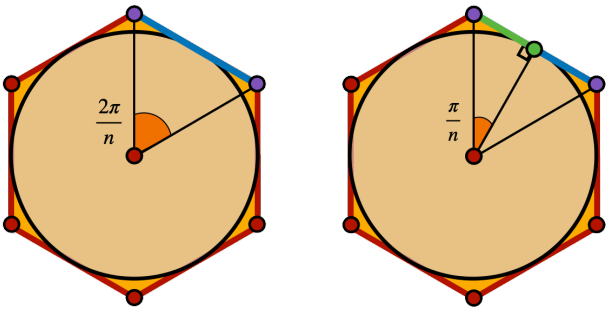


Figure 14.16.: Circumscribed n gons are the smallest n-gons containing the unit circle.

14.4. Euclid's Axioms 4 & 5

The final two of Euclid's postulates mention angles. Now that we have constructed them within our new foundations, we can finally attempt to prove these two!

The fourth postulate states *all right angles are equal.* Of course, by *equal* Euclid meant *congruent* as he often did. In order to be precise, it helps to spell everything out a bit better.

Proposition 14.5 (Euclids' Postulate 4)**.** *Given the following two configurations: - A point p, and two orthogonal unit vectors u_p, v_p based at p - A point q, and two orthogonal unit vectors a_q, b_q based at q*

There is an isometry ϕ of $\mathbb{E}^2$ which takes p to q, takes u_p to a_q, and v_p to b_q.

Exercise 14.12. Prove Euclid's forth postulate holds in the geometry we have built founded on calculus.

Hint: there's a couple natural approaches here.

- *You could directly use Exercise 32.28 to move one point to the other and line up one of the tangent vectors. Then deal with the second one: can you prove its either already lined up, or will be after one reflection?*
- *Alternatively, you could show that every right angle can be moved to the "standard right angle" formed by $\langle 1, 0\rangle, \langle 0, 1\rangle$ at O. Then use this to move every angle to every other, transiting through O*

At long last - we are down to the final postulate of Euclid - the Parallel Postulate, in its original formulation, also mentions angles and so could not be formulated in our new geometry until now.

Proposition 14.6 (The Parallel Postulate)**.** *Given two lines L_1 and L_2 crossed by another line Λ, if the sum of the angles that the L_i make with Λ on one side are less than $\tau/2$, then the L_i intersect on that side.*

Of course, we do not *need* to prove this to finish our quest: we have already proven the equivalent postulate of Playfair/Proculus. But, bot for completeness and the satisfaction of directly grounding the Elements in our new formalism, I cannot help but offer it as an exercise.

Exercise 14.13. Prove the parallel postulate.
Hint: try the special case where the crossing line Λ makes a right angle with one of the others (say L_1). Use isometries to move their intersection to O, the crossing line Λ to the y-axis, and L_1 to the $x-$axis. Now you just need to prove L_2 is parallel to the x-axis if and only if it intersects the y axis in a right angle.

14.5. Conformal Maps

We've already seen that isometries preserve the angles between any two tangent vectors in the plane. But these are not the only maps with this property. In general, an angle preserving map is called *conformal*

Definition 14.9. A map $F : \mathbb{E}^2 \to \mathbb{E}^2$ is *conformal* if it preserves all infinitesimal angles in the plane. That is, if u, v are two tangent vectors at p

$$\text{Angle}(u, v) = \text{Angle}\left(DF_p(u), DF_p(v)\right)$$

Remark 14.3. Recall that by default we read angles counterclockwise: this is important in the definition of conformality. For example, PICTURE is not conformal as it sends an angle of θ to an angle of $\tau - \theta$. (Alternatively, reading clockwise we may say *negative* θ. Maps that preserve angles after reversing their sign are called *anti-conformal*)

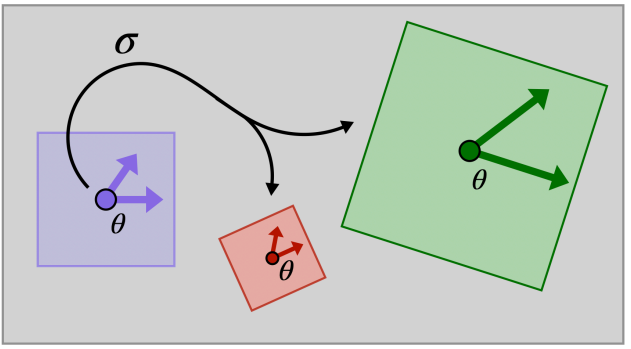


Figure 14.17.: A conformal map preserves all angles, though it may distort lengths.

Because we have a simple relationship between angles and the dot product, we can formulate this in an easy-to-compute way.

Corollary 14.7. *A map $F : \mathbb{E}^2 \to \mathbb{E}^2$ is conformal if for every pair of vectors u, v based at p we have*

$$\cos\theta = \frac{u \cdot v}{\|u\|\|v\|} = \frac{DF_p(u) \cdot DF_p(v)}{\|DF_p(u)\|\|DF_p(v)\|}$$

We won't have much immediate need for this material on conformal maps - as we are primarily concerned with Euclidean isometries at the moment, which we already

know to preserve angles! But, when we study maps of spherical geometry and especially hyperbolic geometry, being able to tell when a map is conformal will be of great use - so we provide some material here to reference in the future.

Example 14.8 (Complex Squaring is Conformal)**.** The complex squaring operation $z \mapsto z^2$ can be written as a s real function on the plane in terms of x, y as

$$S(x, y) = (x^2 - y^2, 2xy)$$

This function is conformal everywhere except at O, which we verify by direct calculation.

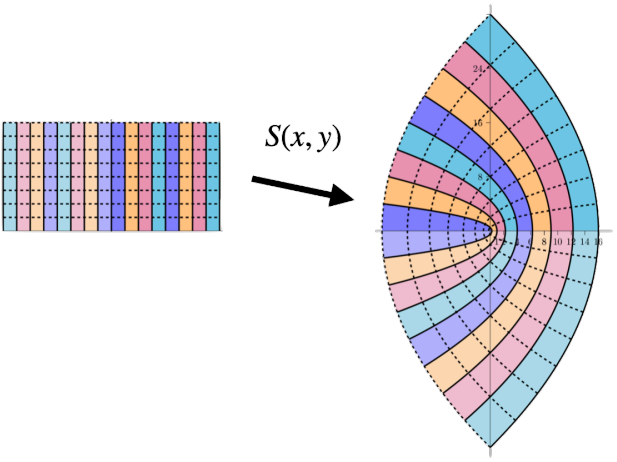


Figure 14.18.: The complex squaring function is conformal: it sends all the right angles of the grid right angles.

The derivative matrix at $p = (x, y)$ is

$$DF_p = \begin{pmatrix} 2x & -2y \\ 2y & 2x \end{pmatrix}$$

So, now we just need to take two vectors $u = \langle u_1, u_2$ and $v = \langle v_1, v_2 \rangle$ based at p, apply the derivative, and see what the resulting angle is!

$$DF_p(u) = \begin{pmatrix} 2xu_1 + 2yu_2 \\ -2yu_1 + 2xu_2 \end{pmatrix} \qquad DF_p(v) = \begin{pmatrix} 2xv_1 + 2yv_2 \\ -2yv_1 + 2xv_2 \end{pmatrix}$$

After a lot of algebra, we can find the length of these two vectors

$$\|DF_p(u)\| = \sqrt{4(x^2 + y^2)(u_1^2 + u_2^2)} \qquad \|DF_p(v)\| = \sqrt{4(x^2 + y^2)(v_1^2 + v_2^2)}$$

And we can also find their dot product:

$$DF_p(u) \cdot DF_p(v) = 4(x^2 + y^2)(u_1 v_1 + u_2 v_2)$$

Thus, forming the quotient that measures the cosine of the angle between them, we can cancel a factor of $4(x^2 + y^2)$ from both the top and bottom!

$$\begin{aligned}\frac{DF_p(u) \cdot DF_p(v)}{\|DF_p(u)\|\|DF_p(v)\|} &= \frac{4(x^2 + y^2)(u_1 v_1 + u_2 v_2)}{\sqrt{4(x^2 + y^2)(u_1^2 + u_2^2)}\sqrt{4(x^2 + y^2)(v_1^2 + v_2^2)}} \\ &= \frac{u_1 v_1 + u_2 v_2}{\sqrt{u_1^2 + u_2^2}\sqrt{v_1^2 + v_2^2}} \\ &= \frac{u \cdot v}{\|u\|\|v\|}\end{aligned}$$

But this still isn't the easiest condition to check, as we have to test it for *all pairs of vectors u, v* at every point! Luckily, we can use the linearity of the dot product to help us come up with an easier means of checking for conformality.

Theorem 14.4 (Testing for Conformality)**.** *A map $F : \mathbb{E}^2 \to \mathbb{E}^2$ is conformal if it satisfies the following two conditions:*

- *It sends $\langle 1, 0\rangle_p$ and $\langle 0, 1\rangle_p$ to a pair of orthogonal vectors, at each point.*
- *These vectors $DF_p(\langle 1, 0\rangle)$ and $DF_p(\langle 0, 1\rangle)$ have the same nonzero length.*

Proof.

□

Example 14.9 (Complex Squaring is Conformal)**.** We can re-check that the squaring map $S(x, y) = (x^2 - y^2, 2xy)$ is conformal using the theorem above: since the derivative at $p = (x, y)$ is

$$DF_p = \begin{pmatrix} 2x & -2y \\ 2y & 2x \end{pmatrix}$$

we simply apply this to the standard basis vectors and see

$$DF_p(\langle 1, 0\rangle) = \big(2x, 2y\big) \qquad DF_p(\langle 0, 1\rangle) = \big(-2y, 2x\big)$$

These two vectors are orthogonal as their dot product is zero. And, they are both the same length: $2\sqrt{x^2 + y^2}$. This length is nonzero unless $(x, y) = O$, so S is conformal everywhere except O.

But wait! We can do even better than this: say that ϕ sends $\langle 1, 0\rangle_p$ to the vector $\langle a, b\rangle_{\phi(p)}$. Then we know (via Theorem 14.4) that $\langle 0, 1\rangle$ must be sent to the $\tau/4$ rotation of this! So, $D\phi_p(\langle 0, 1\rangle) = \langle -b, a\rangle_{\phi(p)}$. But if we know where $D\phi$ sends both of the standard basis vectors, we know its matrix!

Corollary 14.8. *The map $\phi : \mathbb{E}^2 \to \mathbb{E}^2$ is conformal if and only if its derivative matrix has the form*

$$D\phi_p = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

for some a, b at each point p of the plane.

Example 14.10 (Complex Squaring is Conformal)**.** We can re-re-check that the squaring map $S(x, y) = (x^2 - y^2, 2xy)$ is conformal using the theorem above: just taking the derivative

$$DF_p = \begin{pmatrix} 2x & -2y \\ 2y & 2x \end{pmatrix}$$

we see the diagonal terms are equal and the off diagonals are negatives of one another. Thus, its conformal by Corollary 14.8.

Exercise 14.14 (Complex Exponentiation is Conformal)**.** The complex exponential e^z can be written as a real function on the plane in terms of x, y as

$$E(x, y) = (e^x \cos y, e^x \sin y)$$

Prove that E is a conformal map.

Exercise 14.15. Prove that if a map F is conformal and preserves the length of *at least one vector at each point* (say, it sends $\langle 1, 0\rangle_p$ to a unit vector at $F(p)$), then F is an isometry.

15. Area

Now that we know how to measure angle and orthogonality, we can make sense of infinitesimal areas.

Definition 15.1 (Infinitesimal Area in $\mathbb{E}^2$)**.** An infinitesimal area at a point p is a region in the tangent space $T_p\mathbb{E}^2$.

Because the tangent space is a *linear space*, we will primarily be interested in infinitesimal areas described by polygons: the most common of which will be parallelograms as they are defined by two vectors. This suggests a natural means of *measuring* infinitesimal area: just as we took the Pythagorean theorem as the definition of infinitesimal length, we may take the area of a parallelogram (Exercise 15.5) as the definition of infinitesimal area.

Remark 15.1. It may seem like we we are bringing in a *new concept to our geometry* here - something that can't be defined in terms of our starting point which only allowed the measurement of infinitesimal lengths. But - as we will show in the final section of this chapter, this is not the case. We can derive this formula for dA from a set of requirements mentioning only lengths and angles (angles of course, are also defined in terms of lengths).

Definition 15.2 (Measuring Infinitesimal area: dA)**.** The function dA is an infinitesimal area measure on $T_p\mathbb{E}^2$, which takes in two vectors u, v and returns the area of the parallelogram spanned by them:

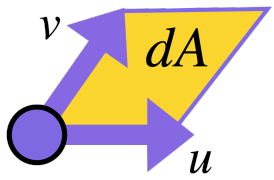


Figure 15.1.: Area in the tangent space.

The most common (and useful!) parallelograms that we will encounter are rectangles, due to our use of x, y coordinates on the plane. Here infinitesimal area is quite simple: if the length is dx and the height is dy, we have an infinitesimal rectangle with area the product of base and height:

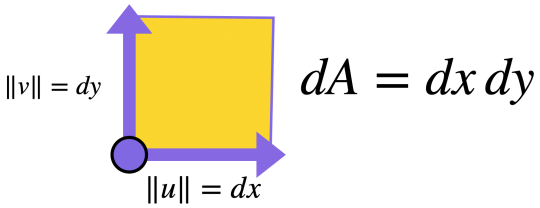


Figure 15.2.: The area of rectangles in the tangent space to a point

Just like lengths, once we have a means of measuring the infinitesimal notion, we can zoom back out to recover what we are really after via integration.

Definition 15.3 (Area in $\mathbb{E}^2$)**.** If R is a region in $\mathbb{E}^2$, its area is given by the following integral expression

$$\text{Area}(R) = \iint_R dA = \iint_R dxdy$$

15.1. Iterated Integrals

How do we compute an integral over a 2-dimensional region, with respect to the infinitesimal area $dxdy$? Looking at a finite approximation gives one answer - we could sum along rows first, adding up all the little areas with the same y coordinates at once. Then we could add up all the total area of each row. In the limit, this tells us to *integrate x first*, and then *to integrate the result with respect to y.

$$\iint_R dA = \iint_R dxdy = \int_{\text{y-slices}} \left(\int_{x-slices} dx \right) dy$$

Conversely, we could have instead sliced our approximation into *columns* (integrating dy first), and then added up the area of the columns (integrating the results dx). This would give

$$\iint_R dA = \iint_R dydx = \int_{\text{x-slices}} \left(\int_{y-slices} dy \right) dx$$

Thus, thanks to the fact that dA factors as a product, an area integral really is just two one dimensional integrals performed in succession! And to evaluate explicit areas, all one needs to do is find a way to measure the length of the x-slices or y-slices of a region.

In practice, this will be our main means of calculating area. It becomes especially tractable when the region R can be described in terms of single-variable functions, where everything reduces to a 1-dimensional integral!

Theorem 15.1 (Area Between Two Curves). *Let f, g be functions with $g(x) < f(x)$ on $[a, b]$. Define R as the region*

$$R = \{(x, y) \mid x \in [a, b],\ y \in [g(x), f(x)]\}$$

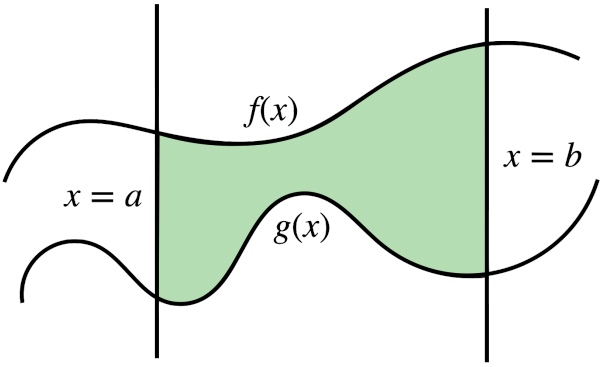


Figure 15.3.: The region between $f(x)$ and $g(x)$.

Then its area can be computed via

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

Proof. Then at each fixed x, the vertical slice through the region is the interval $[g(x), f(x)]$, and so the area integral can be written as an iterated integral: first over $[g(x), f(x)]$ for a fixed x, then over $x \in [a, b]$

$$\text{Area}(R) = \iint_R dA = \int_{[a,b]} \int_{[g(x),f(x)]} dy\,dx$$

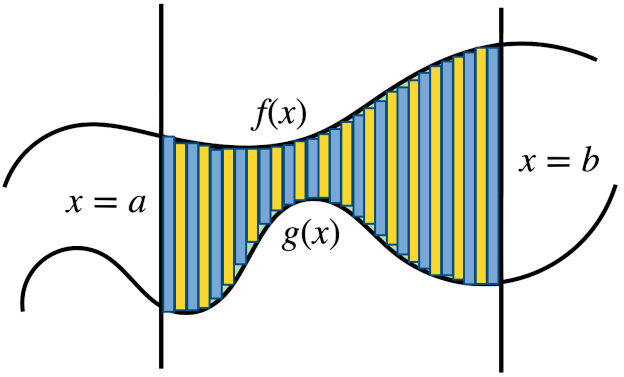


Figure 15.4.: The area integral interpreted as an iterated integral, with $\int dy$ done first.

The inner integral here is straightforward to evaluate: there are no y's at all - so by the fundamental theorem we have

$$\int_{[g(x),f(x)]} dy = y\Big|_{g(x)}^{f(x)} = f(x) - g(x)$$

Substituting back in gives the result:

$$\text{Area}(R) = \int_{[a,b]} f(x) - g(x)\, dx$$

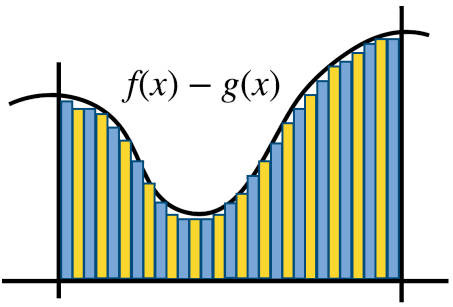


Figure 15.5.: The area between two curves is just the length of all the slices, added up (integrated).

□

This is how we can define rigorously the *area of a circle*: we know (for example) that the unit circle has equation $x^2 + y^1 = 1$, and so its top half can be written $y = \sqrt{1 - x^2}$ and the bottom half by $y = -\sqrt{1 - x^2}$. Thus the area is

$$\int_{-1}^{1} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} dy dx = \int_{-1}^{1} 2\sqrt{1 - x^2} dx$$

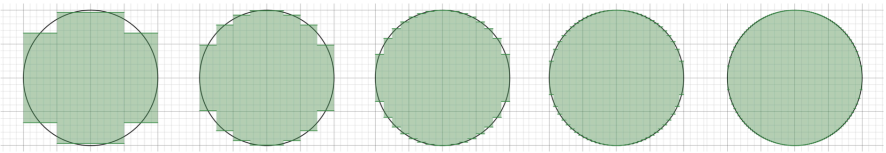


Figure 15.6.: Riemann sums for the integral defining the area of a circle with 4,8,16,32, and 64 bars, respectively.

Corollary 15.1 (Defining π)**.**

$$\pi = \int_{-1}^{1} 2\sqrt{1 - x^2} dx$$

Anytime we can describe a region with functions, we are back to calculus. Sometimes this is impossible for an entire region all at once, but we can break it into smaller regions, each of which are described by functions.

Example 15.1 (Area Between Piecewise Curves)**.** Compute the area between the curve $y = \frac{1}{2}x^2$ and the piecewise curve below, for $x \in [0, 2]$.

$$f(x) = \begin{cases} x^2 & x < 1 \\ x & x \geq 1 \end{cases}$$

Drawing the region, we decide to divide it into two regions R_1 and R_2, with R_1 the portion with $x \in [0, 1]$ and R_2 the portion with $x \in [1, 2]$.

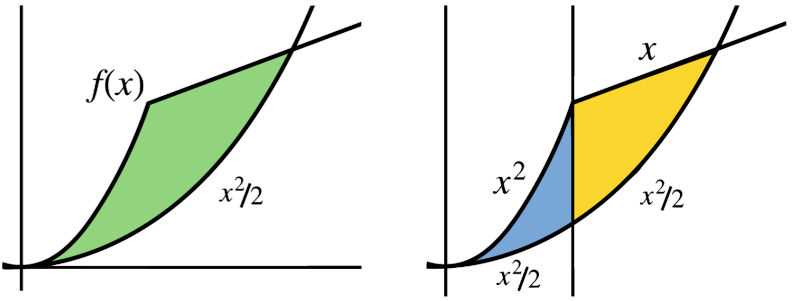


Figure 15.7.: The region R, and its division into two simpler regions R_1 and R_2.

In each of these regions we can specify the boundaries as functions of x, allowing us to express them via single variable integrals (Theorem 15.1)

$$\text{Area}(R_1) = \int_0^1 x^2 - \frac{1}{2}x^2\,dx = \int_0^1 \frac{1}{2}x^2\,dx = \frac{1}{2}\frac{1}{3} = \frac{1}{6}$$

$$\text{Area}(R_2) = \int_1^2 x - \frac{1}{2}x^2\,dx = \frac{1}{2} - \frac{1}{6} = \frac{1}{3}$$

The total area is the sum of these,

$$\text{Area}(R) = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}$$

We can use this definition of area to compute the area of a triangle in the plane, since we now know how to describe straight lines as affine curves.

Exercise 15.1 (Area of Right Triangle With Calculus)**.** Use calculus to find the area between the x-axis, the y-axis, and the linear equation with y-intercept $(0, h)$ and x-intercept $(b, 0)$.

Exercise 15.2 (Area of a General triangle)**.** Set up an area integral to measure the area of a triangle with vertices O, $(L, 0)$, and (p, q) (assume L, p and q are positive numbers: it will be a piecewise area between curves).

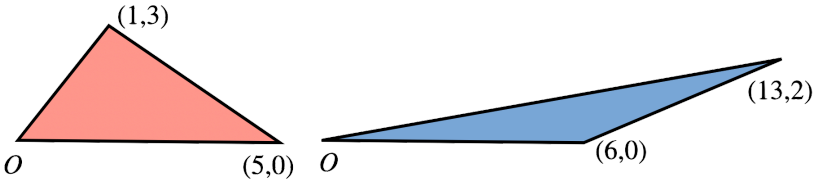


Figure 15.8.: Triangles for various p, q, L.

Show the result gives you half the base times the height.

Similarly, we can make quick work of some impressive results of archimedes, after checking that $y = x^2$ actually describes a parabola (in Exercise 13.8)

Exercise 15.3 (Quadrature of the Parabola with Calculus)**.**

- Write down a formula for the area of the triangle whose third vertex lies at (x, x^2)
- Use calculus to find the point x where the inscribed triangle has maximal area. Then show that Archimedes was right: the slope of the tangent line to the parabola at this point is exactly the same as the slope of the line segment forming the triangle's base!
- Finally, compute the area of the parabolic segment (via integration, as the area between two curves). Show that its exactly 4/3rds the area of the triangle!

(*Hint: instead of finding the height of the triangle to use* $\frac{1}{2}bh$*, can you use the fact that the determinant of a matrix calculates the area of a parallelogram whose sides are the column vectors, and that the area of a the triangle you want is half a parallelogram?*)

15.2. Isometries & Similarities

Now that we know *how* to evaluate an area integral, its time to study some of its properties. Our first question with every new concept we define should be *how does this concept interact with isometries?* So we investigate this below.

Theorem 15.2 (Isometries Preserve Area)**.** *Let R be a region in the plane, and ϕ an isometry. Then*

$$\text{Area}(\phi(R)) = \text{Area}(R)$$

Proof. Let R be a region in the plane, and at each point $p \in R$ consider the unit orthogonal vectors $e_1 = \langle 1, 0\rangle_p$ and $e_2 = \langle 0, 1\rangle_p$ defining the unit area square used in the computation of dA. Since isometries preserve infinitesimal lengths and angles, ϕ takes this to another infinitesimal unit square based at $\phi(p)$, also of unit area.

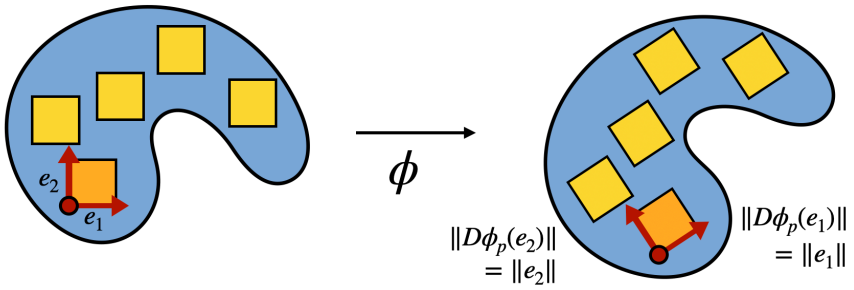


Figure 15.9.: Angles and infinitesimal lengths are preserved, so infinitesimal squares are sent to squares of the same area.

Thus, the integral $\iint_{\phi(R)} dA$ is adding up the exact same areas as $\iint_R dA$, and they are equal. □

Theorem 15.3 (Similarities Scale Area)**.** *Let R be a region in the plane, and σ an similarity with scaling factor k. Then*

$$\text{Area}(\sigma(R)) = k^2 \text{Area}(R)$$

Proof. Running a similar argument to the above, we see that the infinitesimal unit square defined by $e_1 = \langle 1, 0 \rangle_p$ and $e_2 = \langle 0, 1 \rangle_p$ at each point is taken to a square with side lengths k (since similarities uniformly scale infinitesimal lengths, but still preserve angles).

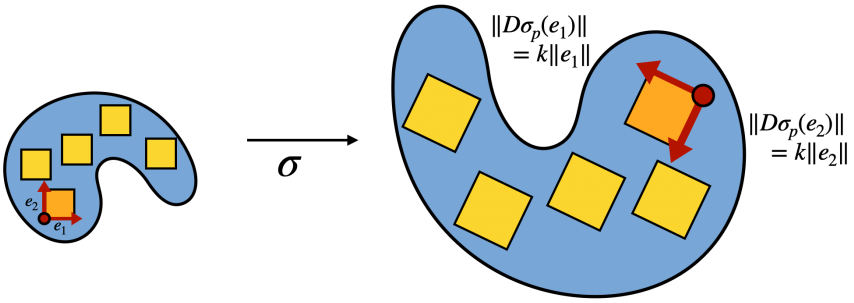


Figure 15.10.: Angles are preserved but infinitesimal lengths are scaled by k. Thus infinitesimal areas are scaled by k^2.

The area of such a square is k^2, so the integral defining $\text{Area}(\phi(R))$ counts an area of k^2 every time the integral defining $\text{Area}(R)$ counts a unit area. Thus, the total area is k^2 times the original. □

15.3. Area and General Mappings

Both isometries and similarities are rather special: they send every infinitesimal unit square to another square, possibly scaled in size by a constant factor.

In this section we are interested in discovering what happens to an area under a general map $\mathbb{E}^2 \to \mathbb{E}^2$. First, let's consider a conformal map ϕ. This map takes infinitesimal squares to squares, but they no longer all need to be the same size.

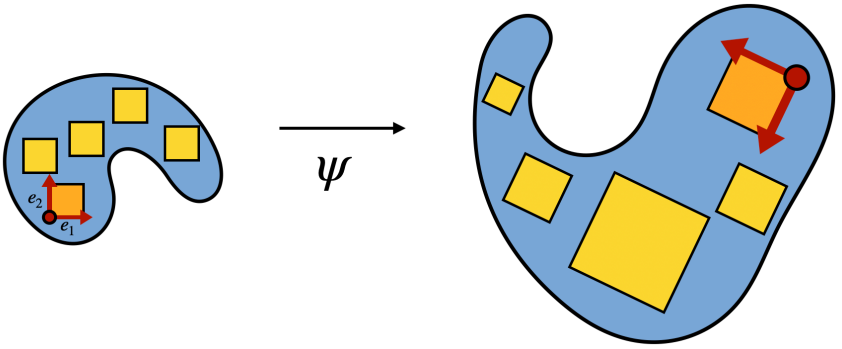


Figure 15.11.: Conformal maps take infinitesimal squares to squares, but the size of the square can differ across the region.

Indeed, by Corollary 14.8 we know that at each point $p \in \mathbb{E}^2$ the sides of such an infinitesimal square are $\langle a, b\rangle$ and $\langle -b, a\rangle$ - each of length $\sqrt{a^2+b^2}$ so the total infinitesimal area is scaled up from 1 by $a(x,y)^2 + b(x,y)^2$. (Here we've written a and b as functions of x, y to emphasize that they may take different values at different points of the plane).

Thus the area of the region $\phi(R)$ can be computed starting from R, but multiplying each infinitesimal area by this factor:

$$\text{Area}(\phi(R)) = \int_R (a(x,y)^2 + b(x,y)^2)\, dxdy$$

Example 15.2 (Area under the map $z \mapsto z^2$)**.** The *squaring map* from complex analysis can be written as a function of real coordinates x, y, as

$$S(x,y) = (x^2 - y^2, 2xy)$$

This map takes the unit square $R = \{(x,y) \mid x \in [0,1],\ y \in [0,1]\}$ to the region $S(R)$ depicted below.

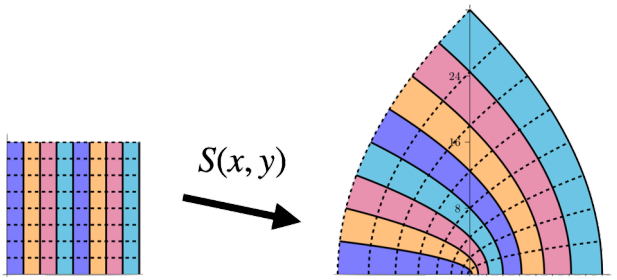


Figure 15.12.: The image of the unit square under the complex squaring map.

Using all that we've learned, we can actually compute the area of this region without having to even describe it explicitly! We know that at each point (x, y), the derivative map is

$$DF_p = \begin{pmatrix} 2x & -2y \\ 2y & 2x \end{pmatrix}$$

Thus the change in area for the infinitesimal square based at (x, y) is $4(x^2 + y^2)$. This allows us to compute the area as

$$\text{Area}(S(R)) = \iint_R 4(x^2 + y^2)dxdy$$

Which we can now just do as an iterated integral:

$$\begin{aligned}
\iint_R 4(x^2 + y^2)dxdy &= \int_0^1 \left(\int_0^1 4(x^2 + y^2)\, dx \right) dy \\
&= \int_0^1 4\left(\frac{x^3}{3} + xy^2 \right) \Big|_0^1 dy \\
&= \int_0^1 \frac{4}{3} + 4y^2\, dy \\
&= \left(\frac{4}{3}y + \frac{4}{3}y^3 \right) \Big|_0^1 \\
&= \frac{8}{3}
\end{aligned}$$

Finally, let's consider a general map $F : \mathbb{E}^2 \to \mathbb{E}^2$ of the plane. We know F does not need to preserve infinitesimal lengths or angle, and so takes takes squares in the tangent space to rectangles or parallelograms.

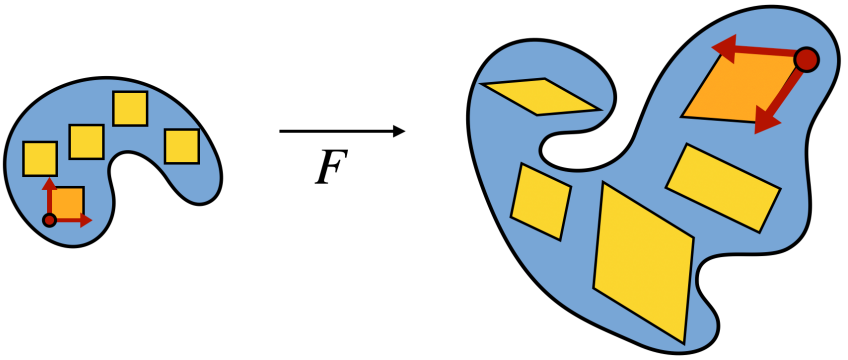


Figure 15.13.: A general mapping need not preserve angles or lengths, and so will take the original infinitesimal squares defining $dA = dxdy$ to parallelograms of various sizes and shapes.

But we can figure out from this what F does to infinitesimal areas: it takes the unit area spanned by $\langle 1, 0\rangle$ and $\langle 0, 1\rangle$ to the parallelogram spanned by $DF_p(\langle 1, 0\rangle)$ and $DF_p(\langle 0, 1\rangle)$!

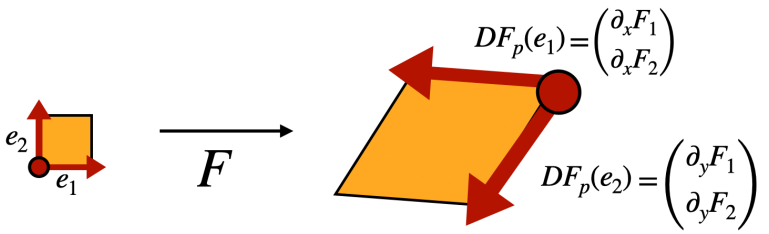


Figure 15.14.: The area of the parallelogram determined by these two vectors is just the determinant of DF!

And we know how to calculate the area of a parallelogram using the vectors determining its sides (Exercise 32.21) - this is just the determinant of the derivative matrix.

Theorem 15.4. *If $F : \mathbb{E}^2 \to \mathbb{E}^2$ is any differentiable mapping, F takes the unit infinitesimal area in $T_p\mathbb{E}^2$ to the area*

$$|\det DF_p| = \begin{vmatrix} \partial_x F_1 & \partial_y F_1 \\ \partial_x F_2 & \partial_y F_2 \end{vmatrix} = \partial_x F_1 \partial_y F_2 - \partial_y F_1 \partial_x F_2$$

This quantity is called the Jacobian of F *at* p.

Much like we have done for isometries, similarities, and conformal maps before; this lets us compute the area of a region $F(R)$ as an integral directly over the starting region R itself! At each point $p \in R$ we just insert the area scaling factor for how much F changes the area of an infinitesimal square based there: the Jacobian.

Theorem 15.5. *Let $R \subset \mathbb{E}^2$ be a region in the plane, and $F : \mathbb{E}^2 \to \mathbb{E}^2$ some mapping that takes R to a new region, $F(R)$. Then*

$$\text{Area}(F(R)) = \int_{F(R)} dA = \int_R |\det DF| dxdy$$

We can use this to find areas that seem difficult at first: for example, we will be able to calculate the area of an ellipse in terms of the area of a circle (we'll find the circles' area in the next section).

Exercise 15.4. The map $F(x, y) = (ax, by)$ takes points on the unit circle to the points of the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$. (Confirm this with algebra!) Thus, it takes the unit disk $D = \{(x, y) \mid x^2 + y^2 \leq 1\}$ to the interior of this ellipse: call this region E.

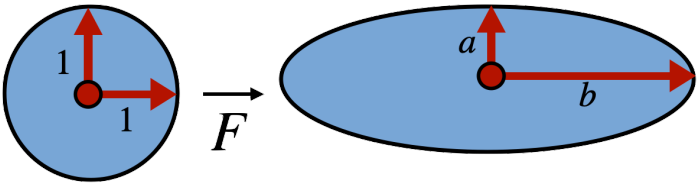


Figure 15.15.: Stretching a circle into an ellipse.

Computing the Jacobian we see F scales areas by a factor of ab:

$$DF = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \implies |DF| = \det \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} = ab$$

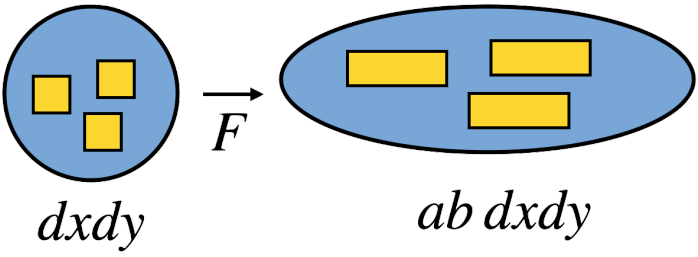


Figure 15.16.: This stretches all infinitesimal areas by a factor of $|DF| = ab$

Thus we can calculate the area of the ellipse E as

$$\begin{aligned}\text{Area}(E) &= \iint_E dA \\ &= \iint_{F(D)} dA \\ &= \iint_D |DF|\, dxdy \\ &= \iint ab\ dxdy \\ &= ab \iint_D dxdy \\ &= ab\,\text{Area}(D)\end{aligned}$$

We will see in shortly that $\text{Area}(D) = \pi$, so that immediately gives $\text{Area}(E) = \pi ab$.

One thing to be careful about: while all isometries preserve area, not all area-preserving maps are isometries! Take any determinant 1 matrix on the plane and use it as a linear map. This preserves area of all subsets (as derivative is itself, and so determinant of the derivative is 1). But does not preserve lengths: try a hyperbolic like

$$\begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

15.4. The Jacobian, Abstractly

This optional section gives a second means of deriving the jacobian, instead of taking the fact that we already understand the area of a parallelogram. We could instead ask, what sort of behavior do we want the function dA to have; and try to *derive* its formula from such a list.

Instead of being explicit about what number dA assigns to every area, we can attempt to be more austere and just declare that dA **assigns *the unit square*** $\langle 1, 0\rangle, \langle 0, 1\rangle$ ***unit area***.

To get further than this, we need a proposal about how dA interacts with scalar multiplication. If v, w are two vectors and we multiply one of them by k, this should increase the area they span by k: that is,

$$dA(kv, w) = kdA(v, w)$$

For vector addition, we analogously propose that the area spanned by $u + v$ and w is the same as the area spanned by u, w and v, w

$$dA(u + v, w) = dA(u, w) + dA(v, w)$$

We've illustrated the case of addition and scalar multiplication in the first vector above, but of course it should not depend which vector we are talking about, so we propose that dA is ***linear in each of its input vectors***.

We have one final thing to consider: what is the relationship between $dA(v, w)$ and $dA(w, v)$? A natural thought is that these both describe the same area, an so should certainly be assigned the same number! But this is ignoring a useful piece of information: that switching between (v, w) and (w, v) negates the *sense* of angles, essentially *flipping* the parallelogram. To allow dA to record this information, we may impose that ***switching the input vectors negates the result***. (Multi)-linear functions with this property are called *alternating*

$$dA(v, w) = -dA(w, v)$$

In fact, these properties alone are enough to fully determine the function dA! And, evaluating it on an arbitrary pair of input vectors, we see the usual formula for the determinant is forced on us.

Exercise 15.5. Using only the following facts about the function $dA(v, w)$, derive the standard formula for the determinant

$$dA\left(\binom{a}{c}, \binom{b}{d}\right) = ad - bc$$

- dA evaluates to 1 on the square $\langle 1, 0\rangle, \langle 0, 1\rangle$.
- dA is alternating: $dA(v, w) = -dA(w, v)$.
- dA is linear in both the first and second argument.

16. π

One of the great mysteries of mathematics is the ubiquity of certain particular numbers. It scarcely matters what field of mathematics you are working in, if you think hard enough and dig deep enoughs you'll inevitably run into the mysterious number

$$2.718281828459\ldots$$

apearing in your calculations. (This will even happen to us, in this class, not too far from now!) This number appears in everything from finance to probability, to differential equations, group theory, real analysis, and non-euclidean geometry. But this is not even the craziest of the numerical conspiracies: the true king of almost magically ubiquitous numbers is

$$3.141592653589\ldots$$

We have already met this number in this class, but our brief encounter (as the area of the unit circle) does not provide much evidence or intution for *why* this should appear everywhere in mathematics! That is the goal of this chapter: we will see that a collection of rather remarkable properties of Euclidean space make it so that many conceptually-different mathematical quantities are all (1) constant and (2) have values directly related to π. Its easiest to explain all of this via examples, so instead of further discussion let's just dive right in.

16.1. Circle Constants

16.1.1. The Length Constant

For any circle C in the plane we can define its *length factor* to be the ratio of its circumference to its radius: that is, how many times do we need to lay out the radius to equal the circumference? If we write $\text{circ}(C)$ to denote the circumference, or arclength of the circle C and $\text{rad}(C)$ to denote the radius, the quantity we are interested in here is just their ratio:

$$\tau(C) = \frac{\text{circ}(C)}{\text{rad}(C)}$$

If we write $C_{p,r}$ for the circle centered at $p \in \mathbb{E}^2$ of radius r, we alternatively could express this length ratio as

$$\tau(C_{p,r}) = \frac{\text{length}(C_{p,r})}{r}$$

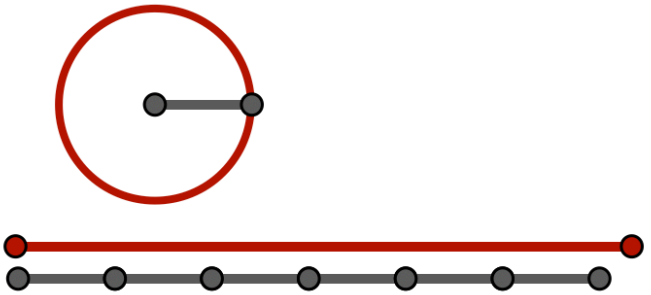


Figure 16.1.: The length factor of a circle is the number of radii needed to make the circumference.

We use the letter τ for this ratio as its value for the unit circle is τ as we have defined it previously:

$$\tau := \tau(C_{O,1}) = \text{length}(C_{O,1})$$

Both the numerator and denominator of this definition *depend on the circle C being considered* - so there's no a priori reason to assume that this ratio should be *independent of the choice of circle.* Indeed - we will see very shortly in both spherical and hyperbolic geometry the analog of $\tau(C)$ takes different values for different circles!

But, as an incredible consequence of the existence of *isometries* and *similarities* of the Euclidean plane, it turns out that here this number is a constant!

Theorem 16.1 (Length Factor is Constant)**.** *The ratio of a circles circumference to its radius is a constant, independent of the circle.*

Proof. Let $C_{p,r}$ be any circle in the plane - centered at some point p and of some radius r. Now let T be the translation which takes p to the origin O. Isometries preserve distances, and thus send circles to circles. This means $T(C_{p,r})$ is a circle of radius r (a distance) centered at O: in symbols

$$T(C_{p,r}) = C_{O,r}$$

Isometries also do not change the lengths of curves (Theorem 11.2), so we know that $\text{length}(C_{p,r}) = \text{length}(C_{O,r})$. And since it doesnt change distances (like the radius: Proposition 12.3) we see that $C_{p,r}$ and $C_{O,r}$ have the same length ratios:

$$\tau(C_{p,r}) = \frac{\text{length}(C_{p,r})}{r} = \frac{\text{length}(C_{O,r})}{r} = \tau(C_{O,r})$$

Now we will show that $C_{O,r}$ has the same length factor as the unit circle, and thus our original circle had the same length factor as the unit circle! To do so, we use the similarity $\sigma(x, y) = (Rx, Ry)$. This has scaling factor r, and so scales all lengths of curves (Proposition 11.3), and all distances (Exercise 12.4) by r. Thus, σ takes the unit circle $C_{O,1}$ to the circle $C_{O,r}$ and also takes $\text{length}(C_{O,1})$ to $r\,\text{length}(C_{O,1})$. Because both the circumference of the circle and the radius got scaled by r the length factor is unchanged:

$$\tau(C_{O,1}) = \frac{\text{length}(C_{O,1})}{1} = \frac{r\,\text{length}(C_{0,1})}{r} = \frac{\text{length}(C_{O,r})}{r} = \tau(C_{O,r})$$

Stringing all the equalities together, we see

$$\tau(C_{p,r}) = \tau(C_{O,r}) = \tau(C_{O,1})$$

Thus every circle has the same length factor as the unit circle, so the length factor is constant. □

Definition 16.1 (The Circle Length Constant)**.** The constant ratio of the circumference to the radius of a circle to its radius is

$$\tau = \frac{\text{length}(C_{p,r})}{r}$$

In Exercise 32.41 we found a good approximation to this following the method of archimedes:

$$\tau \cong 96 \cdot 2 \cdot \sqrt{2 - \sqrt{2 + \sqrt{2 + \sqrt{2 + \sqrt{2 + \sqrt{3}}}}}} \approx 6.282904\ldots$$

Thus in any mathematical problem involving a circle's length, the number $\tau(C) = \tau = 6.28\ldots$ is bound to show up: this is just the circumference measured in units of radii!

16.1.2. The Area Constant

We've found that similarites force the length factor of circles to be a constant - and this explains at least some occurences of a geometric constant appearing in mathematics. But lengths aren't the only important quantity related to circles out there! It's equally natural to consider their area.

Here it doesn't make sense to measure a circles area in units of radii, since radii are a length and area is....not a length. Instead its more natural to measure area in units of *radii squared*: how many squares with side length the radius does it take to fill up a circle? For a given circle C, we will call this *area factor* $\pi(C)$:

$$\pi(C_{p,r}) = \frac{\text{Area}(C_{p,r})}{r^2}$$

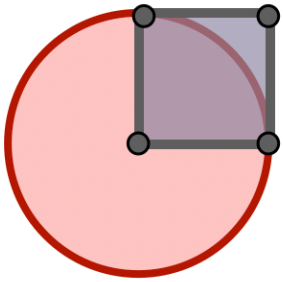


Figure 16.2.: The area factor of a circle is the number of squared radii needed to completely fill it's area.

We use the letter π for this as we have already defined π to be this ratio for the *unit circle* $C_{O,1}$:

$$\pi := \pi(C_{O,1}) = \text{area}(C_{O,1})$$

Again, this fraction involves quantities related to the particular circle $C_{p,r}$ in both the numerator and denominator, so its totally conceivable that its value would depend on the particular circle being considered! (And, in spherical and hyperbolic geometry, it will).

But perhaps after seeing the crucial role of similarities in the argument for the constancy of τ, perhaps you already have a sneaking suspicion that the analogous trick will prove π to be constant here.

Theorem 16.2 (Area Factor is Constant)**.** *The ratio of a circle's area to its radius squared is constant, independent of the circle*

Proof. The proof here is nearly identical to the length factor case, except we need to use the fact that similarites scale *area* by the square of their similarity factor (Theorem 15.3), instead of the fact that they linearly scale length.

Again since isometries don't change distances or areas we can move an arbitrary circle $C_{p,r}$ to the origin, $C_{O,r}$ and know that

$$\pi(C_{p,r}) = \pi(C_{O,r})$$

Next we see that $\pi(C_{O,r})$ is the same as the area factor of the unit circle, using the similarity $\sigma(x, y) = (rx, ry)$, which sends $C_{O,1}$ to $C_{O,r}$ and $\text{area}(C_{O,1})$ to $r^2\text{area}(C_{O,1})$.

$$\pi(C_{O,1}) = \frac{\text{area}(C_{O,1})}{1} = \frac{r^2\text{area}(C_{0,1})}{r^2} = \frac{\text{area}(C_{O,r})}{r^2} = \pi(C_{O,r})$$

Stringing these together we see that every circle's area ratio is the same!

$$\pi(C_{p,r}) = \pi(C_{O,r}) = \pi(C_{O,1})$$

□

Definition 16.2. For any circle C in the Euclidean plane, the ratio of its area to squared radius is a constant denoted by

$$\pi = \frac{\text{Area}(C_{p,r})}{r^2}$$

This tells us that we should expect yet another constant to be popping up throughout mathematics: anytime a discussion of circles and their areas show up, we will run into π as the natural *conversion factor* from radius squared to area!

Archimedes could have went on to estimate the value of π by calculating the *area* of an inscribed polygon or circumscribed polygon, by adding up the area of triangles.

Exercise 16.1 (π via inscribed areas.)**.** The area of a triangle is half its base times it's height. Can you calculate the area of a polygon that *circumscribes* the circle to get an approximation of π? Try starting with a hexagon. Then, can you find a way to use trigonometric identities to double the number of sides repeatedly, like we did for circumference?

However, Archimedes did not do this…he did something much more clever.

16.1.3. Equality

Alone these two facts don't point towards a *single, unified constnat* showing up in mathematics, rather they suggest we should be seeing two different values, π and τ, in two different circumstances: area and length respectively.

It was already known to Euclid that these two constants exist: in CITE PROP Euclid shows the circumference of a circle is always proportional to its radius, and in CITE PROP he shows the area is always proportional to the radius squared. But it wasn't until the work of Archimedes that we discovered the truly astounding fact that these two constants are *related to one another*! Recall the main result of the *measurement of the circle*

Theorem 16.3. *The area of a circle is equal to that of a triangle whose base is the circle's circumference, and whose height is the circles radius.*

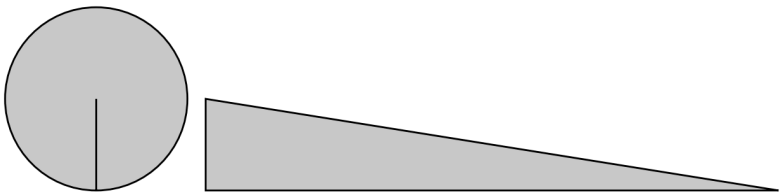


Figure 16.3.: Archimedes' measurement of the circle

Because we know the area of a triangle to be half its base times its height, this tells us that

$$\pi r^2 = \frac{1}{2}(\tau r)(r)$$

Or, cancelling the factors of r and re-arranging,

$$\tau = 2\pi$$

That is, the two circle constants are just *integer multiples of one another*! This means whether we are interested in lengths or areas, so long as we are doing mathematics that invovles a circle this constant is going to appear. (This also explains why you see so many formulas with a 2π in them: this is really the length constant τ! But since they are rationally related we've just chosen one of them, π to write everything in terms of.)

As we have been doing throughout this section of the book, a good exercise is to *prove archimedes observation* using modern techiques. We will give two approaches here, one based on integration, and another on differentiation.

16.1.3.1. Integration

Both lengths and areas in our modern version of geometry are calculated via integrals, so it's no surprise that the values of π and τ themselves are integrals. Indeed, as we saw in Corollary 14.2 we can write the circumference of the unit circle as

$$\tau = \int_{-1}^{1} \frac{2}{\sqrt{1-t^2}} dt$$

And, from the area chapter (Corollary 15.1) we saw we can express its area as

$$\pi = \int_{-1}^{1} 2\sqrt{1-x^2} dx$$

In this homework excercise we will use familiar calculus techniques (just u-substitution!) to relate these integrals to one another (without evaluating either!) giving a modern proof of Archimedes' theorem.

Exercise 16.2. Prove that

$$\int_{-1}^{1} \frac{1}{\sqrt{1-t^2}} dt = \int_{-1}^{1} 2\sqrt{1-x^2} dx$$

Thus showing thtat $\frac{\tau}{2} = \pi$.

Hint: Do u-substitutions to the integrals to make them into the same integral. The goal isn't to evaluate them and get a number! This is just a Calc II problem - but a tricky one, so here's one outline you could follow:

(1) Rewrite the area integrand $\sqrt{1-x^2}$ as $\frac{1-x^2}{\sqrt{1-x^2}}$. Use properties of integrals to break this into two integrals, and see

$$\pi = \tau - \int_{-1}^{1} \frac{2x^2}{\sqrt{1-x^2}} dx$$

(2) Now we just have to evaluate this new integral: Do the u-substitution $u = \sqrt{1-x^2}$ to this, to show that

$$\int_{-1}^{1} \frac{2x^2}{\sqrt{1-x^2}} dx = \int_{-1}^{1} 2\sqrt{1-u^2} du = \pi$$

(This u-sub requires some work: you'll need at some point to solve for x in terms of u!)

(3) Now just assemble the pieces! You never completed a single integral, but you still managed to prove that $\tau = 2\pi$.

16.1.3.2. Differentiation

To give a third proof of this fundamental equality, we will start wtih the formula defining the area of a circle of radius r:

$$\text{area}(C_{p,r}) = \pi r^2$$

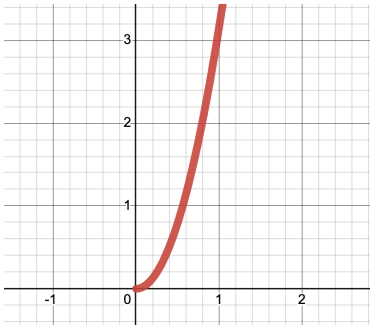


Figure 16.4.: Area of the circle as a function of radius.

Let's think a bit about the derivative of this function: this is easy to compute by hand

$$\frac{d}{dr}\text{area}(C_{p,r}) = \frac{d}{dr}\pi r^2 = 2\pi r$$

but what does it mean? For this, we need to return all the way to the fundamentals, and think about the *definition of the derivative.* To unclutter the notation, below I am going to write area(r) for the area of a circle of radius r (the area doesn't depend on the center point after all!)

$$\frac{d}{dr}\text{area}(r) = \frac{\text{area}(r+h) - \text{area}(r)}{h}$$

The numerator here is a *difference of areas* - between the area of a disk with radius $r+h$ and a disk of radius r. This is what you get if you *remove a disk* of radius r from a disk of radius $r+h$, so this is the area of a thin circular ring.

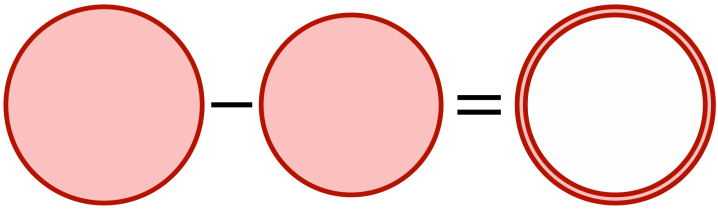


Figure 16.5.: The difference between a disk of radius $r + h$ and a disk of radius r is a ring of thickness h.

What's the area of this ring? In the limit as h becomes infinitesimally small (as we take the limit to become the actual derivative) we can calculate infinitesimally: imagine the circular ring is made of a bunch of tiny squares, whose height is h: since their other sides fit together to form the circumference, the sum of their bases is τr. Thus,

$$\text{area}(r + h) - \text{area}(r) \approx (\tau r)h$$

With this approximation becoming exact as $h \to 0$. But in the derivatrive we *divide by h*, and are left with just τr!

$$\frac{d}{dr}\text{area}(r) = \tau r = \text{circ}(r)$$

This is an incredibly cool fact: so we should box it off as a theorem for future reference!

Theorem 16.4. *The derivative of the* area function *for circles of radius r is the* circumference function

$$\frac{d}{dr}\text{area}(r) = \text{circ}(r)$$

But now we are all but complete with our third way of proving $\tau = 2\pi$. We know the area function is $\text{area}(r) = \pi r^2$, and so we can take its derivative to get $2\pi r$. Similarly, we know the circumference formula is τr, so this relathionship simplifies to

$$2\pi r = \tau r$$

And, canceling the r (or evaluating at the unit circle, $r = 1$) gives the result!

16.1.4. Trigonometric Substitution

The intimate relationship between τ and π is so fundamental that I cannot help but offer yet *another proof* of this result. This may seem wasteful but is in fact often a useful thing to do in mathematics - as different proofs generalize to different situations easier.

Here, we will focus directly on the area integral of Corollary 15.1,

$$\pi = \int_{-1}^{1} 2\sqrt{1-x^2}dx$$

and try to directly evaluate it via the fundamental theorem of calculus (finding an antiderivative, and plugging in the endpoints). To do a little pre-emptive simplifcation, we may notice that the integrand is an *even function* of x so we may instead choose to integrate on half the domain, say $[0,1]$, and double the result:

$$\pi = 4\int_{0}^{1} \sqrt{1-x^2}dx$$

Now, we perform a rather clever substitution to the integral. Because we rigorously studied the trigonometric functions we recall that $\cos^2\theta + \sin^2\theta = 1$, and thus if $x = \sin\theta$ we could simplify $1-x^2$ as

$$1 - x^2 = 1 - \sin^2\theta = \cos^2\theta$$

Thus, $\sqrt{1-x^2}$ simply becomes $|\cos\theta|$. And, because we computed the derivative of $\sin\theta$ and $\cos\theta$ in the chapter on angles, we know that

$$dx = d(\sin\theta) = \cos\theta d\theta$$

The last portion of the integral we need to convert are the bounds. The lower bound of $x = 0$ means we seek θ with $x = \sin\theta = 0$. From our defintion of sin and cos, we see this happens at $\theta = 0$ since sin is the y coordinate, and $\theta = 0$ corresponds to the starting point $(1,0)$. Next, for the top bound $x = 1$ we seek the θ value with $x = \sin\theta = 1$. This occurs along the positive y axis, so a quarter turn around the circle, or $\theta = \tau/4$. Putting all these pieces together, we see

$$\int_{0}^{1} \sqrt{1-x^2}dx = \int_{0}^{\frac{\tau}{4}} |\cos\theta|\cos\theta d\theta = \int_{0}^{\frac{\tau}{4}} \cos^2\theta\, d\theta$$

It appears we aren't doing much better: we didn't know the antiderivative of $\sqrt{1-x^2}$ which is what set all of this off, but we also don't know the antiderivative of $\cos^2\theta$!

However, the reason this type of substitution is powerful is that there arent many *square root identities* out there we can use to change how a function is represented, but there are plenty of *trigonometric identities*.

Indeed, from the angle sum identity we derived,

$$\cos(a+b) = \cos a \cos b - \sin a \sin b$$

by setting $a = b = \theta$ and using the pythagorean identity $\cos^2\theta + \sin^2\theta = 1$, one can show that

$$\cos^2\theta = \frac{1+\cos 2\theta}{2}$$

Exercise 16.3. Derive this identity.

This lets us rewrite our integral

$$\int_0^{\frac{\tau}{4}} \cos^2\theta\, d\theta = \int_0^{\frac{\tau}{4}} \frac{1+\cos 2\theta}{2} d\theta = \frac{1}{2}\int_0^{\frac{\tau}{4}} d\theta + \frac{1}{2}\int_0^{\frac{\tau}{4}} \cos 2\theta\, d\theta$$

The first of these integrals is straightforward: its $\tau/4$. For the second integral, we can u-sub $u = 2\theta$ to get

$$\int_0^{\frac{\tau}{4}} \cos 2\theta d\theta = \frac{1}{2}\int_0^{\frac{\tau}{2}} \cos u\, du$$

But now - *finally* - we know the antiderivative! Since the derivative of sine is cosine, we can compute

$$\int_0^{\frac{\tau}{2}} \cos u du = \sin u\Big|_0^{\frac{\tau}{2}} = \sin\left(\frac{\tau}{2}\right) - \sin(0) = 0 - 0 = 0$$

All that work for zero!! But, putting it all together, we see

$$\int_0^1 \sqrt{1-x^2} dx = \int_0^{\frac{\tau}{4}} \cos^2\theta d\theta = \frac{1}{2}\frac{\tau}{4}$$

And going back to the very beginning we recall that π was exactly four times this integral. Thus

$$\pi = 4\frac{1}{2}\frac{\tau}{4} = \frac{\tau}{2}$$

Our fourth independent derivation that $\tau = 2\pi$.

16.2. Sphere Constants

To talk about things in spheres and cylinders rigorously, we neeed to study a bit of 3-dimensional Euclidean geometry. We will return to this in more detail later in the course, but here we only give a slight taste, as it is important to our overall discussion of π.

Just as a circle was the set of points a fixed distance (the radius) from a fixed point (the center), a **sphere is defined just as a circle** except now in three dimenions. We found the equations for distance minimizing curves in 2D to be affine, and that let us find the distance formula $\text{dist}((x,y),(h,k)) = \sqrt{(x-h)^2+(y-k)^2}$ and consequently the fomrula for a circle $(x-h)^2+(y-k)^2 = r^2$. All of this carries through with no changes in three dimensions, where the distance formula becomes

$$\text{dist}\left((x,y,z),(h,k,\ell)\right) = \sqrt{(x-h)^2+(y-k)^2+(z-\ell)^2}$$

And consequently, the sphere centered at (h,k,ℓ) of radius r has the formula

$$(x-h)^2+(y-k)^2+(z-\ell)^2 = r^2$$

The *surface area* of the sphere is defined exactly as we have done in the plane, by dividing its surface into infinitesimal parallelograms dA, and then integrating the area of these parallelograms to get the final answer. The volume is defined analogously, except we now need a notion of infinitesimal volume in 3-dimensions. Volume of an infinitesimal 3d rectangle is given by length times width times height, or in symbols

$$dV = dxdydz$$

and so three dimensional volumes are calculated by *three iterated integrals* instead of the *double iterated integrals* for area.

16.2.1. Fundamental Constants

Let $S_{p,r}$ denote the sphere of radius r centered at p, just as we did for $C_{p,r}$ for the circle. Like in two dimensions, we can define an *area ratio* and a *volume ratio* for the sphere, comparing each quantity to the relevant power of r.

Theorem 16.5 (Surface Area Ratio is Constant)**.** *The ratio*

$$\frac{\text{area}(S_{p,r})}{r^2}$$

is constant, and independent of the sphere considered.

Proof. The proof strategy here is exactly analogous to what we did for the length and area constants of a circle: we prove that every sphere has the same surface area ratio by using isometries and similarities to relate it to the unit sphere. First, we translate $S_{p,r}$ to the origin, which does not change lengths or areas. Then, we use a similarity to scale the unit sphere to the sphere of radius r. This scales the surface area by r^2 (as similarities scale areas by r^2) and it scales length by r. Thus

$$\frac{\text{area}(S_{O,1})}{1} = \frac{r^2\text{area}(S_{O,1}}{r^2} = \frac{\text{area}(S_{0,r})}{r^2} = \frac{\text{area}(S_{p,r})}{r^2}$$

□

Theorem 16.6 (Volume Ratio is Constant)**.** *The ratio*

$$\frac{\text{vol}(S_{p,r})}{r^3}$$

is constant, and independent of the sphere considered.

Proof. Run the same proof as above, but now notice that when we scale volume, infinitesimal volume is measured by $dxdydz$, so if each is scaled by r we get

$$rdxrdyrdz = r^3dxdydz$$

Thus volumes are scaled by r^3 under a similarity, and so

$$\frac{\text{vol}(S_{O,1})}{1} = \frac{r^3\text{vol}(S_{O,1}}{r^3} = \frac{\text{vol}(S_{0,r})}{r^3} = \frac{\text{vol}(S_{p,r})}{r^3}$$

□

Just as we gave names τ for the length constant and π for the area constant of circles, we may be tempted to give names to these two new fundamental constants that we just discovered. And, temporarily we will do so, but these names will not stick around for long - we will instead find both to be related to the circle constants! To keep things concise during their breif existence we will name the surface area constant C_{Surf} and the volume constant C_{Vol}.

16.2.2. Relationship to π.

The work Archimedes was most proud of we have barely discussed yet in this class. In his book *The Sphere and the Cylinder*, Archimedes managed to find a relationship between the formulas for the surface area and volume of a sphere, and relate them to those of a cylinder. This was of course a big deal because the volume of a shape with curved sides had never been caclulated before, but it was an even bigger deal the *form*

that the answer took. Specifically, Archimedes found that for surface area, the area of the sphere is exactly equal to the area of the round side of a the smallest cylinder that can enclose it (whose radius is the same as the spheres, and whose height is the sphere's diameter)

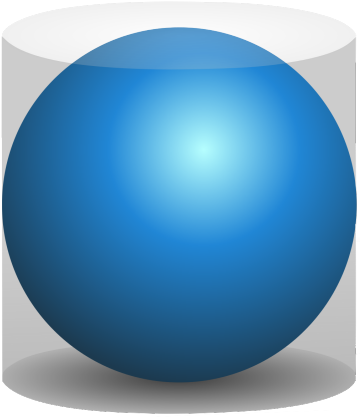


Figure 16.6.: The cylinder and the sphere have the same surface area.

In modern notation, we would write this relationship with a formula. A cylinder is a rolled up rectangle, and so we can calculate its area with base times height. The base is the circumference of the cylinder (so, τr since this is a circle!) and the height is $2r$. Thus

$$\text{area}(S_{p,r}) = (\tau r)(2r) = 2\tau r^2$$

Because we already proved τ is related to π via $\tau = 2\pi$ we often instead see this written as

$$\text{area}(S_{p,r}) = 4\pi r^2$$

This tells us immediately the value of the surface area constant, as its definition is just the surface area over r^2!

Theorem 16.7 (Value of the Surface Area Constant)**.**

$$C_{Surf} = \frac{\text{area}(S_{p,r})}{r^2} = 4\pi$$

We will prove this using modern tools below, but I'll postpone the proof until we talk about Archimedes *other* great discovery - calculating the volume of a sphere.

Through an ingenious argument by slicing, Archimedes showed that the *volume* of the sphere is the same as the volume of the following complicated sounding shape: the volume in between the cylinder enclosing the sphere (from above), and the double cone that fits inside it:

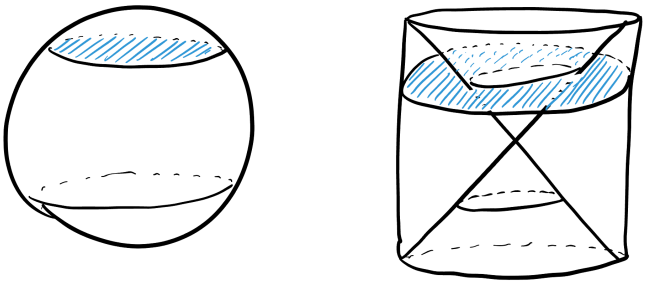


Figure 16.7.: Archimedes' calculation of the volume of a sphere by comparing its slices with the complement of a cone in a cylinder.

Archimedes original argument was by slicing: he imagined slicing each of these shapes by a plane at different heights, and he showed that at any given height z, the cross sections of the two shapes had the same area.

Exercise 16.4. Confirm Archimedes claim: show the slice of a sphere and cylinder-minus-cone at height z have the same areas, for any z.

Then Archimedes noted that the *volume* of a region is the integral of the area of its slices (of course, not using these words, as they were not to be invented for another 1800 years!) and so two shapes with all the same cross sectional areas must have the same volumes.

He next computed the volume of a cylinder to be the area of its base times its height, and the area of a cone to be 1/3 its base times height. This gave him the formula

$$\text{vol}(S_{p,r}) = (\pi r^2)(2r) - \frac{1}{3}(\pi r^2)(2r) = \frac{2}{3}(\pi r^2)(2r) = \frac{4}{3}\pi r^3$$

But this immediately gives us the value of the sphere volume constant in terms of π: thus all the constants for circles and spheres are just rational multiples of a single mysterious number!

Theorem 16.8 (Value of the Volume Constant)**.**

$$C_{Vol} = \frac{\text{vol}(S_{p,r})}{r^3} = \frac{4}{3}\pi$$

Archimedes found this fact so striking and so beautiful that he asked for a Sphere and a Cylinder (the key ideas in this proof) to be engraved on his tombstone. As far as we can tell in the historical record, his wish was heeded when he died in 212BCE- but his grave was quickly forgotten to those living on his native island of Syracruse.

However, in 75BCE, the great Roman orator Cicero was visiting Syracruse and searched out Archimedes - then already known as the greatest mathematical mind in history, and found it due to this carving. In his own words:

> "Once, while I was superintendent in Syracuse, I brought out from the dust Archimedes, a distinguished citizen of that city. In fact, I searched for his tomb, ignored by the Syracusans, surrounded on all sides and covered with brambles and weeds. The Syracusan denied absolutely that it existed, but I possessed the senari verses written on his tomb, according to which on top of the tomb of Archimedes a sphere with a cylinder had been placed. But I was examining everything with the eyes ... And shortly after I noticed a small hill not far emerged from the bushes. On it there was the figure of a sphere and a cylinder. And I said immediately to the Syracusans "That's what I wanted!" > Cicero, 75 BCE

16.2.3. Modern Computation

Having seen the beauty of the results which we are after, we will now seek to prove them with modern (calculus-based) methods. We find the volume of the sphere by slicing into disks, and we know the area of a disk from slicing it into line segments!

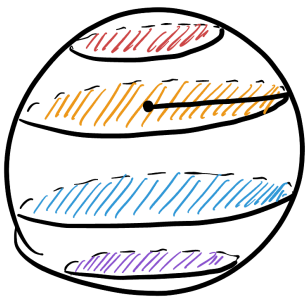


Figure 16.8.: Slicing a sphere allows us to calculate volume by integrating the area of the slices.

Because we know the number we are after is *constant* we are welcome to work directly with the *unit sphere* so that extra letters like radii don't complicate our lives. Call this sphere $S = S_{O,1}$. Then

$$\text{vol}(S) = \iiint_S dV = \int_{-1}^{1} \left(\iint_{C_r} dA \right) dz$$

Where C_r is the circle of radius r that we get by slicing horizontally at height z. Because we know the area of a circle formula is πr^2 we can subsitute this into our integral, and reduce it to a single integration!

$$\text{vol}(S) = \int_{-1}^{1} \pi r^2 \, dz$$

All that remains is to figure out the radius of the slice at height z. This is easiest to do by looking at a *side view* where we can use the distance formula in a plane:

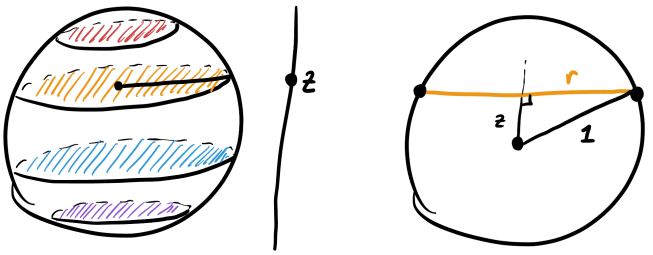


Figure 16.9.: The radius of a slice at height z satisfies $r^2 + z^2 = 1$ so $r = \sqrt{1 - z^2}$.

Alternatively we may just do this algebraically, and note that if $x^2 + y^2 + z^2 = 1$ then $x^2 + y^2 = 1 - z^2$, so at height z the points (x, y) lie in a circle whose radius-squared is $1 - z^2$, or

$$r(z) = \sqrt{1 - z^2}$$

Now plugging this into the area-of-a-disk formula, we can continue our integration by slicing:

$$\begin{aligned}\text{vol}(S) &= \int_{-1}^{1} \pi \left(\sqrt{1-z^2}\right)^2 \, dz \\ &= \pi \int_{-1}^{1} 1 - z^2 \, dz \\ &= \pi \left(z - \frac{z^3}{3}\right)\Big|_{-1}^{1} \\ &= \frac{4}{3}\pi\end{aligned}$$

Because a homothety multiplies each infinitesimal length by its scaling factor, it increases the infinitesimal volume by the *cube* of the scaling factor. Thus, scaling up from the unit sphere to a sphere of radius r scales this as

$$\text{vol}(S_{p,r}) = \frac{4}{3}\pi r^3$$

Now, we can apply everything we learned thinking about circles to give a quick modern derivation of the area constant: area is the derivative of volume!

PICTURE

The reasoning goes through exactly analogously here: the difference quotient $\text{vol}(r+h) - \text{vol}(r)$ is a thin spherical shell of thickness h, so its volume is approximately the surface area of the shere times h, and this approximation becomes exact as $h \to 0$. Thus

$$\text{area}(S_{p,r}) = \frac{d}{dr}\text{vol}(S_{p,r}) = \frac{d}{dr}\frac{4}{3}\pi r^3 = 4\pi r^2$$

16.3. Higher Dimensions

What about the fourth dimension? Can we fiigure out how spheres work there? The fact that lines are given by affine eqautions holds true in all dimensions, which allows us to write down the distance formula in $4D$ and the equation of a sphere exactly as before.
To keep things simple we can start again with the 4-dimensional unit sphere, which is described by

$$x^2 + y^2 + z^2 + w^2 = 1$$

Let's call this sphere H (or $H_{O,1}$) for *hypersphere*. We wish to find H's volume by slicing, where we take *three dimensional slices* with constant w: these slices will intersect

the 4-dimensional ball by *solid three dimensional balls* much as we sliced the 3D ball into filled in 2d circles, and sliced circles into intervals!

It's going to get difficult to keep dimensions straight here, so I'm going to start subscripting our volumes: I'll write vol_3 for the usual three dimensional volume we know and love, and I'll write vol_4 for the new four dimensional hypervolume. This slicing tells us

$$\begin{aligned}\text{vol}_4(H) &= \int_{-1}^{1} \left(\iiint_{S_r} dxdydz \right) dw \\ &= \int_{-1}^{1} \text{vol}_3(S_r) dw \\ &= \int_{-1}^{1} \frac{4}{3}\pi r^3 \, dw\end{aligned}$$

This leaves us once again with a single integral to do! And all we need is the relationship between the radius r and the height w, which is exactly the same as in the dimension below:

$$r(w) = \sqrt{1 - w^2}$$

In theory, all we have to do now is plug this in and integrate! In practice this integral is a bit more challenging than we have come across before (though nothing that you haven't seen already in a Calculus II course)

$$\text{vol}_4(H) = \frac{4\pi}{3} \int_{-1}^{1} \left(\sqrt{1 - w^2} \right)^3 dw$$

This integral requires a trigonometric substitution to complete. It's perhaps easier to deal with the bounds if we first realize the integral is an even function, and so we could instead just integrate on $[0, 1]$ and double the result:

$$\frac{4}{3} \int_{-1}^{1} \left(\sqrt{1 - w^2} \right)^3 dw = \frac{8\pi}{3} \int_{0}^{1} \left(1 - w^2 \right)^{\frac{3}{2}} dw$$

Now we can make the substitution $w = \sin\theta$, where we find $w = 0$ corresponds to $\theta = 0$ and $w = 1$ corresponds to $\theta = \tau/4$ (as $\sin \tau/4 = 1$). After this substitution we have

$$\begin{aligned}\text{vol}_4(H) &= \frac{8\pi}{3}\int_0^{\tau/4}\left(1-\sin^2\theta\right)^{\frac{3}{2}}d(\sin\theta)\\ &= \frac{8\pi}{3}\int_0^{\frac{\tau}{4}}(\cos^2\theta)^{\frac{3}{2}}\cos\theta d\theta\\ &= \frac{8\pi}{3}\int_0^{\frac{\tau}{4}}\cos^4\theta d\theta\end{aligned}$$

Now we have yet more work, as we have arrived at the integral of the fourth power of cosine. This requires some trigonometric work with the double/half angle identites we proved:

Exercise 16.5 (Integrating $\cos^4(\theta)$)**.** Use the identity $\cos^2 x = \frac{1}{2}(1+\cos 2\theta)$ twice to show that

$$\cos^4(\theta) = \frac{3}{8} + \frac{\cos 2\theta}{4} + \frac{\cos 4\theta}{8}$$

Then use this to confirm that

$$\int_0^{\frac{\tau}{4}}\cos^4\theta = \frac{3}{8}\frac{\tau}{4}$$

Putting this together with the above, we finally reach our answer (using that $\tau = 2\pi$)

$$\text{vol}_4(H) = \frac{8\pi}{3}\frac{3}{8}\frac{\tau}{4} = \frac{\pi\tau}{4} = \frac{\pi^2}{2}$$

This is the first time that our constant has *not* been a rational multiple of π, but instead a rational multiple of π^2! Since homotheties scale four dimensional volumes by a factor of r^4, we get that the full volume formula for a hypersphere of radius r

Theorem 16.9 (Volume of the Hypersphere)**.** *The volume of the 4-dimensional hypersphere of radius r is*

$$\text{vol}_4(H_{p,r}) = \frac{\pi^2}{2}r^4$$

From this we can get the three dimensional *surface area* by differentiation. Again to keep things straight, I'll write area_3 for the three dimensional analog of surface area in 4D space, and area_2 for the usual 2D area in 3D space that we have thus far been just calling area.

Theorem 16.10 (Surface of the Hypersphere)**.**

$$\text{area}_3(H_{p,r}) = \frac{d}{dr}\text{vol}_4(H_p, r) = \frac{d}{dr}\frac{\pi^2}{2}r^4 = 2\pi^2 r^3$$

Thus, the 3-dimensional surface area constant for hyperspheres is $2\pi^2$: also a multiple of π^2 because it arose from differentiating volume.

Exercise 16.6. Find the volume and surface area constants for the 5-dimensional sphere via integration by slicing (for volume) and then differentiation (for surface area).

16.4. A Surprise in Even Dimensions

If you complete Exercise 16.6 above, you'll find that the 5-volume has a rather strange-looking constant out front:

$$\text{vol}_5 = \frac{8}{15}\pi^2 r^5$$

What can we do with this information? Carry on the march to higher dimensions of course! If we try to find the volume of the unit 6-sphere by slicing, (say the axis we slice along is called w again, for convenience) we can write

$$\begin{aligned}\text{vol}_6 &= \int_{-1}^{1} \text{vol}_5(\sqrt{1-w^2})dw \\ &= \int_{-1}^{1} \frac{8}{15}\pi^2\left(\sqrt{1-w^2}\right)^5 dw \\ &= 2\frac{8}{15}\pi^2\int_{0}^{1}\left(\sqrt{1-w^2}\right)^5 dw\end{aligned}$$

Unfortunately this time (again!) we cannot get rid of the square root since 5 is an odd power, and we must resort to a trigonometric substitution $w = \sin\theta$. Skipping the now-familiar steps,

$$\int_0^1 \left(\sqrt{1-w^2}\right)^5 dw = \int_0^{\frac{\tau}{4}} \cos^6\theta\, d\theta$$

Now we need only expand out $\cos^6$ via trigonometric identities and integrate:

Exercise 16.7. Confirm, similarly to a previous exericse that

$$\int \cos^6\theta\, d\theta = \frac{5}{16}x + \frac{15}{64}\sin(2\theta) + \frac{3}{64}\sin(4\theta) + \frac{1}{192}\sin(6\theta)$$

And thus, that the definite integral we are after is

$$\int_0^{\frac{\tau}{4}} \cos^6\theta\, d\theta = \frac{5}{16}\frac{\tau}{4}$$

Plugging this back into our original expression we get some almost magical cancellation of all these constants:

$$\begin{aligned}\text{vol}_6 &= 2\frac{8}{15}\pi^2\frac{5}{16}\frac{\tau}{4}\\ &= \frac{\pi^2}{3}\frac{\tau}{4}\\ &= \frac{\pi^2}{3}\frac{\pi}{2}\\ &= \frac{\pi^3}{6}\end{aligned}$$

Theorem 16.11 (Volume of the 6-Sphere)**.** *The volume of the six dimensional sphere of radius r is*

$$\frac{\pi^3}{6}r^6$$

From here - if we were feeling brave - we could calculate the volume of the seven-dimensional ball by slicing (which would not need a trig sub, as the slices are 6 dimensional and the sixth power will get rid of the square root) yeilding

$$\text{vol}_7 = \frac{16}{105}\pi^3$$

Then use this to calculate the volume of the 8-dimensional ball by slicing (which will now need another trig sub, which will introduce another factor of π through the bound $\tau/4$). The result here has some ugly calculation and marvelous cancellations, ending with

$$\text{vol}_8 = \frac{\pi^4}{24}$$

A pretty interesting pattern is arising here - using vol_2 for the *two dimensional volume* (area) of a circle, we have

$$\text{vol}_2 = \pi \qquad \text{vol}_4 = \frac{\pi^2}{2} \qquad \text{vol}_6 = \frac{\pi^3}{6} \qquad \text{vol}_8 = \frac{\pi^4}{24}$$

It appears that the volume of the $2n$ dimensional ball is $\pi^n/n!$. Incredibly, this turns out to be correct:

Theorem 16.12 (Even volumes)**.**

$$\text{vol}_{2n} = \frac{\pi^n}{n!}$$

One way to prove this is to continue the process we have been doing, with the trig subs and all, but via induction (and being clever, realizing we only need to know the *constant term* of $\cos^{2n}(\theta)$ - all the rest integrate to zero every time!)

But there's an alternative way - one can try to integrate via slicing over two dimensions at once, and get a *recurrence relation* relating the volume in dimension n to the volume in dimension $n-2$:

Proposition 16.1.

$$\text{vol}_n = \frac{\pi}{n}\text{vol}_{n-2}$$

If you're interested in doing this - come talk to me in office hours! But now for the truly strange part: what is the sum of the volumes of all the even dimensional balls?

$$\sum_{n\geq 0} \text{vol}_{2n} = \sum_{n\geq 0} \frac{\pi^n}{n!} = e^{\pi}$$

WHAT?! This is the series expasion of e^x evaluated at π. But it gets even crazier. What if we add up the volumes of the spheres of radius r? This multiplies each term by r^{2n} (since they are even dimensional spheres) and equals

$$\sum_{n\geq 0} \frac{\pi^n}{n!} r^{2n} = \sum_{n\geq 0} \frac{(\pi r^2)^n}{n!} = e^{\pi r^2}$$

Why in the world is the sum of the volume of all the even dimensional balls what you get by plugging the area of the circle into the exponential function?! I have no idea...

Part IV.

The Sphere

17. Foundations

After a rather deep dive into the foundations and history of plane geometry, we are ready to leave the familiar behind and explore other worlds! The first new geometry we will consider is.....well.....actually also familiar: its the sphere. We've even met this geometry in our discussion of π, where we noted that using analogous arguments to what we did in the plane, the distance formula in three dimensions is a natural generalization of the pythagorean theorem, which provides an equation for the sphere.

Definition 17.1 (The Sphere (Points))**.** The (unit) sphere is the set of points $(x, y, z) \in \mathbb{R}^3$ lying at distance 1 from the origin.

$$\mathbb{S}^2 = \{(x, y, z) \in \mathbb{E}^3 \mid x^2 + y^2 + z^2 = 1\}$$

It's important to remember that by *sphere* mathematicians usually mean the surface, not the interior (we will call the interior of the sphere the *ball*). Thus, $\mathbb{S}^2$ is two dimensional, which is why we denote it this way, and call it the *two-sphere*.

The sphere has been studied since ancient times: we came across it most recently while analyzing the work of Archimedes, but it became of particular importance outside of mathematics around the same time, when Eratosthenes calculated the circumference of the Earth (quite accurately). But in both of these contexts we are picturing the sphere *extrinsically*, from the perspective of three-dimensional beings that could hold it in their hands.

Remark 17.1. Centuries earlier around 450BCE, the pre-Socratic philosopher Anaxagoras correctly postulated that the earth was a sphere, floating freely in the vacuum. But no one knew its size!

The big change in perspective here is that we are going to think of the sphere *as a geometry all on its own*, just like we did for the plane! We will work with coordinates in three dimensions to make our lives easier, but the surrounding 3-dimensional space is of no interest or consequence to us: the only space that is "real" is the surface of the sphere itself.

In some sense we are very used to this: as this is how we actually live our lives! Since evolution did not grace the great apes with wings, we humans spend almost all of our time walking around on the *surface* of a large sphere, unable to meaningfully interact with the totality of the 3-dimensional space it is embedded in. However, this

isn't totally helpful, as our two main ways of sensing the world around us, *sight* and *sound* depend on the physics of 3-dimensional space, and are not constrained to the sphere.

For me it is helpful to think about spherical geometry as the geometry a mathematically gifted-ant would discover if it lived its entire life on an orange weak and downward pointed eyes only able to perceive its immediate vicinity on the peel. What curves on the orange would the ant call lines? How would the ant measure angles and distances? Does the ant's mathematics contain the pythagorean theorem?

Remark 17.2. The first treatments of spherical geometry as a true intrinsic geometry in its own right come not from silly thought experiments about ants of course, but rather from *navigation using the stars*, where the celestial sphere modeled the sky, and spherical trigonometry was first developed.

17.1. Calculus on ^{2}

Having put all the work into understanding a modern, calculus-based approach to geometry in the plane, we will reap significant benefits here by seeing how many of the ideas remain *conceptually the same* on the sphere. Our infinitesimal foundations all rely on being able to take derivatives, so the first thing we should wonder is *what is the derivative of a curve on the sphere*? Happily, because the sphere lives in $\mathbb{E}^3$ and we understand Euclidean calculus well, we can directly borrow that notion:

Definition 17.2 (Calculus on the Sphere)**.** The sphere inherits its notion of calculus from the 3-dimensional space it lives in: if γ is a curve on the sphere then $\gamma(t) = (x(t), y(t), z(t))$ and $\gamma' = \langle x', y', z' \rangle$.

To really get things moving, we need to define a notion of *tangent space* to each point on the sphere. This space should be the set of all infinitesimal tangent vectors to curves to the sphere. Here we need to put a little more thought in than we did for the plane, where we just noted that the derivative to a planar curve was also a 2-dimensional vector, so the tangent space at each point should be another copy of the plane. Why? Well here we have represented points on the sphere with *three coordinates*, and so tangent vectors also have *three coordinates*. But this doesn't mean the tangent space at each point is three dimensional! Indeed, there are many three dimensional vectors at each point which are not tangent to any curve on the sphere.

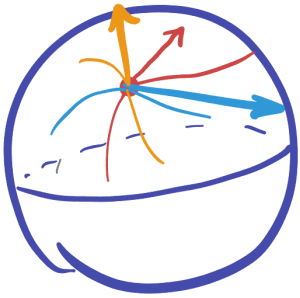


Figure 17.1.: Tangent vectors to ^{2} at a point are the derivatives of curves passing through that point.

Proposition 17.1 (Tangents to Curves on the Sphere)**.** *If γ is a curve lying on the surface of the sphere passing through a point $\gamma(t) = p$, then its tangent vector $\gamma'(t)$ is orthogonal to p in $\mathbb{E}^3$.*

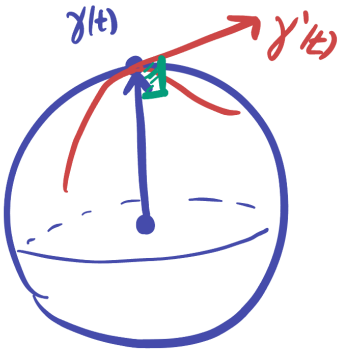


Figure 17.2.: Tangent vectors to a curve on the sphere through $p \in {}^2$ are orthogonal to the point p, thought of as a vector from the origin in $\mathbb{E}^3$.

Proof.

□

This argument relied on the observation that the dot product has its own *product rule*, which is a straightforward algebraic computation from its definition.

Exercise 17.1 (Product Rule for Dot Product)**.** Let $f(t) = \langle f_1(t), f_2(t), f_3(t)\rangle$ and $g(t) = \langle g_1(t), g_2(t), g_3(t)\rangle$ be two vector functions. Prove that the dot product satisfies the product rule:

$$\frac{d}{dt}\left(f(t)\cdot g(t)\right) = f'(t)\cdot g(t) + f(t)\cdot g'(t)$$

Definition 17.3 (The Sphere (Tangent Vectors))**.** If $p \in {}^2$, then the tangent space $T_p{}^2$ is the set of all vectors in $\mathbb{E}^3$ which are orthogonal to p:

$$T_p{}^2 = \{q \in \mathbb{E}^3 \mid p\cdot q = 0\}$$

In coordinates, if $p = (p_1, p_2, p_3)$,these are the points $\langle x, y, z\rangle$ such that $p_1x + p_2y + p_3z = 0$.

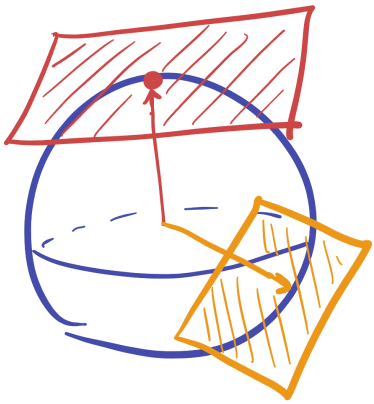


Figure 17.3.: Tangent spaces to the sphere are linear subspaces of $\mathbb{E}^3$ containing all vectors perpendicular to the position.

17.2. Geometry on ^{2}

Now that we have points and tangent vectors, we need to bring the actual *geometry* into the picture. In our original development of $\mathbb{E}^2$ we encoded all of geometry via the notion of an infinitesimal length. We then went on to develop all the higher level concepts like lengths of curves, and eventually angles - before discovering that the we could measure angles easily with the dot product! But since we can *also* measure infinitesimal lengths with the dot product, we saw that we could *alternatively* take this as the basis of all of geometry. We will take this bold new approach here with the sphere.

Definition 17.4 (The Sphere's Dot Product)**.** If $v = \langle v_1, v_2, v_3 \rangle$ and $w = \langle w_1, w_2, w_3 \rangle$ are two tangent vectors on the sphere based at a point p their dot product is computed using the standard dot product on $\mathbb{E}^3$:

$$v \cdot w := v_1 w_1 + v_2 w_2 + v_3 w_3$$

This gives rise immediately to our notion of infinitesimal length:

Definition 17.5 (Infinitesimal Length on ^{2})**.** Given a vector $v = \langle v_1, v_2, v_3 \rangle$ in the tangent space $T_p{}^2$, the infinitesimal length of v is the square root of its dot product with itself:

$$\|v\| = \sqrt{v \cdot v} = \sqrt{v_1^2 + v_2^2 + v_3^2}$$

Thus, each tangent space comes with an infinitesimal version of the pythagorean theorem, just like we had for $\mathbb{E}^2$! Remember what tangent spaces are all about: they're encoding the result of a limiting process of *infinite zoom*: the fact that we see the pythagorean theorem here on the tangent space is just the statement that zooming in on a point, a sphere appears to be flat! This we are quite used to from living on the surface of the earth. The magic we will see soon is that in fact *all of spherical geometry* can be recovered from this infinitesimal flatness.

Remark 17.3. Some people are *too impressed by this fact*, and have the mistaken impression that the earth actually is flat!

To define the length of a curve on ^{2} we follow the exact approach from $\mathbb{E}^2$, and use integration to promote thse infinitesimal lengths to finite ones.

Definition 17.6 (Lengths of Curves on ^{2})**.** Length of a curve is the integral of its infinitesimal lengths:

$$\text{length}(\gamma) = \int_I \|\gamma'(t)\| d$$

Now for angles, our new foundations make everything *much* easier! Instead of working hard (to define an angle as the arclength of the unit circle in the tangent space, spanned by two tangent vectors at a point), we instead note that we already *know* how this is related to the dot product in Euclidean space, and we know the tangent space IS euclidean (Definition 17.5). Thus, we can take the relation to the dot product as our *definition*:

Definition 17.7 (Angles on ^{2})**.** The angle between two vectors on the sphere is defined using the inner product:

$$\measuredangle(v, w) = \arccos\left(\frac{v \cdot w}{\|v\|\|w\|}\right)$$

Where here arccos can be calculated by the integral expression we derived in Proposition 14.2 (or by your calculator, which does this faster!)

Exercise 17.2. Consider the curves $\alpha(t) = (\cos t, \sin t, 0)$ (the equator of the sphere), and $\beta(t) = (0, \sin(t), \cos(t))$ (a line of longitude). Prove that they

- Intersect each other at the $t = \pi/2$
- Form a right angle at their point of intersection.

17.3. Isometries of ^{2}

Our fundamental tool for working with Euclidean space was *isometries*. In our development of the geometry, we tried to seek out as many isometries early on as we coould, and then continually used them to make our lives easier: moving points to the origin, lines to the x-axis, and so on.

The same approach will prove benificial on the sphere: it'll be nice to be able to move points to the north pole, or circles to the equator when we desire. So, let's track down some isometries! But first - what is an isometry here? We defined an isometry before as a function which preserved infinitesimal lengths, but that was because infinitesimal lengths were the foundation of our geometry. Now we've decided to take the dot product as our foundations so, perhaps we should change our definition of isometry here too?

Definition 17.8 (Isometries on ^{2})**.** An isometry of ^{2} is a function $\phi : {}^2 \to {}^2$ which preserves the dot product. Precisely, this means that if $p \in {}^2$ is a point and $v, w \in T_p{}^2$ are tangent vectors, then

$$v \cdot w = (D\phi_p v) \cdot (D\phi_p w)$$

However, it doesn't actually matter which we take as our definition (preserving infinitesimal length, or the dot product) they pick out precisely the same class of maps! In practice, when we want to prove something is an isometry, we will either show it preserves the dot product, or that it preserves infinitesimal lengths, whichever is easier. This perhaps surprising claim is justified by a result:

Theorem 17.1. *A function $f : {}^2 \to {}^2$ (or $\mathbb{E}^2 \to \mathbb{E}^2$, or $\mathbb{E}^3 \to \mathbb{E}^3$...) preserves all infinitesimal lengths if and only if it preserves the dot product.*

One direction of this theorem is straightforward: if a map ϕ preserves the dot product, then it certainly preserves infinitesimal lengths! After all, preserving the dot product means that for any vector v, we have

$$v \cdot v = (D\phi_p v) \cdot (D\phi_p v)$$

But length is just the square root of this expression, so this immediately implies $\|v\| = \|D\phi_p v\|$. The perhaps more surprising direction is the reverse: *if a map preserves all infinitesimal lengths, then it actually preserves the dot product.* The trick here is to show that it's actually possible to compute the dot product of two vectors using infinitesimal lengths (the reverse of what we did above!)

Exercise 17.3 (Dot Products from Lengths)**.** Prove that if v, w are two vectors then the following equation is true:

$$\|v + w\|^2 = \|v\|^2 + \|w\|^2 + 2\langle v, w\rangle$$

Solve this for the dot product (moving all the other terms to the other side of the equation), and then prove the following fact: if $\|u\| = \|D\phi_p u\|$ for *all vectors*, then $v \cdot w = (D\phi_p v) \cdot (D\phi_p w)$ (Hint: apply $D\phi$ to the equation!)

Because isometries are defined using the same basic machinery here as Euclidean space (preserving infinitesimal quantities) the theorems we proved there about their composition and inversion carry over without any change:

Theorem 17.2. *The composition of any two isometries of the sphere is an isometry, and the inverse of any isometry of the sphere is an isometry.*

So to find isometries of the sphere we just need to track down functions on $\mathbb{E}^3$ that preserve the dot product. But in linear algebra at least, such functions already have a name!

Definition 17.9. If A is a linear map $\mathbb{E}^n \to \mathbb{E}^n$ such that preserves the dot product $(Av) \cdot (Aw) = v \cdot w$, then A is called an *orthogonal matrix*.

Example 17.1. The linear map $(x, y, z) \mapsto (x, y, -z)$ is represented by an orthogonal matrix:

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

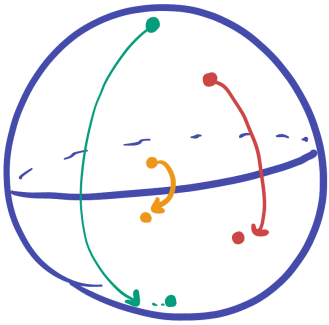


Figure 17.4.: The isometry $(x, y, z) \to (x, y, -z)$ on the sphere.

Exercise 17.4. A *permutation matrix* is a square matrix where every row and column has exactly one "1", and the other entries are zero. Prove the following permutation matrix is an orthogonal matrix:

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

These maps preserve the dot product on $\mathbb{E}^3$, but we need a little more than that to be sure they are isometries! Isometries of *the sphere* need to actually be maps $^2 \to ^2$.

Corollary 17.1. *If $\phi(x) = Ax$ is an orthogonal transformation of $\mathbb{E}^3$, then ϕ sends the unit sphere to the unit sphere.*

Proof. Since A is orthogonal it preserves the dot product. THus it preserves infinitesimal lengths, and so it preserves distances in $\mathbb{E}^3$. This means if $p \in ^2$ (so that p is distance 1 from the origin O) then $\phi(p)$ is also on the sphere (its distance 1 from $\phi(O)$, but ϕ sends the origin to itself, because its a linear map). □

Putting these facts together gives the following powerful theorem telling us tons of isometries of the sphere! (In fact, these are *all* the isometries of the sphere. But we don't need that here)

Theorem 17.3. *If A is an orthogonal matrix, then the function $p \mapsto Ap$ is an isometry of the sphere.*

This theorem gives us access to tons of isometries: all we need to do is track down orthogonal 3×3 matrices. We've already seen a couple explicit examples above (the reflection $(x, y, z) \mapsto (x, y, -z)$ and the permutation matrices in the examples), but it

will prove useful to dig a little deeper and try to figure out what kind of matrices are orthogonal. The following theorem of Linear Algebra gives us a complete classification:

Theorem 17.4. *A matrix is an orthogonal matrix if and only if all of its columns are unit vectors, and each column is orthogonal (hence the name) to every other.*

Now that we know the algebraic description of isometries (3×3 number squares where all the columns are orthonormal) we turn to the *geometry*: what do isometries of the sphere do?

The two most useful properties of isometries by far in $\mathbb{E}^2$ were the ability to move points around, and the ability to rotate any tangent vector to any other: these were the properties we called *homogenity* and *isotropy*. It was these two properties that gave the plane its incredible symmetry.

The sphere is of course very symmetric looking as well, and we are used to from our experience in day-to-day life with the ability to rotate a sphere any which way we like. But now we should *prove it*:

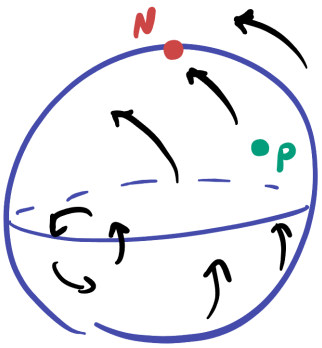


Figure 17.5.: Any point can be sent to the north pole of the sphere, or equivalently, you can send any point to the north pole via an isometry.

Proposition 17.2 (Any Point Moves to the North Pole)**.** *Let $N = (0, 0, 1)$ denote the north pole of the sphere, and p and arbitrary point on the sphere. Then there is an isometry of ^{2} which moves N to p. (And thus its inverse moves p to N).*

Proof. We will find an orthogonal matrix A so that the isometry $\phi(x) = Ax$ takes N to p. Since $N = (0, 0, 1)$, applying a linear map A to the vector N gives us the third colum of A. So, to begin to assemble such a map we will make its third column be p:

$$A = \begin{pmatrix} * & * & p_1 \\ * & * & p_2 \\ * & * & p_3 \end{pmatrix}$$

Now we just need to find values for six missing entires so that the all columns are orthogonal and unit length. In fact, there are many ways to do this! And we don't need any explicit solution we just need to know of their *existence*. So we will work column-by-column.

Call the second column of this matrix $u = (u_1, u_2, u_3)$. We know this must be orthogonal to p, so we have an equation this must satisfy:

$$u \cdot p = u_1 p_1 + u_2 p_2 + u_3 p_3 = 0$$

This is a single linear equation in three variables, and so it has many solutions (a two-dimensional space of solutions, in fact)! Taking any solution, we can rescale it to unit length, and use that as our second column.

Now for the first column, we have three unknowns (its three entires): but we have two equations - it must dot product with both the second and third column to zero. This still has an infinite number of solutions (in linear-algebra-speak, there's one 'free variable'), and choosing any solution and normalizing it gives a viable first column. □

Remark 17.4. The argument I give here is a *soft* or qualitative argument: we prove the existence of something without actually computing it. If you would like to actually *compute a specific matrix* that takes N to p (which is often useful in real-world applications of spherical geometry), you can do so by starting with any two vectors u, v where that $\{u, v, p\}$ is linearly independent, and apply the Gram-Schmidt process.

Theorem 17.5 (The Sphere is Homogeneous)**.** *Given any two points p and q on the sphere, there is an isometry taking p to q:*

Proof. Let N be the north pole of the sphere. Then by Proposition 17.2, we can find an isometry $|phi$ taking N to p, and another isometry ψ taking N to q. We will apply our by-now-standard trick, and compose one of these with the inverse of the other!

Specifically, the map ϕ^{-1} is an isometry which takes p to N, and ψ takes N to q so the composition $\psi \circ \phi^{-1}$ takes p to q, as desired. □

Next, we wish to see the sphere is also *isotropic*. We will do this in two parts (just like we did for $\mathbb{E}^2$)! First, we show that you can rotate the sphere about some specific point, and then we use homogenity to show we can actually do this at any point.

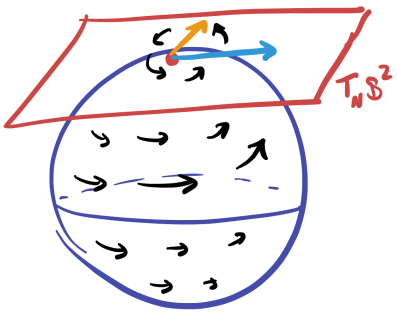


Figure 17.6.: Rotating the sphere about a point.

Proposition 17.3. *Let N be the north pole, and v be any unit vector in $T_N{}^2$. Then there exists an isometry ϕ of the sphere which fixes N and takes $\langle 1,0,0\rangle \in T_N{}^2$ to v.*

Proof. First, what sort of a vector is v? Its a unit vector in $T_N{}^2$, but what set of vectors is this? By Definition 17.3, its the set of vectors orthogonal to $N = (0,0,1)$. That is, the vectors $\langle v_1, v_2, 0\rangle$: its a horizontal Euclidean plane! So, $\langle v_1, v_2\rangle$ is a unit vector in this plane, and we want to rotate $\langle 1,0\rangle$ to this vector, and we know a matrix in the plane (from Euclidean geometry!) that does this:

$$\begin{pmatrix} v_1 & -v_2 \\ v_2 & v_1 \end{pmatrix}$$

How can we write down a transformation of $\mathbb{E}^3$ which does this to the horizontal plane and fixes the vertical direction (thus fixing N)? We can just insert it as the top 2×2 block of the matrix:

$$A = \begin{pmatrix} v_1 & -v_2 & 0 \\ v_2 & v_1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Its easy to see that this takes $\langle 1,0,0\rangle$ to v: as v is the first column of this matrix! So all we need to see is that this is actually an isometry: that A is an orthogonal matrix.

But this is likewise straightforward: we can take the dot product of any two columns and see we get zero (try it!) and, each column is unit length (because v was by hypothesis, and $(0,0,1)$ is). □

Exercise 17.5. Use Proposition 32.3 and Theorem 32.1 to show the sphere is isotropic: that given any point $p \in {}^2$ and any two unit vectors $v, w \in T_p{}^2$, there exists an isometry of ^{2} fixing p and taking v to w.

(Hint: first show you can do this when p is the north pole! Then use homogenity and a *conjugation*)

Now we have access to isometries that can move any point to any other point, and also rotate any vector to any other vector. This prepares us to prove the analog of the Euclidean theorem Exercise 32.28.

Exercise 17.6. Let p, q be any two points on the sphere, and v a unit vector at p and w a tangent vector at q. Then there is an isometry of [2] taking (p, v) to (q, w).

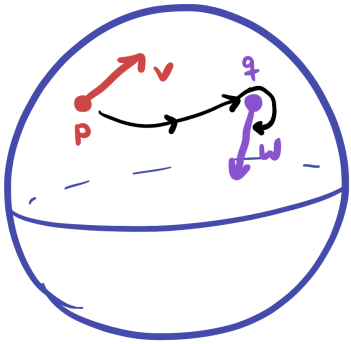


Figure 17.7.: The sphere is homogeneous and isotropic.

These are essentially *all* the facts that we will need about isometries of the sphere! But we would be remiss to not mention one very useful dichotomy between isometries of the sphere: the familiar groups of rotations vs reflections. Like everything else we've studied in this section, this concept is also captured infinitesimally (ie with linear algebra).

Definition 17.10. An isometry of the sphere $\phi(x) = Ax$ is a *reflection* if the det $A = -1$ and is a *rotation* if det $A = 1$.

This lets us see computationally that the matrix in Example 17.1 and **?@exm-permutation-orthogonal** are both reflections, whereas the matrix we created in **?@prp-sphere-homogeneous-step** is a rotation.

Exercise 17.7. Prove that you can find a rotation which takes N to any point p of the sphere.

(Hint: our earlier construction produces an isometry, but we don't know if its a rotation or reflection. If it *is* a reflection, can you modify it somehow so that it becomes a rotation, without changing the fact that it sends N to p?)

18. Lines & Circles

Now that we've made it through the fundamentals of spherical geometry, we are ready to move on from the infinitesimal to the actual, finite geometric properties we are usually interested in.

In this section, again much of the details will be similar to what we have already seen in the Euclidean plane, and because of those similarities we will be able to make rather fast progress. However the actual statements we can *prove* will start to differ - signaling something is truly different about this geometry. We will take up the cause of this difference - curvature - in the following chapter.

18.1. Lines

In Euclidean geometry we considered several distinct definitions of the term *line*, and then proved that all three definitions pick out the same class of curves. This allowed us the freedom to freely switch between the three definitions,

- Length Minimizing
- Straightest
- Lines of Symmetry

when convenient. The same holds for the sphere: when we are seeking the fundamental curves of this geometry we can either look for length minimizers, or for curves that do not turn, or curves that are fixed by an isometry.
I will often call a curve satisfying any of these three equivalent properties a *line*, because these curves play the same role in the theory of the sphere that our original lines do in the plane. But theres a more ancient term which originated with the sphere, and is now commonly used for this generalization of *line* all across mathematics.

Remark 18.1. Careful readers will notice that here I am just *claiming* that all three definitions remain the same on the sphere, we have not yet *proved it.* We will prove it in time, but it will be best to wait until we have developed some more tools, so we can avoid difficult and unenlightening integrals.

Definition 18.1 (Geodesic)**.** A geodesic on the sphere is a curve satisfying any of the three equivalent properties which defined lines in the plane.

The term *geodesic* is Greek, originally deriving from γεωδαισία, or *division of earth*, as a line across the surface of the earth divides it in two. This grew into *geodesy* or *measurement of the earth* in english, and then to *geodesic* in mathematics.

18.1.1. Curves Fixed by Isometries

Because we just spent all this effort dealing with isometries it will turn out to be easiest to discover which curves are lines using the *lines of symmetry* definition. Recall that we say a point p is *fixed* by an isometry ϕ if $\phi(p) = p$. Analogously, we say that an entire curve γ is fixed by ϕ if for each value of t, the point $\gamma(t)$ is fixed by ϕ - like the line a mirror sits on when it reflects the plane. What is the analog on ^{2}?

Example 18.1. The equator $(x, y, 0)$ of the sphere is a geodesic.

Proof. Consider the first isometry of the sphere we met, $(x, y, z) \mapsto (x, y, -z)$. Which points are fixed by this? Well, if z is nonzero the point is not fixed, as its sent to a point with third coordinate $-z$. However, whenever $z = 0$ this point is fixed by the isometry! Thus, the set of points $(x, y, 0)$ on ^{2} is fixed, making it a line of symmetry, and thus a geodesic. □

This is the first major difference between the sphere and the plane: we found a geodesic on the sphere, but that geodesic closes up, and is finite in length!

Corollary 18.1. *Euclid's Postulate 2 is false for the sphere, as line segments cannot be extended indefinitely: once you extend a segment of the equator to length $\tau = 2\pi$, it closes up!*

From our perspective as 3-dimensional beings looking at the sphere one easy way to describe the equator is that its the intersection of a *plane* through the origin with the sphere. Such things are called *great circles* (for a reason we'll understand better shortly)

Definition 18.2 (Great Circle)**.** A great circle is a curve on the sphere which is the intersection of ^{2} with a plane passing through the origin in $\mathbb{E}^3$.

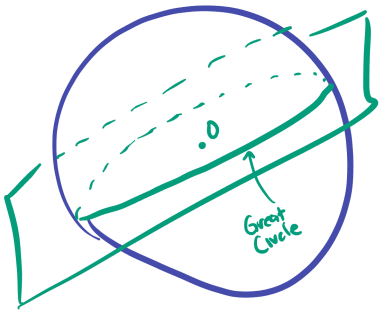


Figure 18.1.: A great circle is the intersection of the sphere with a plane through the origin.

So far, we know that one great circle is a geodesic. But we also know tons of isometries of the sphere! And just like we did for Euclidean space (where we started only knowing the x-axis was a line) we can use these to find all the other lines:

Theorem 18.1 (Great Circles are Geodesics)**.** *Let C be any great circle on ^{2}. Then C is a geodesic.*

Proof. Just like we did for the equator, our goal is to find an isometry of ^{2} which fixes the great circle C. Our strategy will mirror when we did this in Euclidean space - we'll find an isometry that takes C to the equator, and then use the fact that we know how to reflect in the equator to figure out how to reflect in C.

But how to do this? Well, the isometries we found for ^{2} are all *orthogonal transformations*, which take pairs of *orthogonal vectors* to other *orthogonal vectors* (in fact, they preserve the entire dot product, and thus the angle between any two vectors of course). Since the natural language we have available for talking about isometries discusses *orthogonality*, how could we describe the equator using this language?

The equator is all the points which are orthogonal to the north pole $N = (0, 0, 1)$! Similarly, our great circle C is the intersection of ^{2} with some plane P - and this plane has a unit normal vector $v \in \mathbb{E}^3$, where we can describe C as the points of the sphere which are orthogonal to v.

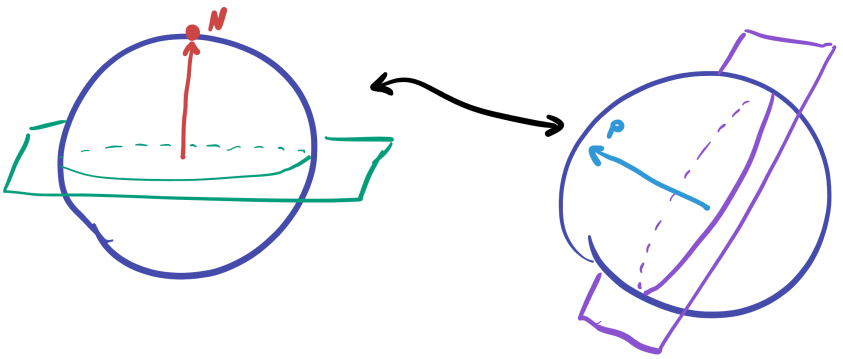


Figure 18.2.: Every great circle is a geodesic, as it can be moved via an isometry to the equator.

With this simple observation, we are already almost done! We know that we can find an isometry which takes N to v (Proposition 17.2), so call such an isometry ϕ. Because of how we constructed these isometries, we know ϕ is an orthogonal transformation, and preserves angles. Thus, if p is any point of ^{2} orthogonal to C, its sent to a point which is orthogonal to N! This means the circle C is sent to the equator, as required.

Now let $R(x, y, z) = (x, y, -z)$ be the reflection in the equator that we used to prove it was a geodesic. From R and ϕ we can build the isometry

$$\psi = \phi \circ R \circ \phi^{-1}$$

This takes C to the equator, reflects in the equator, and then returns the equator to C. Thus any point on C is unchanged by ψ, so C is a *fixed curve* by this isometry - its a geodesic! □

The realization that great circles are geodesics has another nice corollary - it makes it easy for us to draw a line between any two points of the sphere! This was the content of Euclids' axiom I, so we may say that this still holds on ^{2}.

Remark 18.2. In modern mathematics, the ability to draw a geodesic between any two points of a space remains very important - spaces where you can do this are called *geodesic metric spaces*.

Proposition 18.1. *Given any two points p and q of ^{2}, its possible to draw a geodesic segment connecting them.*

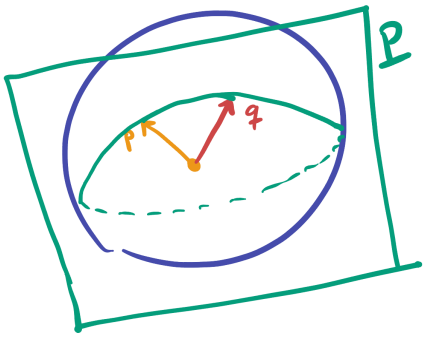


Figure 18.3.: A great circle can be drawn through any two points.

Proof. If p, q are two points of the sphere, form a plane P through the origin containing both p and q. (When p is not exactly opposite q on ^{2} they are linearly independent, so this plane is just their span. If $p = -q$ then you can take any plane you like containing the line of multiples of p). This plane intersects the sphere in a great circle containing both p and q, so there is a geodesic of ^{2} that passes through p and q. □

Instead of describing geodesics as connecting two points, we can also describe geodesics in terms of their starting point and a starting direction. Like we did in Euclidean space, we can use Exercise 32.52 to see that given any point p and any unit tangent vector v on the sphere, there is a geodesic passing through p in direction v (use this exercise to translate the equator, which passes through $(1, 0, 0)$ with tangent $\langle 0, 1, 0\rangle$ to p and v respectively).

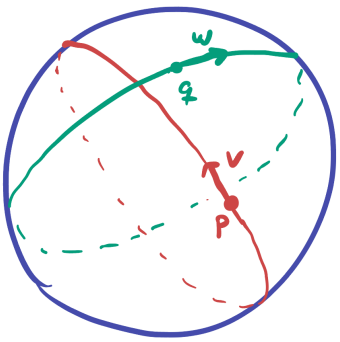


Figure 18.4.: There is a unique geodesic through each point p on the sphere, in each direction v.

18.1.2. Straightness on the Sphere

This section is suggested - but optional - reading, as we have already found all the geodesics above! But, while the definition of *line of symmetry* was the easiest to translate onto the sphere, its worth pausing a bit to talk about how we might define *straightness* here. In Euclidean space we said a curve was straight if its tangent vector did not change in time. But this definition will not do for ^{2}! Indeed, any curve whose tangent vectors do not change in time can be written as an affine equation, and none of these lie on the sphere!

What does "straight" mean on the sphere? It still means "does not turn", but we must be careful since we are working with the sphere inside of $\mathbb{R}^3$, and all curves restricted to the sphere bend with the sphere itself.

Definition 18.3 (Spherical Acceleration)**.** If γ is a curve on the sphere, its spherical acceleration at p is the projection of γ'' onto the tangent space $T_p{}^2$.

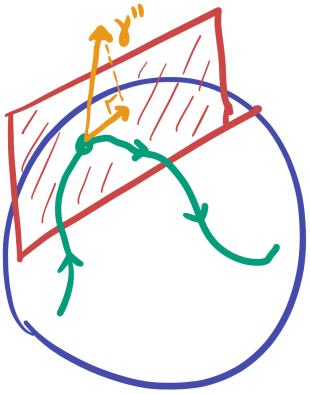


Figure 18.5.: Spherical acceleration of a curve (short yellow vector) is the projection of γ'' onto the tangent space.

This definition is precise, but not *useful* - we would like to have a formula which will let us compute the exact value of the spherical acceleration of any curve. And to get one - we need to do some Euclidean geometry! The key will be the ability to project a vector onto a plane.

Theorem 18.2. *Let P be a plane in $\mathbb{E}^3$ with normal vector $\vec{n}$. Then if $v \in \mathbb{E}^3$ is a vector (a point, thought of as a vector from the origin), the projection of v onto P is given by*

$$\text{proj}(v) = v - \frac{v \cdot n}{n \cdot n} n$$

Proof. Let v be any vector, and P a plane with normal vector n.

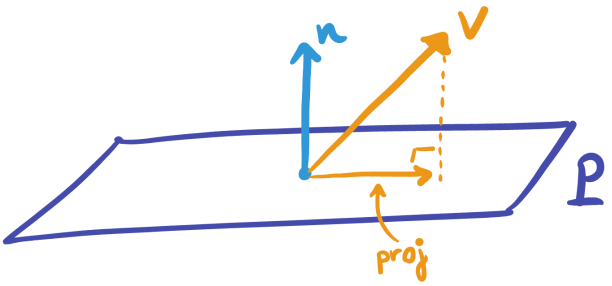


Figure 18.6.: The projection of a vector v onto a plane P

Our goal is to figure out how much of the vector v lies in the plane P, and our approach will be kind of backwards. We will figure out how much of v is *not* in the plane, and the subtract this!

To figure out how much of v is not parallel to the plane P, we need to figure out how much is in the direction of the normal vector n. That is, we wish to compute the projection of v onto the line spanned by n:

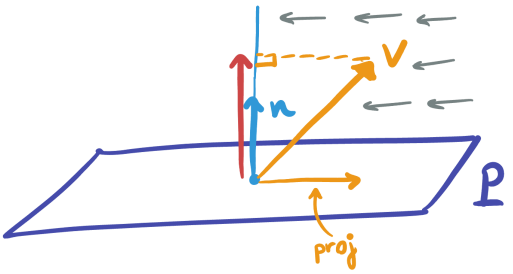


Figure 18.7.: The projection of a vector v onto the normal vector n to a plane.

This is actually a problem of Euclidean plane geometry, which we can solve using angles and the dot product! Let's look just in the Euclidean plane containing v and n, where the projection of v onto n forms a right triangle with hypotenuse v. From here, we can see the quantity we want is $\|v\| \cos\theta$, where θ is the angle between v and n.

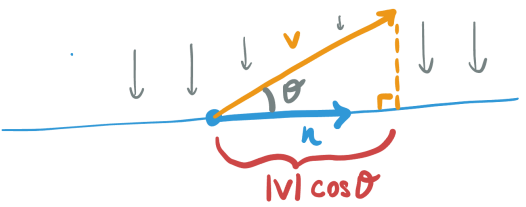


Figure 18.8.: Calculating the projection of v onto the line spanned by n.

But, we also know that $\cos\theta$ is defined in terms of the dot product!

$$\cos\theta = \frac{v \cdot n}{\|v\|\|n\|}$$

Thus, the projection onto n is

$$\|v\|\cos\theta = \|v\|\frac{v \cdot n}{\|v\|\|n\|} = \frac{v \cdot n}{\|n\|}$$

This is the *length* of the projection of v onto n: what we need now is a vector in the direction of n, which has this length. The solution? Just multiply by the unit vector in direction n!

$$\frac{v \cdot n}{\|n\|}\frac{n}{\|n\|}$$

This can be simplified algebraically, since we have two copies of $\|n\|$ in the denominator now:

$$\frac{v \cdot n}{n \cdot n}n$$

Phew! But - this is the amount of the vector *not in the plane*. It's exactly the part of v that we don't care about! To get what we want, we need to subtract this from v:

$$\text{proj} = v - \frac{v \cdot n}{n \cdot n}n$$

□

It'll be useful to note that this formula simplifies a bit if $\|n\| = 1$, to become just

$$\text{proj}(v) = v - (v \cdot n)\, n$$

. This allows us to project onto any plane we wish in $\mathbb{E}^3$, so long as we know its normal vector. But our goal is to project onto the *tangent plane to* ^{2}, so we can do even better, since we know exactly what the tangent spaces are. Indeed, if v is a vector in $\mathbb{E}^3$ based at a point p on the sphere, we can write down a projection map directly to the sphere:

Definition 18.4 (Projecting onto $T_p{}^2$.)**.** If v is a vector in $\mathbb{E}^3$ based at p, then its projection onto $T_p{}^2$ is

$$\text{proj}_2(v) = v - (v \cdot p)p$$

Indeed, at the point $p \in {}^2$, the tangent space is the set of all vectors orthogonal to p - so the normal vector to $T_p{}^2$ is just p itself! This

Corollary 18.2 (Spherical Acceleration)**.** *Given a curve $\gamma(t)$ on ^{2}, its spherical acceleration is*

$$\begin{aligned}\text{acc}_\gamma(t) &= \text{proj}_2(\gamma''(t)) \\ &= \gamma''(t) - (\gamma''(t) \cdot \gamma(t))\,\gamma(t)\end{aligned}$$

Now we can formally define what it means for a curve on the sphere to be straight. Because this process produces an equation that γ has to satisfy to be a geodesic, it is called the *geodesic equation* and versions of it are fundamental to modern geometry - from the sphere to hyperbolic space to black holes and beyond.

Definition 18.5 (The Geodesic Equation for ^{2})**.** A curve on the sphere is a geodesic if its spherical acceleration is zero. That is, γ is a geodesic if $\text{acc}_\gamma(t) = 0$, or

$$\gamma''(t) - (\gamma''(t) \cdot \gamma(t))\,\gamma(t) = 0$$

Remark 18.3. This looks pretty daunting - at least in comparison to what we had to do in Euclidean space! There, our equation for straightness was just $\gamma'' = 0$, which we could solve by hand using Calculus I. This equation howver is much more complicated! If we write out γ in terms of components $\gamma(t) = (x(t), y(t), z(t))$ we can expand this equation all the way out into a system of three equations, for x, y, z:

$$\gamma'' = (\gamma'' \cdot \gamma)\gamma$$

Writing $\gamma(t) = (x(t), y(t), z(t))$ we can fully write this out as a vector equation

$$\begin{pmatrix} x'' \\ y'' \\ z'' \end{pmatrix} = \left(\begin{pmatrix} x'' \\ y'' \\ z'' \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} \right) \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

And, this vector equation is really just a system of three equations, which I'll write out below (expanding the dot product, which is the common factor multiplied to all of them):

$$\begin{Bmatrix} x'' = (x''x + y''y + z''z)\,x \\ y'' = (x''x + y''y + z''z)\,y \\ z'' = (x''x + y''y + z''z)\,z \end{Bmatrix}$$

In less symmetrical geometries, the only way forward from here is to actually *solve* this system of coupled differential equations! However, luckily for the sphere there's plenty of symmetry, and we can avoid this mess.

Now its our goal to show that the collection of curves which are straight on the sphere is the same as the collection which are fixed by some symmetry: that just like in Euclidean space, these two notions of *geodesic* coincide!

Theorem 18.3 (Straight Curves on the Sphere)**.** *All great circles are straight - they have zero spherical acceleration.*

Proof. First, we start with the equator $e(t) = (\cos t, \sin t, 0)$. We need to show that $\text{acc}_e(t) = 0$, by computing

$$\text{acc}_e = e'' - (e'' \cdot e)e$$

Computing we see $e'' = -e$ since its components are sinusoids, and so $e'' \cdot e = (-e) \cdot e = -(e \cdot e) = -1$ because e lies on the unit sphere. Plugging this in, we see that $\text{acc}_e = e'' - (-1)e = e'' + e$ But now we can use once more that $e'' = -e$! Thus

$$\text{acc}_e = e'' + e = -e + e = 0$$

So, the equator experiences no spherical acceleration, and thus is straight as claimed.

Next, we need to show this for an arbitrary great circle $c(t)$. Let P be the plane containing this great circle, and p be its normal vector. Then since we know there is an isometry of ^{2} taking N to p, we have an isometry taking $c(t)$ to the equator $e(t)$, and thus its inverse is an isometry taking $e(t)$ to $c(t)$: because this isometry is an orthgonal transformation I will represent it by its matrix A, so we can write

$$c(t) = Ae(t)$$

Where the juxtaposition Ae is matrix multiplication. Now we wish to use the fact that we know e is straight, to show that c is too. Our overall goal is to compute $\text{acc}_c(t)$, so let's write this out:

$$\begin{aligned} \text{acc}_c &= c'' - (c'' \cdot c)c \\ &= (Ae)'' - ((Ae)'' \cdot (Ae))(Ae) \end{aligned}$$

To simplify, we need to compute the second derivative of the composition $Ae(t)$. Let's just take one derivative at a time: by the chain rule

$$(Ae(t))' = DA_{e(t)}e'(t)$$

and, since A is a linear map it is its own derivative: $DA_{e(t)} = A$! Thus this simplifies to the statement "you can pull A out of the derivative", and similarly upon differentiating once more:

$$(Ae(t))' = Ae'(t) \implies (Ae(t))'' = Ae''(t)$$

Plugging this into our spherical acceleration formula, we find

$$\text{acc}_c = Ae'' - ((Ae)'' \cdot (Ae))Ae$$

There's still plenty of simplification to be done! The first order of business is to deal with the dot product here. Since A is an isometry it preserves the dot product by definition:$(Av) \cdot (Aw) = v \cdot w$. Applying this in our case we can remove the As above, resulting in

$$Ae'' - (e'' \cdot e)Ae$$

Now, whatever $e'' \cdot e$ is, its just a number for each value of t. Sine A is a linear map, we can pull it inside the map, and then use linearity to combine everything together:

$$\begin{aligned} Ae'' - (e'' \cdot e)Ae &= Ae'' - A((e'' \cdot e)e) \\ &= A(e'' - (e'' \cdot e)e) \end{aligned}$$

But now we are done! Look at what we are applying the linear map A to here: its nothing other than the spherical acceleration of the equator e! And we already know that e is straight, so this is zero. Finally, since A is a linear map it sends zero to zero, and so

$$\text{acc}_c = A(\text{acc}_e) = A(0) = 0$$

Thus c is also straight! □

This argument was pretty long and algebraic, more so than many of the more geometric arguments we've given throughout the course. I wanted to present it this way to show the power of all the tools we built: we managed to prove something about the curve c using nothing but properties of isometries and calculus: this is how many arguments in modern geometry proceed.

18.1.3. Distance on the Sphere

Now that we know the geodesics of the sphere, we can really get geometry moving by deriving the *distance function*.
But there is one subtlety we have to confront first. In Euclidean space, we proved that given any two points there was a *unique* line segment connecting them. And then we took the length of this line segment as the distance between them. But this seemingly simple fact is **false** on the sphere! Here geodesics are *circles*, and given two generic points there are actually two ways to connect them with segments of a great circle - going around the circle one way, or the other!

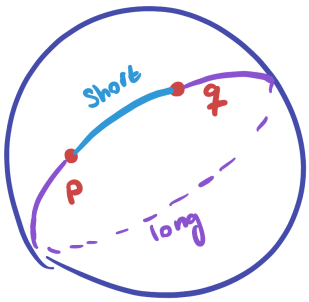


Figure 18.9.: There are at least *two* geodesics between any two points of the sphere! Generically, one is shorter than the other.

Exercise 18.1. True or false, between any two points on the sphere there are *exactly two* geodesic segments connecting them. (Can there ever be more? If so, when and how many?)

Of course, in general one of these is shorter than the other, and this does not pose any big theoretical problem. We just have to amend our terminology a bit, as in the plane we found the distance between two points was the length of **the** line segment connecting them.

Definition 18.6. The distance between two points p,q on the sphere is the length of the *shortest* curve connecting them. This is the length of the shorter of the geodesic segments defined by the great circle passing through p and q.

Believe it or not, we've already done all the rest of the hard work, and the distance formula is sitting here waiting for us to realize it! First, let's consider two points on the equator. Since the equator is just the unit circle in the Euclidean plane $(x, y, 0)$, these two points determine two vectors on the unit circle, and what we want to know is the length of arc between them.

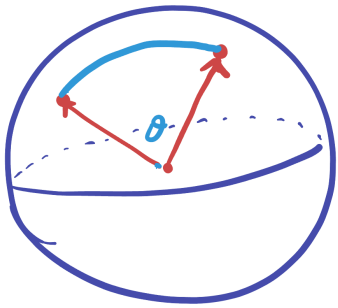


Figure 18.10.: Distance on the sphere is an arclength of a great circle. Thus, distance is an angle!

But this is the *definition of angle*! So, in spherical geometry, we have an beautiful relationship between distance and angle:

The distance between p and q on ^{2} is the same as the angle between them as vectors in $\mathbb{E}^3$.

Even better, we spent plenty of time working out exactly how to measure angles quantitatively, in the end discovering a nice relationship between the angle θ and the dot product. Furthermore, while we developed all of this material just in the plane, isometries do not change lengths (and thus do not change angles), so we can take this result from the Equator and apply it to any great circle on ^{2} via an isometry:

Theorem 18.4. *let p, q be two points on ^{2}. Then the distance between them is equal to*

$$\text{dist}(p, q) = \arccos(p \cdot q)$$

Just like the distance formula in $\mathbb{E}^2$ was the key to unlocking the rest of geometry (circles, trigonometry, and π), so is this distance formula, to unlocking the rest of the geometry of the sphere.

18.2. Circles

A circle is defined to be the set of points which are the same distance (the radius) from a fixed point (the center). We wish to study these curves in spherical geometry, now that we have the distance function available to us.

Like many things, its easiest to start working about a familiar point (the origin in $\mathbb{E}^2$, the north pole in ^{2}) and generalize beyond that using isometries.

Proposition 18.2 (Circles about the North Pole). *The circle of radius r about $N = (0, 0, 1)$ in ^{2} is given by the points*

$$C_{N,r} = \{(x, y, \cos(r)) \in {}^2\}$$

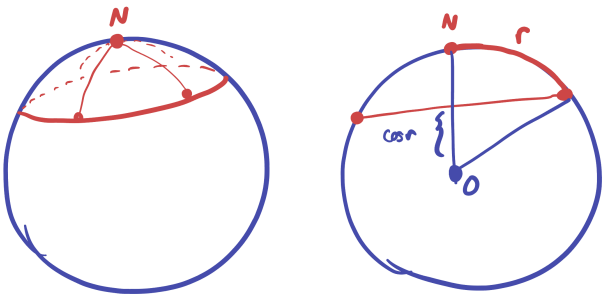


Figure 18.11.: A circle about the north pole N of radius r, and a "cross section view" showing why all z coordinates are $\cos r$.

Proof. We just need to see these are the points which lie at distance r from N along ^{2}. The crucial point is that we are measuring distances *along the sphere*, not in $\mathbb{E}^3$, so

$$\begin{aligned} d((x, y, z), N) &= \arccos((x, y, z) \cdot N) \\ &= \arccos((x, y, z) \cdot (0, 0, 1)) \\ &= \arccos(z) \end{aligned}$$

Thus, for (x, y, z) to lie on the circle it must (1) be a point on the sphere and (2) have $\arccos(z)$ equal to the radius - equivalently $z = \cos(r)$. □

Corollary 18.3 (Circles on the Sphere). *Circes are intersections of planes which* do not pass through the origin *with the sphere.*

We see this is true in the case centered on N, as the circle consists of all points with a fixed constant z coordinate: this is just a horizontal plane intersecting the sphere! To argue the general case, we use isometries: we know that any great circle can be taken to any other by an orthogonal transformation of $\mathbb{E}^3$ - and these are linear maps so they take planes to planes.

Thus, starting from a circle around the north pole cut out by a horizontal plane, moving this plane by an orthogonal transformation takes this plane to another plane, and its intersection to another circle!

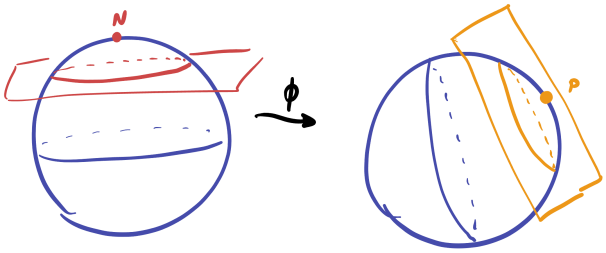


Figure 18.12.: Every circle on ^{2} is the intersection with some plane not through the origin.

We have come across two distinct curves on the sphere which can be described as the intersections of the sphere with planes. First, we had the great circles, corresponding to planes through the origin, which are the spherical analog of straight lines. These second ones are planes *not* through the origin, which correspond to geometric *circles*. What is going on here - how can these two different classes of curves seem so similar? After all, just shifting a plane slightly downwards can turn it from something curved (a circle) to something straight (a geodesic).

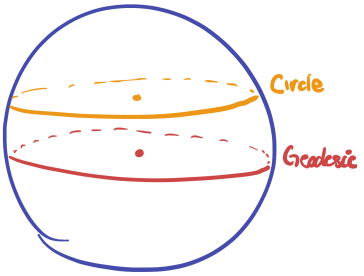


Figure 18.13.: On ^{2} both circles and geodesics can be found as intersections of planes with the sphere.

In fact, this is not as strange as it seems at first, and there's a sort of analog happening already in the Euclidean plane. Here, as a circle's radius grows larger and larger, the circle itself appears 'straighter' nearby. The difference is that this only becomes exact in the limit where the circle's radius goes to infinity, whereas on the sphere, this straightening of circles happens at a finite radius: $r = \tau/4$.

18.2.1. Area and Circumference

Now that we know what the circles on ^{2} are, its time to get quantitative and study their radii and circumference. In Euclidean space, we used isometries to show that any two

circles of the same radius could be brought to one another. This same argument goes through without change on the sphere:

Exercise 18.2. There is an isomery taking any circle of radius r to any other.

This was our first step to understanding the length constant τ: because isometries do not change lengths, this told us that every circle of radius r has the same circumference. And now, we know the same thing for the sphere! But the key step to understanding π and τ was figuring out how circles of different radii were related.

Theorem 18.5. *The only similarities of the sphere are isometries: there are no maps which non-trivially uniformly stretch infinitesimal distances.*

Proof. Let σ be such a similarity, which scales all infinitesimal lengths by k. Then σ preserves the dot product, and sends infinitesimal squares to other squares, expanding their area by k^2. This doesn't sound like a problem, until you remember that the entire spherical universe has a *finite area*!

The area of the unit sphere ^{2} is 4π, and if $\sigma : {}^2 \to {}^2$ is the supposed similarity above, it would take this area to an area $4\pi k^2$. But - this map is supposed to take the sphere to the sphere (it can't take it to a larger sphere in $\mathbb{E}^3$: that's a different space)!

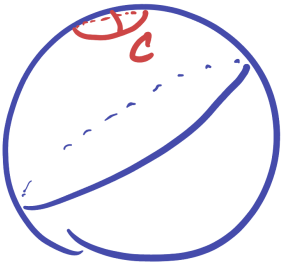
Thus, since the image is the same sphere we know that its area is still 4π, so $4\pi k^2 = 4\pi$, or $k^2 = 1$. Thus, the only possibility is $k = 1$ (as k is the scaling factor, a positive number), which corresponds to isometries. □

This simple fact - that ^{2} does not have any similarities - has profound consequences for its geometry. First off, it assures there's no hope of generalizing our proof that circumference/radius is a constant. Of course, this does not guarantee that its *not* a constant (perhaps we just need a different style of argument). To see what's really going on, we need to do some computations!

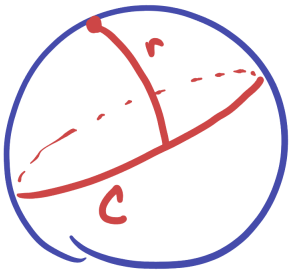
Before diving into the general case, its helpful to look at a couple of special cases: we will consider circles of very small radius, great circles, and circles of very large radius.

If the radius r is very small, then the circle itself fits within a very small region of ^{2}, and small regions of space are well-approximated by their tangent plane. Thus, we expect that small circles are very close to being Euclidean circles (think about drawing a circle on the sidewalk, which is a small portion of the spherical earth). Being nearly Euclidean, we expect their ratios of circumference to diameter to be very close to the Euclidean value of τ, or $2\pi \approx 6.28$.

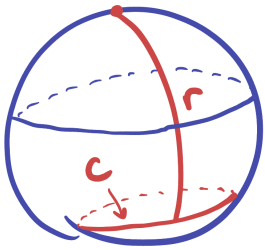
Now, what about a great circle? For specificity, let's consider the equator as a circle about the north pole. The radius is a quarter of a way around the sphere (half way would be from the north pole to the south pole, and the north pole to the equator is half of that). But the circumference is one full revolution around the sphere: so this



(a) A small circle has circumference-to-radius close to $2\pi \approx 6.28$



(b) A great circle has circumference-to-radius equal to 4.



(c) Large radii circles have circumference-to-radius as small as you like.

means that circumference over radius is $\frac{\tau}{\tau/4} = 4$! This shows us a very important fact - the sphere's analog of τ (the ratio of circumference to radius) cannot be a constant! It takes on values slightly larger than 6 for small circles, but takes on the exact value of 4 for a great circle.

How small can this ratio get? Consider a circle of a very large radius: close to π - so the radius runs from the north pole all the way down to something close to the south pole. The set of all points which lie at this distance from N is short curve near the south pole. So here the ratio of circumference to radius is a small circumference to a big radius - its a very small number!

These qualitative considerations not only show us that the spherical analog of τ is not a constant, but also that we expect it to be able to take on any values between 0 and 6.28. But to get more precise than this we actually need to sit down and do some calculations...

Proposition 18.3. *Show the circumference of the circle of radius r is*

$$2\pi \sin(r)$$

Proof. We know already that the circle of radius r about the north pole N is contained in the plane $z = \cos(r)$. But this is a *Euclidean plane*! So, we can measure this circles circumference using Euclidean geometry.

Figure 18.15.: Because circles on ^{2} lie on *Euclidean planes*, we can use Euclidean geometry to calculate their circumference.

Say its radius in the plane is d (note its radius on ^{2} is r, but this lies on the sphere, not on the plane). Then we can use our understanding of Euclidean circles to see $C = 2\pi d$. But what is d? Looking at a side view of our configuration, d is the opposite side of a right triangle with angle r inside a unit circle:

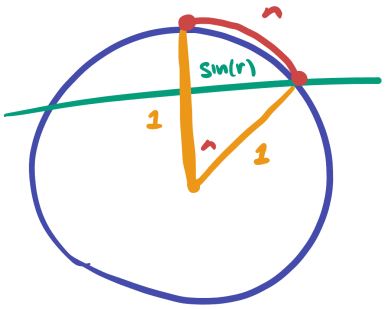


Figure 18.16.: The Euclidean radius is the sine of the spherical radius, since the spherical radius is an angle (distance equals angle on ^{2})

Thus, $d = \sin(r)$ and so $C = 2\pi \sin(r)$ as claimed. □

Corollary 18.4 (There is no length constant.)**.** *There is no single number like τ which is a universal constant for all circles on the sphere, like there was for circles on the plane.*

Proof. Recall we defined the function $\tau(r)$ to be the circumference of a circle of radius r, divided by its radius. In Euclidean space we found this was a constant, independent of circle. But on the sphere, we see

$$\tau(r) = \frac{2\pi \sin(r)}{r}$$

which is not constant! □

This pattern continues for area, where we show there is also no analog of π by seeing that areas of circles do not grow quadratically with radius. But first - how do we find the area of a circle on the sphere? We need to to use integration, to add up all the small infinitesimal area elements! Here the easiest way to do this is to invert our relationship between area and circumference discovered in Theorem 16.4. This same logic applies on the sphere, showing circumference to be the *derivative* of area.

$$\frac{d}{dr}\text{area}(r) = \lim_{h \to 0^+} \frac{\text{area}(r+h) - \text{area}(r)}{h} \approx \frac{\text{circ}(r) \cdot h}{h} = \text{circ}(r)$$

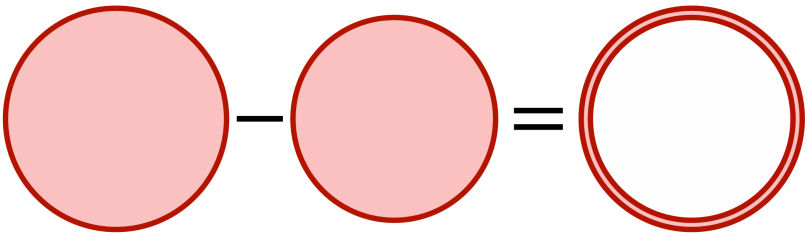


Figure 18.17.: The numerator of the derivative of area describes a thin ring, whose area is approximately the circumference of the circle times the difference in radii.

Thus, if $A'(r) = C(r)$, we can recover the formula for area via integration:

$$A(r) = \int_0^r C(r)dr$$

Proposition 18.4 (Area of a Circle)**.** *The area of a circle of radius r on ^{2} is*

$$A(r) = 4\pi \sin^2(r/2)$$

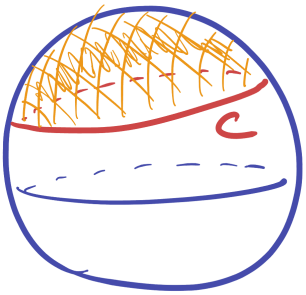


Figure 18.18.: When measuring the area inside a circle on ^{2}, we mean the area of the spherical cap whose boundary is the circle, and which contains the circles center.

Proof. This is just an explicit computation, using the result of **?@exr-sphere-circle-circumference**.

$$A(r) = \int_0^r 2\pi \sin(r) dr$$
$$= -2\pi \cos(r)\Big|_0^r = 2\pi(1 - \cos(r))$$
$$= 4\pi \frac{1 - \cos r}{2} = 4\pi \left(\sqrt{\frac{1 - \cos r}{2}}\right)^2$$
$$= 4\pi \sin^2\left(\frac{r}{2}\right)$$

Where, in the last two lines, we have used the half-angle formula to simplify our original answer, into a form that looks more similar to what we are used to in the Euclidean case. □

Exercise 18.3. Use the series expansion of $\sin x$ to give the first few terms of a series expansion of $A(r)$. Show that the first nonzero term is πr^2: this means when r is small, $A(r)$ is approximately πr^2. What does this mean geometrically?

18.3. Three Dimensions

The *three dimensional* version of spherical geometry is given by the surface of the four dimensional ball, just as the two dimensional sphere is the surface of the ball in three dimensions.

While it is hard to directly picture this space in four dimensions, its possible to compute things directly analgously to what we did above.

Definition 18.7 (Points and Tangent Spaces)**.** The points of ^{3} are the four tuples of length 1 in $\mathbb{E}^4$:

$$^3 = \{(x, y, z, w) \mid x^2 + y^2 + z^2 + w^2 = 1\}$$

A vector $v = \langle v_1, v_2, v_3, v_4 \rangle$ is tangent to ^{3} at $p = (x, y, z, w)$ if it is orthogonal to p:

$$T_p{}^3 = \{v \mid v \cdot p = 0\}$$

Much of the mathematics of ^{2} carries over to ^{3} with little change. In particular - we can find the geodesics by finding the *straight curves*, and see that they are again just great circles.

Theorem 18.6 (Geodesics)**.** *Geodesics on ^{3} are great circles: they are intersections of 3-dimensional hyperplanes in $\mathbb{R}^4$ with the unit sphere. For example $e(t) = (\cos t, \sin t, 0, 0)$ is a geodesic.*

Exercise 18.4. Prove this: write out what it means to have zero tangential acceleration, and prove that $e(t) = (\cos t, \sin t, 0, 0)$ is such a curve.

Because the geodesics are the same class of curves, we can measure *distance* in the same way - its an arc length of a circle, so distance equals angle!

Theorem 18.7. *If $p, q \in {}^3$ then the distance between p and q is given by*

$$\text{dist}(p, q) = \arccos(p \cdot q)$$

Exercise 18.5. Let $N = (0, 0, 0, 1)$ be the north pole of ^{3}. What are teh points of the sphere of radius r about N?

Exercise 18.6. What is the surface area of a sphere of radius r? What is its volume?

19. Curvature

We've seen that in some ways the sphere behaves similarly to the plane, and in other ways its quite different. Qualitatively, the big difference stems from the lack of any similarities other than isometries: this makes there be no universal constant like π or τ, for one. Our goal in this chapter is to *quantify* that difference.

In doing so, we will uncover the precise quantity of *curvature*: which measures how much geometry differs from that of the flat plane. This chapter will be short, but is discovery crucially important to all geometry beyond that of Euclidean space!

19.1. Circumference of Circles

Our first real quantitative difference between the sphere and the plane had to do with the the size of circles, so this is where we begin. We know (Definition 16.1) that for circles in Euclidean space $C = 2\pi r$, and by (**?@exr-sphere-circle-circumference**) that the analog in the sphere is $C = 2\pi \sin(r)$. Circles on the sphere grow *slower* than circles in the plane, as we can see by graphing these two functions.

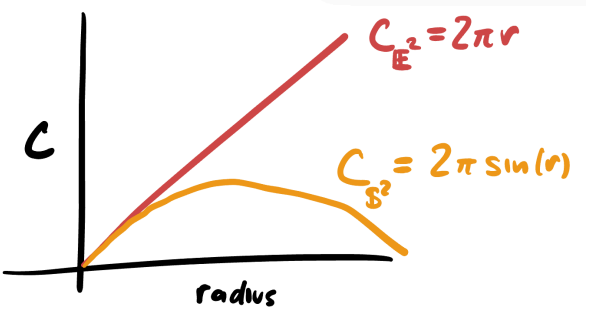


Figure 19.1.: A graph of the circumference of circles as a function of their radius, in Euclidean (red) and Spherical (yellow) geometry.

19.1.1. Limits

But how can we turn this *slower* insight into something quantitative, and infinitesimal? We want to be able to measure the curvature of the sphere *at a point* p, so we

should naturally be looking not at circles of some fixed finite radius, but rather families of circles that are shrinking down centered at the point p. How do these behave? Zooming in on the graph above we reach a first disappointment: it's hard to tell their behavior apart from Euclidean circles!

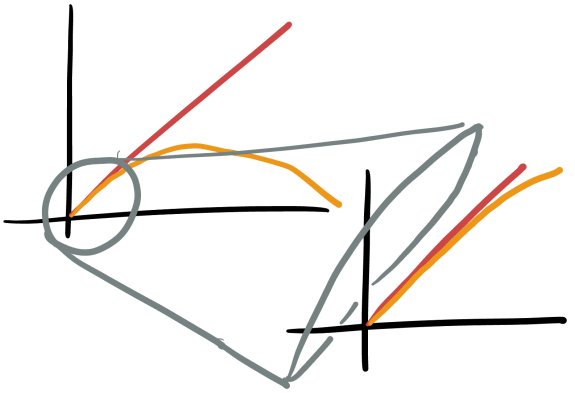


Figure 19.2.: Zooming in on small values of r, spherical geometry looks very much like plane geometry, which is reflected in the fact that it is difficult to tell their circumference functions apart.

This is because the series expansion of $\sin(r)$ starts out with $r - \frac{1}{3!}r^3 + \cdots$ and so $2\pi \sin(r)$ starts out with $2\pi r - \cdots$ - the same as in the Euclidean case! We already knew this - that at small scales the geometry of the sphere looks Euclidean, so what we are more interested in is the *difference* between the two geometries: that is, we care about

$$\lim_{r \to 0} C_{\mathbb{E}^2}(r) - C_2(r)$$

Or, at least - something like this! This can't be the right quantity all alone as when r shrinks, both of these go to zero, and so the limit just gives zero! This problem is reminiscent of when we define the derivative in Calculus I: if we just look at the difference in y values

$$\lim_{h \to 0}(f(x+h) - f(x))$$

the result goes to zero - which is not helpful! This is because we really need to be measuring a *ratio* - how much is this difference changing as x changes (or, in our case, as r changes).

This might suggest we take a look at the quantity

$$\frac{C_{\mathbb{E}^2}(r) - C_2(r)}{r}$$

But if we graph this quantity as $r \to 0$ we see this still goes to zero! In fact, the same happens if we normalize by a denominator of r^2: neither of these lets us see the difference between the sphere and the plane show up in the limit yet.

However, when we normalize by r^3, we actually get something that converges to a finite nonzero number!

Exercise 19.1. Check this, that as $r \to 0^+$ the following limits both exist, and are both equal to zero:

$$\lim_{r\to 0} \frac{C_{\mathbb{E}^2}(r) - C_2(r)}{r} = 0$$

$$\lim_{r\to 0} \frac{C_{\mathbb{E}^2}(r) - C_2(r)}{r^2} = 0$$

But

$$\lim_{r\to 0} \frac{C_{\mathbb{E}^2}(r) - C_2(r)}{r^3} \neq 0$$

what is its value?

Hint: recall that $C_{\mathbb{E}^2}(r) = 2\pi r$ *and* $C_2(r) = 2\pi \sin(r)$*, and L'Hospital's rule.*

This is a number that an inhabitant of the sphere could calculate for themselves: for smaller and smaller values of r they could compute this ratio by measuring things on the sphere, and look at what value the limit is approaching. And, the fact that they do not get zero would tell them definitively that they live somewhere other than the plane!

At this point, we could just *define* this ratio to be the curvature, but its convenient instead to normalize it: we multiply by a normalizing constant so that the curvature of the unit sphere is equal to $+1$:

Definition 19.1 (Curvature)**.** If X is *any* surface and $C_X(r)$ is the function which gives the circumference of the circle of radius r centered at p, then the curvature at p is defined by the limit

$$\kappa(p) = \left(\frac{3}{\pi}\right) \lim_{r\to 0} \frac{C_{\mathbb{E}^2}(r) - C_X(r)}{r^3}$$

A nice feature of this definition is that, being based totally off of lengths, its easy to check that curvature is not changed by isometries.

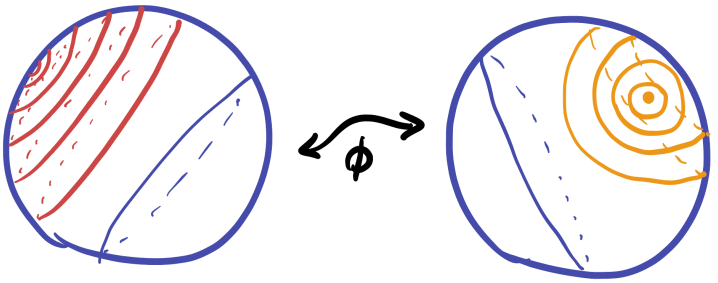


Figure 19.3.: Curvature is an isometry invariant.

Theorem 19.1. *If ϕ is an isometry that takes p to q, then $\kappa(p) = \kappa(q)$.*

Proof. Let ϕ be an isometry taking p to q. Then as ϕ does not change distances, it takes circles of radius r (a distance) based at p to circles of radius r based at q. And, as ϕ doesn't change the length of curves, it does not change the circumference of these circles.

Since the numerator and denominator of the limit expression are built purely using the distance r and circumference (and the Euclidean circumference, which is just $2\pi r$ - a multiple of r.) none of these quantities are changed by isometries, so the limit that we need to evaluate at p is *the exact same limit* as the one we need to evaluate at q: thus we get the same number both times.

PICTURE □

This has a nice corollary for homogeneous spaces: if there's an isometry that takes any point to any other, then the curvature at every point must be the same!

Corollary 19.1. *The unit sphere has constant curvature $+1$, and the Euclidean plane has curvature 0.*

19.1.2. Series & Derivatives

We can also approach this more algebraically than geometrically, after realizing that the correct geometric notion (a normalized difference) looks somewhat like a derivative. In particular, the series expansion of $\sin(r)$ is

$$\sin(r) = r - \frac{1}{3!}r^3 + \frac{1}{5!}r^5 - \cdots$$

so the series expansion of the circumference of a circle on ^{2} is

$$C_2(r) = 2\pi r - \frac{\pi}{3}r^3 + \frac{\pi}{60}r^5 - \cdots$$

And so, we can see that the third series coefficient is exactly what our limit was computing! But what is the third coefficient is a series expansion? Remember Taylor's formula:

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots$$

The third coefficient is just the third derivative divided by 3!. So, this means we can replace our limiting expression with $C'''(0)/3!$, and get a new expression for curvature:

Definition 19.2. Let $C(r)$ be the circumference function for circles of radius r centered at a point p, on some surface. Then the curvature at p is given by

$$\kappa(p) = -\frac{C'''(0)}{2\pi}$$

Proof. This is just a computation, plugging in our new term in place of the limit:

$$\begin{aligned}
\kappa(p) &= \left(\frac{-3}{\pi}\right)\lim_{r\to 0}\frac{C_X(r) - C_{\mathbb{E}^2}(r)}{r^3} \\
&= \left(\frac{-3}{\pi}\right)\frac{C'''(0)}{3!} \\
&= \frac{-1}{\pi}\frac{C'''(0)}{2} \\
&= \frac{-C'''(0)}{2\pi}
\end{aligned}$$

□

19.2. Area of Circles

We could also choose quantify the *curvature* of a space by comparing the area of circles to their Euclidean counterparts. Just like above, we could imagine two separate ways of extracting a quantitative number:

- A normalized limit of the difference between Spherical and Euclidean areas
- A normalized derivative of the Area Function

In both cases, we want to fix the normalizing factors so that the limit exists and the curvature comes out to be +1.

Exercise 19.2 (Curvature and Area: I)**.** Give a limit definition of the form

$$\kappa = \lim_{r \to 0} \frac{\text{area}_2(r) - \text{area}_{\mathbb{E}^2}(r)}{\text{normalizing factor}}$$

which computes the curvature of the sphere. What's the normalizing factor?

Exercise 19.3 (Curvature and Area: II)**.** Give a limit definition of the form

$$\kappa = C \cdot \frac{d^n}{dr^n} \text{area}_2(r) \Big|_{r=0}$$

which computes the curvature of the unit sphere. What is n, what is C?

19.3. Distance Between Geodesics

Besides circles, there's another interesting difference between the sphere and the plane we could try to quantify to measure curvature: how quickly geodesics spread out.

You are not responsible for this material now, but we will come back to it as an example when discussing the differences between spherical and hyperbolic space

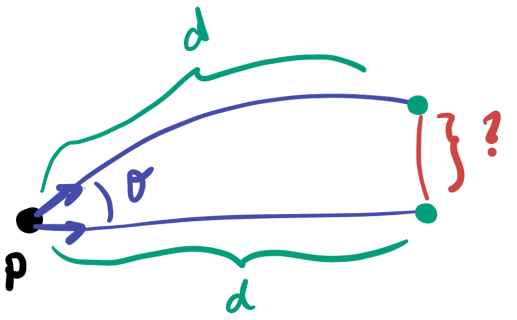


Figure 19.4.: Starting off at the same point and walking away from each other at angle θ, how far away are two travelers after going distance d?

To be precise, say we have two geodesics passing through a point p, which initially make an angle of θ with respect to one another at the point of intersection. After traveling for distance d along each geodesic, how far apart are the resulting points?

The question that turns out to be most interesting mathematically isn't the actual distance for some finite value of θ (as the formulas can get quite messy), but rather a more *infinitesimal notion*, looking only at geodesics right next to our original one. Thus we want to *zoom in* on θ near zero, which means we want to differentiate!

To make this precise, need a definition

Definition 19.3. Given a geodesic γ through some point p at $t = 0$, let γ_θ be the geodesic which makes angle θ with γ at p.

The collection of all geodesics for θ varying within some small interval about 0 is called a *geodesic variation* about γ.

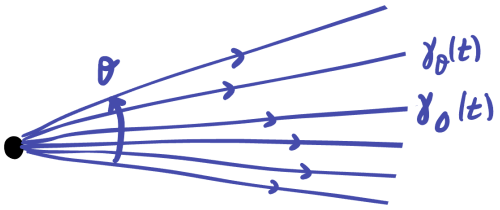


Figure 19.5.: A geodesic variation through p.

Definition 19.4. Given a geodesic γ, we can measure the *infinitesimal spread* of nearby geodesics by taking a geodesic variation $\gamma_\theta(t)$ and differentiating with respect to θ at 0:

$$J(t) = \left.\frac{d}{d\theta}\right|_{\theta=0} \gamma_\theta(t)$$

We can work this out in Euclidean space using the distance formula: if the point is O and we set one geodesic γ off in direction $\langle 1, 0\rangle$, the geodesic at angle θ starts with initial direction $\langle \cos\theta, \sin\theta\rangle$ by definition. Since geodesics are affine functions $p + tv$ we can write down their equations directly from this:

$$\gamma(t) = (t, 0) \qquad \gamma_\theta(t) = (t\cos\theta, t\sin\theta)$$

We see that these are spreading out linearly from one another with time, but how do we quantify this mathematically? By the infinitesimal variation!

$$\begin{aligned} J(t) &= \left.\frac{d}{d\theta}\right|_{\theta=0} \gamma_\theta(t) \\ &= \left.\frac{d}{d\theta}\right|_{\theta=0} (t\cos\theta, t\sin\theta) \\ &= \left.(-t\sin\theta, t\cos\theta)\right|_{\theta=0} \\ &= (0, t) \end{aligned}$$

The magnitude of the geodesic variation tells us how quickly nearby geodesics are spreading out away from γ. Here, in flat space we see

$$\|J(t)\| = \|(0,t)| = t$$

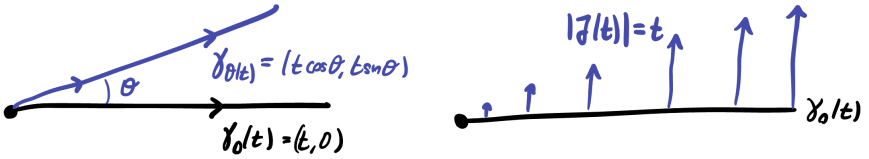


Figure 19.6.: A geodesic variation and its Jacobi field in Euclidean space. $|J(t)| = t$ means that nearby geodesics spread out linearly over time.

But what about on the sphere? Here, we may take our geodesic to be a line of longitude, say

$$\gamma(t) = (\sin t, 0, \cos t)$$

which passes through the north pole $N = (0, 0, 1)$ at time zero, with initial direction

$$\gamma'(0) = \left.\frac{d}{dt}\right|_{t=0} (\sin t, 0, \cos t) = \langle 1, 0, 0\rangle$$

How do we find the geodesics through N making angle θ with γ? By using isometries of course! We can *rotate the sphere* fixing N by angle θ using previous work:

$$\gamma_\theta(t) = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \sin t \\ 0 \\ \cos t \end{pmatrix} = \begin{pmatrix} \cos\theta\sin t \\ \sin\theta\sin t \\ \cos t \end{pmatrix}$$

Now we have our geodesic variation, so all we need to do is differentiate it with respect to θ.

$$\begin{aligned} J(t) &= \left.\frac{d}{d\theta}\right|_{\theta=0} \gamma_\theta(t) \\ &= \left.\frac{d}{d\theta}\right|_{\theta=0} (\cos\theta \sin t, \sin\theta \sin t, \cos t) \\ &= (-\sin\theta \sin t, \cos\theta \sin t, 0)\Big|_{\theta=0} \\ &= (0, \sin t, 0) \end{aligned}$$

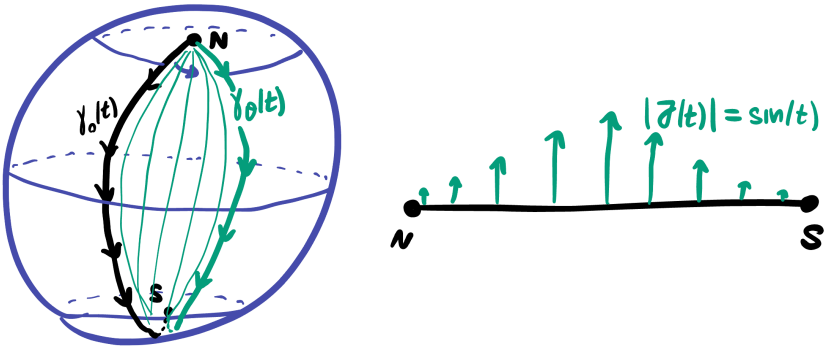


Figure 19.7.: A geodesic variation on ^{2} and its Jacobi field. The fact that $|J(t)| = \sin(t)$ tells us that nearby geodesics that start spreading out eventually come back together, after distance $t = \pi$.

Again, the magnitude of this vector measures the rate at which nearby geodesics are diverging from one another:

$$\|J(t)\| = \|(0, \sin t, 0)\| = \sin t$$

How do we interpret this? Well the sine function first grows until $t = \tau/4$ and then begins to shrink: this means that geodesics *first begin to diverge* then at $t = \tau/4$ begin to converge once more. Of course, we already knew this - because the geodesics are great circles (and we had found their explicit formulas to even compute the variation).

We've seen the qualitative behavior of these variations depends on the curvature: if the curvature is zero, then geodesics spread out linearly, but when its positive

they oscillate sinusoidally between converging and diverging. In fact :::{#thm-jacobi-equation} Let $J(t)$ be an infinitesimal variation along a geodesic. Then if the space we are considering has constant curvature κ, the infinitesimal variation satisfies the differential equation

$$J'' + \kappa J = 0$$

:::

19.4. Spheres of Other Sizes

So far, our entire discussion has been taking place for the unit sphere, but unlike Euclidean space, there are multiple *different* spherical geometries: things behaved differently depending on the radius of the sphere. For each positive real number R we can define *spherical geometry of radius R*, denoted ^2_R, as follows.

Definition 19.5 (Spherical geometry of Radius R.)**.** Let ^2_R denote the set of points which are distance R from the origin in $\mathbb{E}^3$. For each point $p \in {}^2_R$, the tangent space $T_p{}^2_R$ consists of all points in $\mathbb{E}^3$ which are orthogonal to p (definition unchanged from the unit sphere), and the dot product for measuring infinitesimal lengths and angles is the standard dot product on $\mathbb{E}^3$ (also unchanged from the unit sphere).

The development of each of these spherical geometries is qualitatively very similar to that for ^{2}: we can see without any change that $(x, y, z) \mapsto (x, y, -z)$ is an isometry so the equator is a geodesic, and orthogonal transformations are still isometries so all great circles are geodesics.

What changes is the quantitative details: the formulas for length area and curvature. In the next two problems, your job is to redo the calculations that I did for ^{2}, for the geometry ^2_R:

Exercise 19.4 (Circumference and area.)**.** What is the formula for the circumference and radius of a circle of radius r on ^2_R?

Hint: base your circles at $N = (0, 0, R)$ and look back at our arguments from class to see what must change, and what stays the same.

Exercise 19.5. Using the definition of curvature as a limiting ratio of circumferences (Definition 32.4), compute the curvature of ^2_R, and show that

$$\kappa = \frac{1}{R^2}$$

Think about what this relationship says: as a sphere gets bigger in radius, it's curvature (which measures the difference between it and the Euclidean plane) quickly decreases! That is, the bigger a sphere you have, the *more difficult* it is to tell apart from a plane at a point.

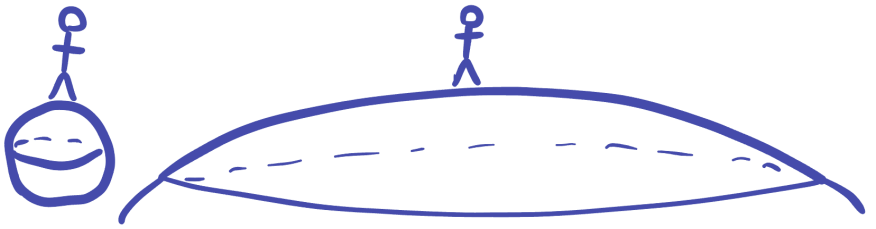


Figure 19.8.: Small spheres have large curvature, making it easy to see. Large spheres have small curvature, making it more difficult to notice.

This mathematical fact has tricked an unfortunate number of people into believing that large spheres like the earth *are flat*. But now we know better:

Example 19.1 (Curvature of the Earth)**.** The earth's circumference is 40 million meters (in fact, the meter was originally defined so that the distance from the equator to the north pole was 10 million meters!). This means the radius of the earth is

$$R = \frac{40,000,000}{2\pi} \approx 6,366,197\text{m}$$

and so the curvature of the earth is

$$\begin{aligned}\kappa &= \frac{1}{R^2} \\ &= \frac{1}{40,528,473,456,935} \\ &\approx 0.00000000000000024674011002723\end{aligned}$$

But while small, 0.00000000000002 is **not zero**, and on large enough scales this small amount of curvature actually has a big effect, on flight paths, air currents, the formation of hurricanes, etc.

20. Polygons

We've made great progress on understanding the sphere: we've discovered the geodesics, found enough isometries to do meaningful work (we can move any point to any other, and any geodesic to any other), analyzed circles and defined both acceleration and curvature. With these tools in hand we are ready to confront some of the most surprising differences between the geometry of the sphere and that of the plane, which are become visible when studying polygons.

An n-gon is a polygon with n vertices (or equivalently, with n edges). In the plane we studied 3−gons (triangles), 4-gons (quadrilaterals), and beyond, but we never mentioned 2-sided shapes. Why? Well 2-gons (or bigons, if you are feeling fancy) do not exist in $\mathbb{E}^2$!

This is because for a two sided shape to exist, its two sides would have to meet each other twice (once at each vertex). And we proved that Euclidean lines are given by affine equations, and such curves can only intersect once, lest they be equal (linear algebra!).

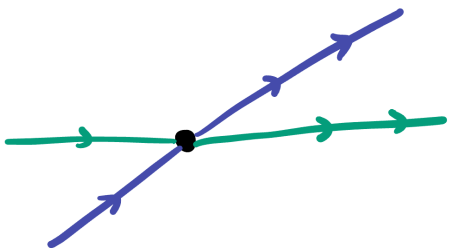


Figure 20.1.: Two lines in $\mathbb{E}^2$ intersect exactly once, if they intersect at all.

However, this basic behavior of lines is very different on the sphere.

Theorem 20.1 (Geodesics intersect twice)**.** *Any two distinct geodesics on* ^{2} *intersect exactly twice.*

Proof. Let C_1 and C_2 be two geodesics on ^{2}. Each is a great circle, and so is described by a plane passing through the origin. But two planes passing through the origin of $\mathbb{E}^3$ must intersect each other in a line! Thus, these two planes have an entire line

through the origin in common, and this line must intersect the sphere in two points. These two points are then intersections between C_1 and C_2.

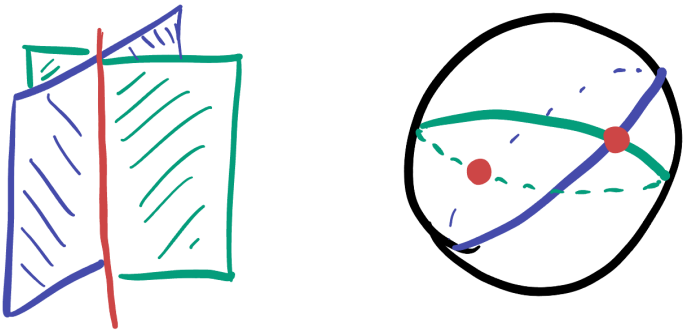


Figure 20.2.: Any two planes through the origin intersect in a line implies that any two geodesics on ^{2} intersect in exactly two points.

There cannot be any more intersections, as we can see also by thinking about the linear algebra of planes in $\mathbb{E}^3$: if our curves had three points of intersection on the sphere, at least one of them would not be a multiple of the others (as the only multiple of a point p which still lies on ^{2} is $-p$). Thus, we have three non-collinear points which lie on both geodesics. But, three points in $\mathbb{E}^3$ fully determine a plane, so if these points lie on both planes, the planes are equal, and so the geodesics themselves are equal (thus not distinct). □

Recall the definition of parallel - we said that two lines were *parallel* if they did not intersect! But all lines on ^{2} intersect: thus there are no parallel lines at all!

Corollary 20.1. *There are no parallel lines on the sphere. Thus, Playfair's axiom is false for* ^{2}.

Playfairs axiom (which stated that given any line, and any point not on that line) is equivalent to Euclid's 5th postulate, assuming the first four. We do not have this equivalence available to us here (because not all of the first four are true in spherical geometry!) so we have to separately ask about the5^{th}:

Exercise 20.1 (Euclid's Fifth Postulate is False on ^{2})**.** Show that Euclids postulate is false by finding a counterexample: give two geodesics on the sphere that intersect a third in angles which sum to π, but nonetheless intersect.

Remark 20.1. Its actually hard to *precisely* make sense of Euclid's postulate on the sphere, in his original wording, as it talks about finding an intersection on *one side* of the crossing line or the other. But on the sphere there are no sides: everything meets up on the back!

But besides answering these interesting foundational questions, realizing that pairs of geodesics intersect each other twice has another important corollary:

Corollary 20.2 (Bigons Exist)**.** *Bigons exist in spherical geometry.*

Thus, we begin our study of polygons not with triangles as we did in Euclidean space, but at an even lower, more basic level: we begin with bigons!

20.1. Bigons

A bigon has two angles, and two sides. At first, we know nothing else about them, so we might give a different name to each side and to each angle, like so:

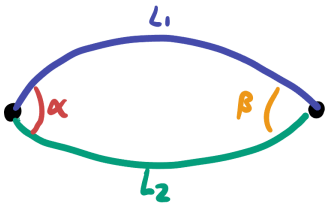


Figure 20.3.: A bigon is a two-sided polygon.

Just like in trigonometry, our goal here is to try to discover relations between the sides and angles of a bigon. However, unlike trigonometry - the relations here turn out to be very simple: there just aren't many ways to make a bigon!

Proposition 20.1. *Both sides of a bigon have length π:*

Proof. We saw in Theorem 20.1 that if one vertex of a bigon is p, then its sides, being geodesics, meet again at the point antipodal to p. Thus, each side of the bigon is exactly half of a great circle, and so has length $\frac{1}{2}2\pi = \pi$. □

Next, we should ask about the bigon's angles: is it possible to have a bigon with two angles of different measures?

Proposition 20.2. *The two angles of a bigon are equal to one another.*

Proof. To make things easier to picture, we can use an isometry to move one of the vertices of our bigon to the north pole (and thus the other to the south pole, since they are antipodal).

Now, the sides of the bigon make a right angle with the equator (since they are great circles going from the north to south pole) and so reflection in the equator sends the bigon to itself, exchanging its two vertices.

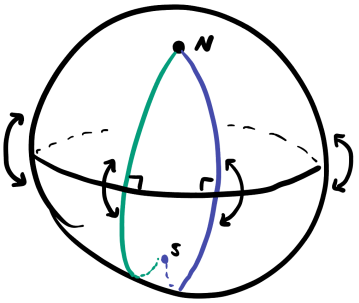


Figure 20.4.: Reflecting a bigon about the equator exchanges its two vertices, showing the two angles must be the same.

But now we are done! Isometries preserve angle, and so if there's an isometry that swaps the vertices of the bigon they must have the same angle. □

Thus bigons only have one free parameter: once you know the angle a bigon has at one vertex, you know everything there is to know to construct the entire bigon.

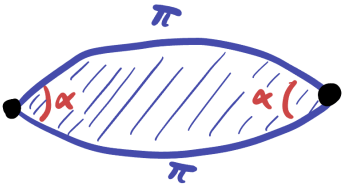


Figure 20.5.: Bigons are determined by their angle measure.

Indeed - up to isometry there is exactly one bigon for every angle $\theta \in (0, 2\pi)$, where the bigon with angle π is exactly one hemisphere of the sphere, and a bigon with angle $> \pi$ covers more than half the sphere. Strange world spherical geometry is, where a polygon can have only two sides and take up more than half of the universe!

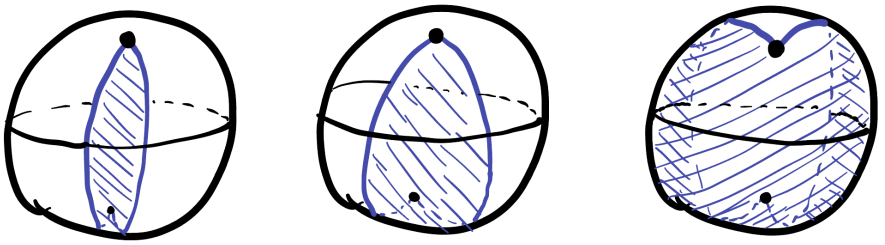


Figure 20.6.: Bigons of small medium and large angle size on ^{2}.

The final geometric quantity we may wish to understand is the *area* of a bigon.

20.2. Triangles

Having discovered literally everything there is to know about bigons, its time to move on to the world of triangles. First, we should be careful and check that triangles even exist! This might sound silly - but our recent experience with bigons should warn us to be extra careful.

Proposition 20.3 (Spherical Triangles Exist)**.** *Any three points not all lying on the same great circle determine a triangle.*

Proof. Let p q and r be any three points on ^{2} not on the same great circle. Draw the shorter geodesic segment connecting p and q (recall, there are two of these, as p and q both lie on a great circle: if p and q are antipodes then choose either segment).

Likewise, draw the shorter segments connecting p to r and q to r. All we need to do to show this forms a triangle is to argue that these two new segments do not cross the first segment.

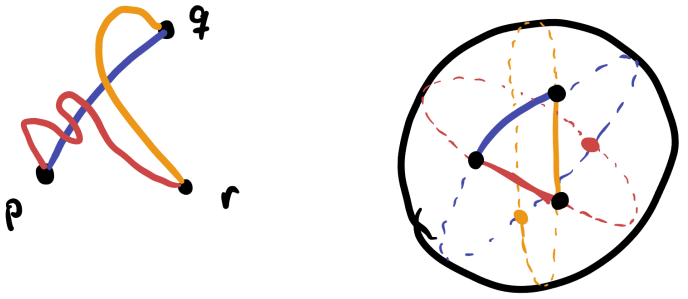


Figure 20.7.: Left: the case we want to show does not happen, where the geodesics intersect more and do not form a triangle. Right: what actually happens, because we know geodesics are great circles.

Of course, they *do* intersect the first segment at its endpoints p and q! But they can't intersect it anywhere else - the entire great circles they define intersect the great circle containing p and q only at these points, and then at their antipodes $-p$ and $-q$.

But, since the segment connecting p and q was the *shorter* of the two geodesic segments, its not long enough to include both p as one of its endpoints, and $-p$ as a point in the interior: then it would stretch more than half way around the sphere! Thus, these other intersections are not on the segment, and the three segments meet only at their vertices, forming a triangle. □

Remark 20.2. Think for a moment from the perspective of an inhabitant of spherical geometry, who are so accustomed to dealing with bigons that when they start trying to understand Euclidean geometry, they don't think twice and just immediately start their theory by investigating bigons. Whatever theorems they prove would be *useless* because they all implicitly are of the form *the existence of bigons implies XXX* and the premise is false: bigons do not in fact exist at all!

Now that we are confident in their existence, we turn to the most surprising - and at the same time the most useful - property of spherical triangles: their area is intimately tied to their angle sum.

Theorem 20.2 (Area of a Spherical Triangle)**.** *The area of a spherical triangle is equal to the angle sum, minus π. In symbols, if a triangle T has angles of measure α, β and γ, then*

$$\text{area}(T) = \alpha + \beta + \gamma - \pi$$

The style of proof here is quite clever, and uses our work with bigons! Indeed, we will cover the sphere with six bigons starting from our triangle, and find the triangle's area by counting area overlaps.

Proof.

□

Exercise 20.2. What is the analogous formula for the area of a triangle on the sphere of curvature κ?

Hint: recall that the map taking the unit sphere to the sphere of radius R is a similarity of $\mathbb{E}^3$, and use what we know about similarities effect on area and angles to deduce this directly from the unit sphere case, without repeating the proof given in class.

This formula has some immediate and surprising consequences. Since polygons have nonzero (and positive!) area, we can use as a powerful tool in proving the *nonexistence* of various objects on the sphere. The strategy goes:

- Assume for contradiction a certain object exists
- break it into triangles
- compute the area of the shape, using these triangles
- find the area is zero or negative: contradiction!
- Thus, the object does not exist.

20.3. Quadrilaterals

Theorem 20.3 (Area of Spherical Quadrilaterals)**.** *The area of a convex spherical quadrilateral is equal to its angle sum minus 2π.*

Proof. Let Q be a quadrilateral on ^{2}, with angles $\alpha, \beta, \gamma, \delta$. Choose two opposite vertices of Q, and draw the line segment connecting them. This segment lies fully inside the quadrilateral (by convexity), and divides it into two triangles T_1 and T_2, dividing the angles $\alpha = \alpha_1 + \alpha_2$ and $\gamma = \gamma_1 + \gamma_2$ between them:

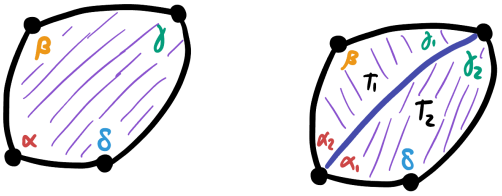


Figure 20.8.: Determining the area of a quadrilateral by decomposing it into triangles.

Now we can compute the area as the sum of the area of the triangles:

$$\begin{aligned}\text{area}(Q) &= \text{area}(T_1) + \text{area}(T_2)\\ &= (\alpha_1 + \beta + \gamma_1 - \pi) + (\gamma_2 + \delta + \alpha_2 - \pi)\\ &= (\alpha_1 + \alpha_2) + \beta + (\gamma_1 + \gamma_2) + \delta - 2\pi\\ &= \alpha + \beta + \gamma + \delta - 2\pi\end{aligned}$$

□

This has some immediate surprising consequences, including the nonexistence of rectangles!

Corollary 20.3 (Rectangles do not exist.)**.**

Proof. Assume that there is a quadrilateral R on ^{2} with four right angles. Then by the above, we can compute the area

$$\text{area}(R) = \left(\frac{\pi}{2} + \frac{\pi}{2} + \frac{\pi}{2} + \frac{\pi}{2}\right) - 2\pi = 0$$

But this is impossible, quadrilaterals cannot have zero area! Thus, we must have been wrong, and a right angled quadrilateral cannot in fact exist. □

Exercise 20.3. What is the analogous formula for the area of a quadrilateral on a sphere of curvature κ?

20.4. Platonic Solids

Besides proving nonexistence results like that for rectangles, the triangle area formula helps us determine what regular polygons can be used to tile the sphere.

Recall we call a polygon *regular* if it has rotational symmetries about its center: in particular this implies that all its sides are the same length, and all its angles have the same measure (since isometries preserve both lengths and angles).

In the Euclidean plane, we know that regular polygons of all side numbers ≥ 3 exist (these are how Archimedes approximated the circle, after all!), but their angles are strictly determined by their number of sides. We proved in a previous homework that the angle sum of an n-gon is $(n-2)\pi$, and if all the angles of a regular n gon are equal, each angle must measure $\theta_n = \frac{n-2}{n}\pi$.

This puts a strong restriction on which regular polygons can be used to tile the plane. To tile the plane, a necessary (but not sufficient) condition is that we need to be able

to fit k copies of each polygon around a vertex, without any gaps or overalps. This tells us that the angles of a polygon that can tile must be $\theta = \frac{2\pi}{k}$.

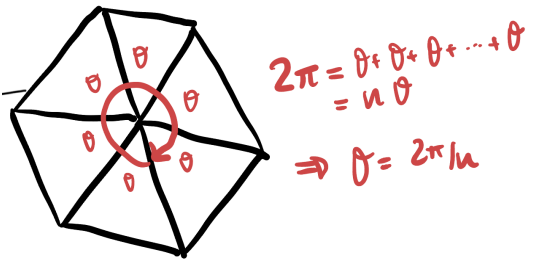


Figure 20.9.: Angles need to be an integer divisor of 2π to fit evenly around a point without gaps or overlap.

Thus, to figure out which polygons even have a chance of tiling the Euclidean plane, we want to know for which n (the number of sides) there the angle θ_n is actually 2π over an integer. We can start listing:

$$\theta_3 = \frac{3-2}{3}\pi = \frac{\pi}{3} = \frac{2\pi}{6}$$

$$\theta_4 = \frac{4-2}{4}\pi = \frac{\pi}{2} = \frac{2\pi}{4}$$

$$\theta_5 = \frac{5-2}{5}\pi = \frac{3\pi}{5}$$

$$\theta_6 = \frac{6-2}{6}\pi = \frac{2\pi}{3}$$

$$\theta_7 = \frac{7-2}{7}\pi = \frac{5\pi}{7}$$

Thus, we see that its possible to fit six triangles around a vertex, four squares around a vertex and three hexagons around a vertex, but as the angles θ_5 and θ_7 aren't even divisions of 2π, there's no nice way to fit pentagons or 7-gons around a vertex, and thus no hope of using them to tile the plane.

This is the start to the classification of regular tilings of the plane, where by what we see from the angle measures, its possible for triangles, squares and hexagons, but impossible for all other shapes!

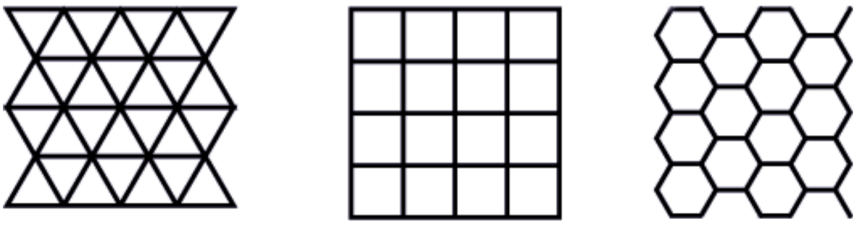


Figure 20.10.: The three regular polygons that tile the Euclidean plane.

Our goal here is to investigate what changes on the sphere.

Exercise 20.4 (Spherical Pentagons)**.**

- Find a relationship between the area A of a spherical regular pentagon and its angle measure α. *Hint: divide the spherical pentagon into five triangles*
- Show that there exists a spherical pentagon whose angle evenly divides 2π: how many of these spherical pentagons fit around a single vertex?
- What is the area of such a spherical pentagon? How many of these pentagons does it take to cover the entire sphere?

The resulting tiling of the sphere is the dodecahedron - one of the Platonic solids discovered by the Greeks (though, usually these are imagined as having flat faces, instead of actually lying directly on the surface of the sphere). This is pretty encouraging, our simple investigation into areas of triangles led us all the way to the dodecahedron! But can it go farther? Can we learn exactly *which* polygons can tile the sphere from such meager data?

Exercise 20.5 (No Tiling by Hexagons)**.** Show that there is no regular hexagon which can tile the sphere.

And, it only gets worse from here:

Exercise 20.6. Prove that for any $n \geq 7$, there are no regular spherical n gons that can tile the sphere.

The problem we run into with hexagons is that their area must be *zero*, and its worth commenting briefly on what that means. Having zero area means the angle sum needed is *equal* to the Euclidean angle sum - and so this is just telling us that the sphere is the wrong spot to be looking for such a tiling; instead it exists in the Euclidean plane!

But what are we learning in the case of 7-gons and above? If we try to find any value of k where the angles would be $2\pi/k$, we get a *negative area*: this means the shapes both don't exist on the sphere and don't exist in Euclidean space. However, we will meet these tilings shortly, in *hyperbolic space*

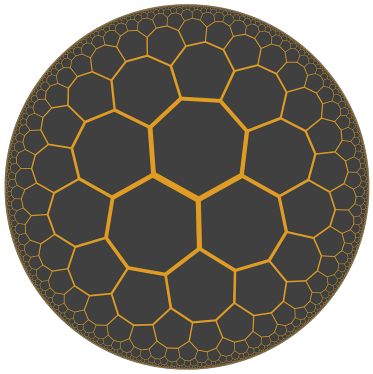


Figure 20.11.: A tiling of the hyperbolic plane by heptagons.

So, we've found a pentagon that tiles ^{2}, and no higher n-gons do. This leaves only three cases left to investigate: the bigons, triangles, and quadrilaterals!

For bigons, the condition that they tile the sphere is just that their angles are $2\pi/n$: this is possible for every $n \geq 2$, so we have an infinite collection of different bigon tilings:

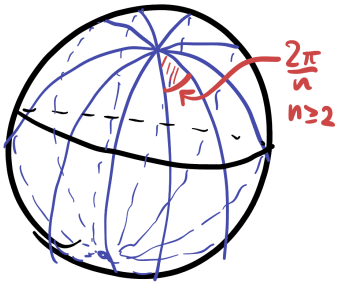


Figure 20.12.: Bigons of angle $2\pi/n$ tile the sphere.

But, these aren't really that interesting: they're just what you get by drawing an n-gon on the equator, and then extending perpendicular geodesics up to the north and

south poles. Indeed, these are so simple that these tilings are often not even counted among the platonic solids!

The more interesting shapes appear when $n = 3$ and $n = 4$.

Exercise 20.7. Prove that there is exactly one quadrilateral that can tile the sphere. How many fit around each corner? How many quadrilaterals does it take to cover the sphere?

Which platonic solid does this correspond to?

Exercise 20.8. There are three different equilateral triangles that can be used to tile the sphere. Find them! For each triangle:

- How many fit around each vertex?
- How many are needed to cover the sphere?
- What platonic solid does this correspond to?

20.5. Trigonometry

We've already gotten an incredible amount of information out of just knowing how to relate *angles* to *area* of spherical triangles. But there is much more to be gained from studying the quantitative relationships between *angles* and *lengths* as well. This is the study of spherical trigonometry!

We will not dive too deeply into this material in this course, as it is a huge topic dating all the way back to the greeks, and navigation by the stars! Instead we aim to just give a taste.

As in Euclidean space, its easiest to start with as simple of triangles as possible. In $\mathbb{E}^2$ these were right triangles, as having a right angle makes a lot of things easier. In the sphere - we can do one better: well, really *two better* - there are triangles which have *three right angles!*

To see these exist - you can make one by starting with a right angle at the north pole, and following both geodesics down to the equator, then stopping and using the segment of the equator connecting the endpoints as the third side. The top angle was right by construction, and these next two are also right angles, as they are the intersection of the equator with geodesics through the north pole (as you showed on your last homework assignment).

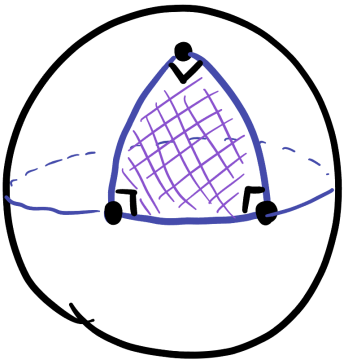


Figure 20.13.: A triangle with three right angles on the sphere.

However, there is not a very interesting theory of the trigonometry of three-right angled triangles: it turns out that up to isometry, this example above is the only one.

Proposition 20.4. *All triply right triangles on* ^{2} *are isometric to one another.*

Proof.

□

Thus we completely understand these right triangles: they all have angles $\pi/2$ (of course!), but they *also* have side lengths $\pi/2$, and they have area $\$A = 3\frac{\pi}{2} - \pi = \frac{\pi}{2}$: every geometric measurement here is equal to $\pi/2$!

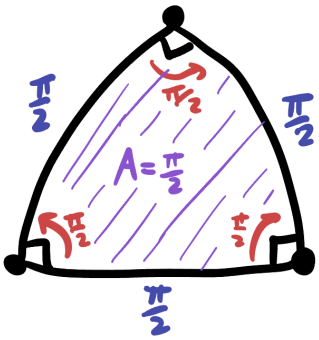


Figure 20.14.: All measurements of a triply right triangle are $\pi/2$.

The next simplest case is that of doubly-right triangles. Let's call the third angle of such a triangle α. These are also quite restricted: take the side opposite α which contains the two right angles, and move it to a segment of the equator by isometries. Now, the other two sides are geodesics which make right angles with the equator: they intersect at the north pole! So, our triangle has two sides of length $\pi/2$.

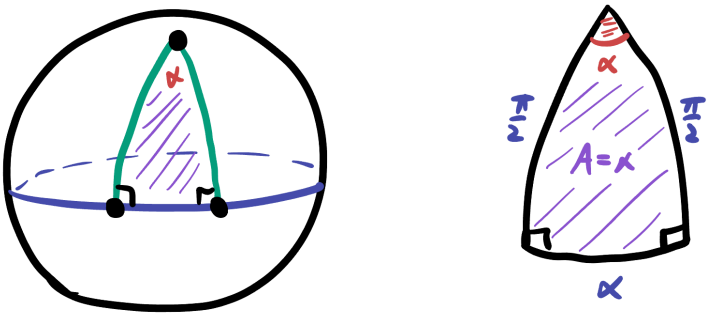


Figure 20.15.: The trigonometry of a doubly-right triangle.

Now that we know this, the area is immediate: this is half of a bigon with angle α, so its area is $\frac{1}{2}(2\alpha) = \alpha$. We can also quickly determine the third side length: the angle at N is α, and so the arclength along the equator (which is a unit circle) is also α.

20.5.1. Right Triangles

Things get both more interesting, and more complicated in the case of triangles with a single right angle. The fundamental trigonometric relationship for a right triangle is how the lenght of its hypotenuse depends on the lengths of its legs. In Euclidean space, this is the famous *Pythagorean theroem*, but in spherical geometry it takes on another form.

Theorem 20.4 (Spherical Pythagorean Theorem)**.** *Given a right triangle on* ^{2} *with side lengths* a, b *and hypotenuse* c*, these three lengths satisfy the equation*

$$\cos(c) = \cos(a)\cos(b)$$

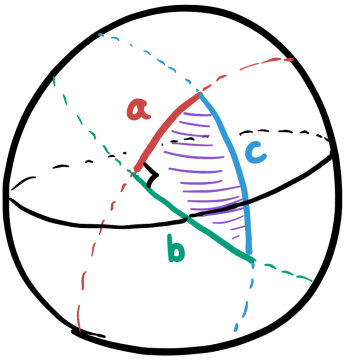


Figure 20.16.: Right triangles on ^{2} have their own analog of the pythagorean theorem, equating the cosine of the hypotenuse to the product of the cosines of the other two sides.

Exercise 20.9 (Deriving The Pythagorean Theorem)**.** Prove that the formula given above really does hold for the legs and hypotenuse of a right triangle on ^{2}, using the distance formula that we've already calculated:

$$\cos \operatorname{dist}(p, q) = p \cdot q$$

*Hint: move your triangle so the right angle is at the north pole, and the legs are along the great circles on the xz and yz plane. Now you can write down exactly what the other two vertices are since you know they are distance a and b along these geodesics from N*K

On a sphere of radius R, a similar formula exists: here to be able to use arguments involving angles we need to divide all the distances by the sphere's radius, but afterwards an argument analogous to the above exercise yields

$$\cos\left(\frac{c}{R}\right) = \cos\left(\frac{a}{R}\right)\cos\left(\frac{b}{R}\right)$$

Its often more useful to rewrite this result in terms of the curvature $\kappa = 1/R^2$

Theorem 20.5 (Pythagorean Theroem of Curvature κ)**.** *On the sphere of curvature κ, the two legs a, b and the hypotenuse c of a right triangle satisfy*

$$\cos\left(c\sqrt{\kappa}\right) = \cos\left(a\sqrt{\kappa}\right)\cos\left(b\sqrt{\kappa}\right)$$

As a sphere gets larger and larger in radius, it better approximates the Euclidean plane. We might even want to say something like in the limit $R \to \infty$ (so, $\kappa \to 0$) the spherical geometry *becomes euclidean.* But how could we make such a statement

precise? One way is to study what happens to the theorems of spherical geometry as $\kappa \to 0$; and show that they become their Euclidean counterparts. The exercise below is our first encounter with this big idea:

Exercise 20.10 (Euclidean Geometry as the Limit of Shrinking Curvature)**.** Consider a triangle with side lengths a, b, c in spherical geometry of curvature κ. As $\kappa \to 0$, the arguments of the cosines in the Pythagorean theorem become very small numbers, so it makes sense to approximate approximate these with the first terms of their Taylor series.

Compute the Taylor series of both sides of

$$\cos\left(c\sqrt{\kappa}\right) = \cos\left(a\sqrt{\kappa}\right)\cos\left(b\sqrt{\kappa}\right)$$

in the limit $\kappa \to 0$, we can ignore all but the first nontrivial terms. Show here that only keeping up to the quadratic terms on each side recovers the Euclidean Pythagorean theroem, $c^2 = a^2 + b^2$.

Like in the plane, we might next hope to discover relationships between the sides of a spherical right triangle and its angle measures. And, indeed we can!

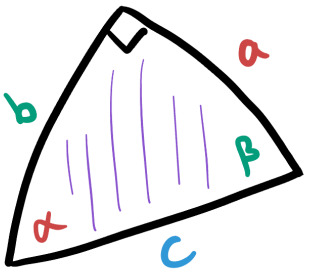


Figure 20.17.: A right triangle with angles α, β and opposite sides a, b.

The corresponding laws of spherical trigonometry are as follows:

Theorem 20.6 (Spherical Trigonometric Relations)**.** *For a right triangle with angles α, β, corresponding opposite sides a, b and hypotenuse c the following relations hold:*

$$\sin\alpha = \frac{\sin a}{\sin c} \qquad \sin\beta = \frac{\sin b}{\sin c}$$

$$\cos\alpha = \frac{\tan b}{\tan c} \qquad \cos\beta = \frac{\tan a}{\tan c}$$

Its instructive to compare these to their Euclidean counterparts: where the $\sin\alpha = a/c$ and $\cos\alpha = b/c$ for instance. The spherical versions have the same ratios, but the lengths are showing up *inside trigonometric functions themselves*!

These can be derived (though we will not, for the sake of time) using the geometry of planes in $\mathbb{E}^3$ - since great circles on the sphere are just intersections of planes through the origin with the sphere.

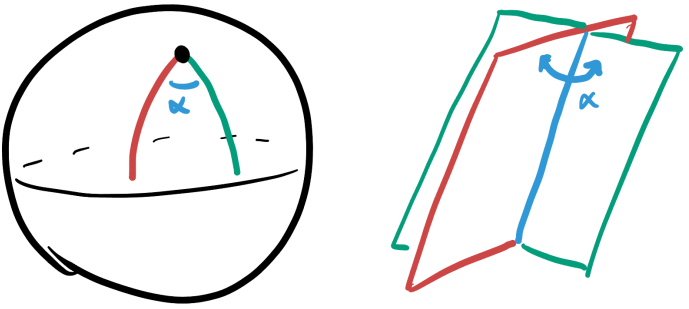


Figure 20.18.: The angles of a spherical triangle are angles between planes in $\mathbb{E}^3$, which lets us use Euclidean trigonometry to derive spherical trigonometric relationships.

Here's a nice derivation, which finds the angles between planes (and thus the angles between great circles) by finding the angles between their normal vectors.

One of the most biggest differences between spherical trigonometry and its Euclidean counterpart is that its possible to derive formulas for the length of a triangles' sides *in terms of only the angle information*! This is impossible in Euclidean space because of the existence of similarities: there are plenty of pairs of triangles that have all the same angles but wildly different side lengths! No so in the geometry of the sphere.

Exercise 20.11. Using the trigonometric identities in Theorem 32.4 together with the spherical pythagorean theorem Theorem 32.2, show that the side length a of a right triangle can be computed knowing only the opposite angle α and the adjacent angle β as

$$\cos a = \frac{\cos\alpha}{\sin\beta}$$

Hint: start with the formula for $\cos\alpha$*. Write out the tangents in terms of sines and cosines, then apply the pythagorean theorem to expand a term. Finally, use the definition of* $\sin\beta$ *to regroup some terms.*

Formulas such as this are incredibly useful for calculating the side lengths of polygons, by dividing them into triangles and using facts that are known about their angles.

Exercise 20.12 (Spherical Trigonometry)**.** Use spherical trigonometry to figure out the side lengths of the pentagon you discovered in the first exercise.

Hint: can you further divide the five triangles you used before, into ten right triangles inside the pentagon?

Part V.

Maps

21. Cartography

The sphere is a 2-dimensional object (its *inside* is three dimensional, but remember the geometry that concerns us is only the surface), but so far we have been studying it using the *three coordinates* of Euclidean 3-space in which it lives. This is sometimes incredibly convenient - it let us describe the geodesics of the sphere in a simple way as intersections of ^{2} with planes, for example. But in other respects it causes extra complication: functions now have three variables, meaning their derivatives are 3×3 matrices (containing 9 numbers), when they we should only require 2×2 matrices (four numbers, less than half!) if we could find a way to really work with the sphere intrinsically as a 2D object.

While the number of variables alone certainly increases complexity, worse computational woes are caused by the fact that the cartesian x, y, z of $\mathbb{E}^3$ just aren't a good fit for the sphere they are well adapted to describe straight objects like lines and planes, but the sphere bends in all three directions, making it so that even using symmetry we can't move things around until something turns into a 1-dimensional problem (the best we can do is move something to a great circle that lies in a *coordinate plane*, like the equator, which sets one variable to zero).

This makes certain computations prohibitively difficult: you may recall that so far out of the three properties we claimed equivalently define geodesics, (length-minimizing, zero acceleration, and fixed by symmetries), we have actually only proven the latter two to be equivalent. The reason is simply that dealing with length integrals of the form

$$\int_a^b \sqrt{x'(t)^2 + y'(t)^2 + z'(t)^2}\, dt$$

lead to some pretty long calculations, and massively increase the complexity of actually doing a the calculation, leading to some pretty unenlightening pages of integrals.

But the good news is, there's another perspective on spherical geometry which addresses these shortcomings of the $\mathbb{E}^3$-based-approach.

- Its naturally intrinsically 2-dimensional, dealing only with 2D vectors and 2×2 matrices.
- It allows (some) geodesics to be described in a simple way, making various length integrals more tractable.

- It can be drawn on a chalkboard or in a notebook, so I don't have to lug a physical sphere up and down to the fourth floor for each class!

And the best feature of this new approach to geometry is that *we are already familiar with it* - humans have been using it to represent spheres since antiquity. It's the study of *maps*.

In this chapter we will look at some historical examples of maps, try to discover the underlying mathematical concepts that tie them all together, and then look at one (particularly simple) map as an example, to see how calculations are done.

21.1. Examples

The idea to try and accurately portray regions of the spherical earth on a portion of the Euclidean plane dates back to antiquity, and in the 2nd century CE Ptolemy wrote a book - *Geographica* containing a prescription of *coordinates* for map-making, and a map of the known world (the Mediterranean basin).

Figure 21.1.: Ptolemy's world map (a redrawing in the 15th century from the original 2nd century coordinates).

Over the intervening centuries many hundreds of different map-making styles have been created, with each map specifically designed to serve certain purposes best. Perhaps the most famous map is the Mercator projection, designed in 1569 by Gerardus Mercator (whose real name was Gerhard Kremer):

Figure 21.2.: The Mercator projection.

This map was originally designed to simplify navigation by ship: it has the very useful property that any *angle* you measure on the map accurately reflects the true angle value on the globe (more about this, later).

While this map became famous for preserving *angles*, it is also infamous for not accurately representing *areas*. In reality Africa is a gigantic continent (the United States approximately fits just into the Sahara Desert!), but here it appears to be about the same size as Greenland: an island which is actually *fourteen times smaller*. To appreciate the amount of distortion here, we can look at an overlay of each country with an accurately-sized version of itself.

Figure 21.3.: The Mercator projection versus the true size of countries.

To interact with such a graphic in real time, visit the website https://www.thetruesize.com/, which allows you to choose a country to drag around the mercator projection, and see its true size relative to other countries on the map!

Of course, its natural that there's distortion to the shape and size of regions on the map: we are trying to *flatten* the curved geometry of the sphere into a region of the plane! After we've developed a bit more mathematical material, we will prove a theorem to this effect. To help visualize the distortions of map, French mathematician Nicolas Auguste Tissot introduced the drawing of small disks on a map - representing the images of small uniformly sized circles on the earth. Such a shape is now known in map-making as a *Tissot's Indicatrix* (plural: *Tissot's Indicatrices*).

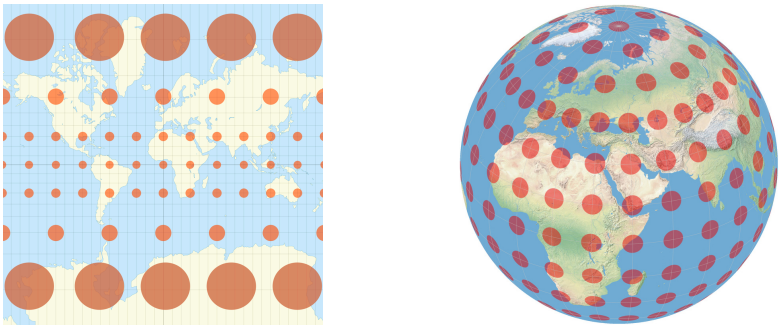


Figure 21.4.: Tissot's Indicatrices on the Mercator Map: each of these orange disks is the same size and shape on the globe.

The pairing of a map with Tissot's Indicatrices gives a helpful way to both visualize the earth while simultaneously being aware of the distortion incurred by this projection. We will adopt this as good style throughout the rest of the text, and anytime we draw a new map for the first time, we will accompany it with Tissots Indicratices (though once we start into the mathematics, we will leave this mapmaker terminology behind and start calling them *map disks*, or *infinitesimal disks*).

A natural question after seeing the Mercator map is, *can you find a map that doesn't mess up areas so much*? And indeed, you can! The map projection below is attributed to John Lambert in 1772 and is usually called the Lambert Cylindrical Projection, but the key idea traces back to Archimedes!

The Archimedes/Lambert map preserves the area of all regions, but it does so at its own cost: now *angles are distorted*! Recall that each of Tissots Indicatrices represents a small *perfect circle* on the globe - so the fact that these are being represented as thinner and thinner ovals near the poles means the map is stretching much more in one direction than in the other.

There are all sorts of area-preserving maps like this. The Archimedes/Lambert projection leaves the area near the equator relatively undistorted, but stretches horizontally near the poles. Instead, the Smyth projection manages to conserve area by stretching *vertically* inside the tropics, and *horizontally* outside:

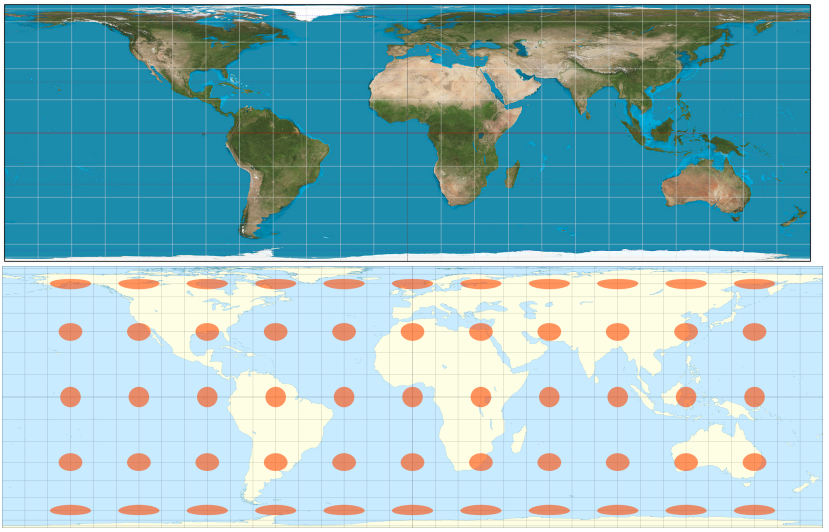


Figure 21.5.: The Archimedes/Lambert Cylindrical Projection map preserves area, but distorts *shape* and *angle* as seen by Tissots indicratices.

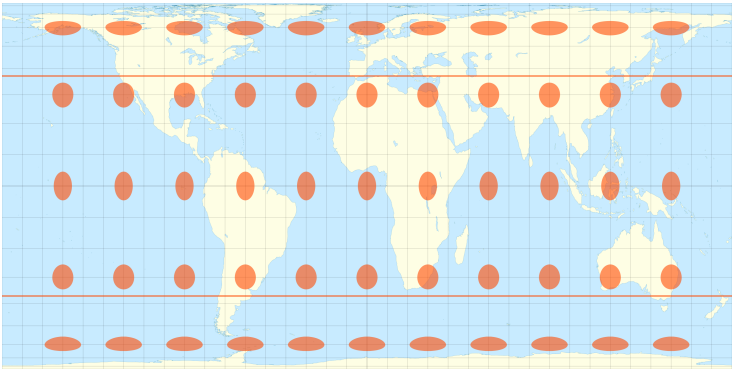


Figure 21.6.: The Smyth Projection is area preserving, stretching equatorial regions vertically and polar regions horizontally.

This vertical stretching makes a map wtih smaller *aspect ratio*, and this trend can be continued: the Tobler projection stretches most of the earth vertically and only the arctic circle horizontally, to make a map which is a perfect square:

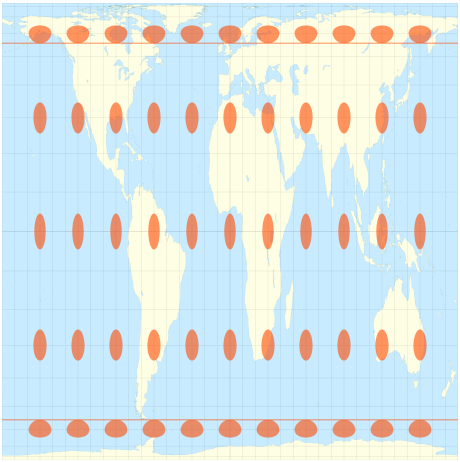


Figure 21.7.: The Tobler Projection

There are also other map styles which choose to preserve angles (like Mercator), such as the *Lambert Cone Projection*

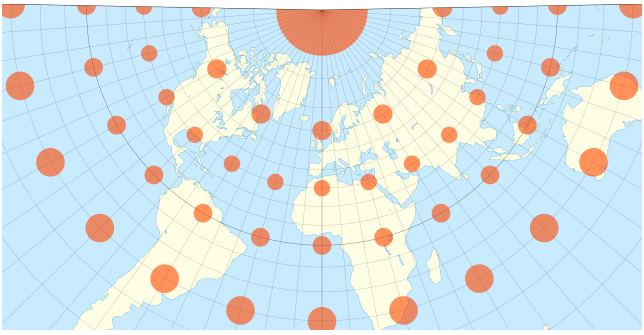


Figure 21.8.: The Lambert Cone Projection

or the *Pierce Projection* which does unequally display area, but attempts to arrange it so the big distortions happen out at sea, and do not affect the landmasses as much.

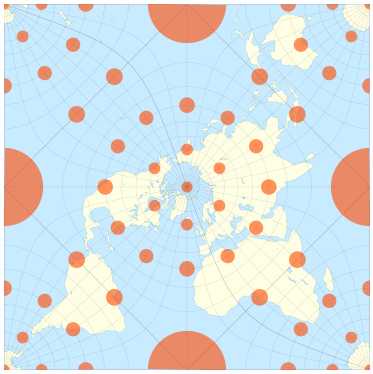


Figure 21.9.: The Pierce projection

Another map in this category is *stereographic projection*, which will be the most important map of all mathematically (see the future chapter by the same name). Having drawn Tissots Indicatrices, you can easily tell these maps scale *distances* non-uniformly over the globe, but at each point scale all distances by the same amount, as circles are sent to circles!

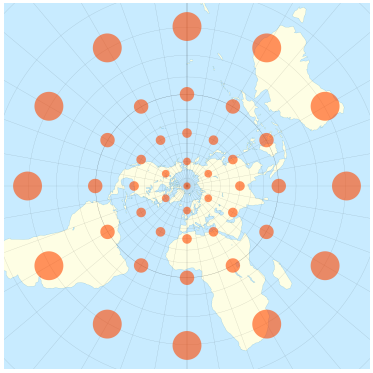
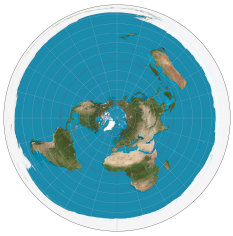


Figure 21.10.: Stereographic projection

Other maps may try to preserve certain *distances*, instead of lengths or angles. The *Azimuthal Equidistant projection* shows the correct distance from the north pole to any point on the map, while distorting both angles and areas. One particularly egregious distortion: the south pole is mapped to an entire ring (a circle of constant distance from the north pole), an unfortunate feature causing some people on the internet to take this map *literally*, and claim there is a great ice wall surrounding our disk shaped world.

So far, all the maps we've looked at have tried to depict the earth in some *reasonable shaped region* of the plane, and just live with the resulting distortion. But one can



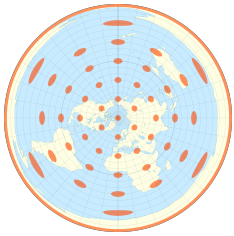


Figure 21.11.: The Azimuthal Equidistant projection: a Flat Earther's favorite map.

instead try to as best preserve angle and area as possible, and instead distort the *shape* of the map itself. Notable maps in this family include the tetrahedral projection,

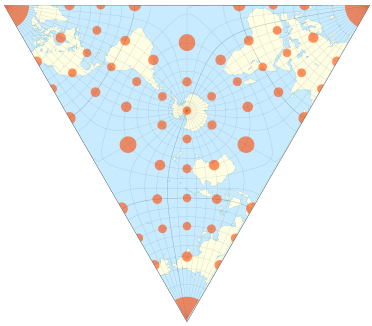


Figure 21.12.: The Tetrahedral projection. Fold the sides of the triangle upwards until the three corners meet at the top, to form a regular tetrahedron.

as well as the Waterman projection, based on a truncated octahedron

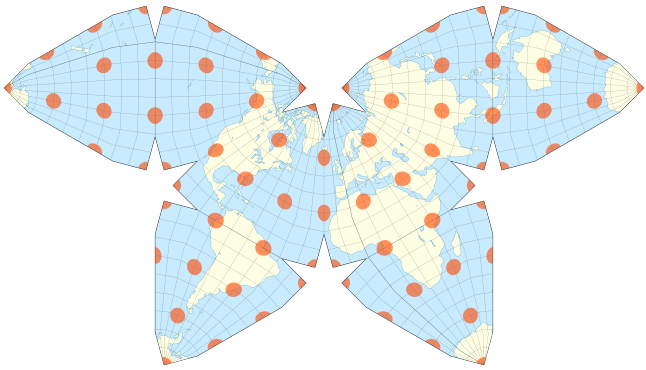


Figure 21.13.: Waterman Projection

or the even wilder Dymaxion projection

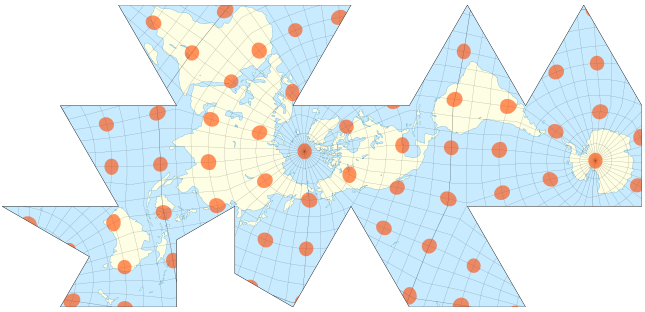


Figure 21.14.: Dymaxion Projection

21.2. Foundations

Alright - so we've now seen a bunch of maps, and spent some time thinking about how to interpret them. But how do we make this subject *mathematical*? To do mathematics we need definitions, and so the first thing we have to do is figure out abstractly, what is a *map*. What features are in common to all of the examples above?

One thing they all have in common is that they are all represented by some region in the Euclidean plane. The other fundamental commonality is that each point in that region represents some unique point of the sphere. Mathematically, pairings of points of one space uniquely with points of another space are modeled by a certain type of function: a *bijection*. So a map is a function!

Definition 21.1 (Map)**.** A map is a region $R \subset {}^2$ of the sphere, and a subset $M \subset \mathbb{E}^2$ of the Euclidean plane, together with a bijection $\phi : R \to M$. We call this map a *chart*, and we call the image of $M = \phi(R)$ the *map of R via the chart ϕ*.

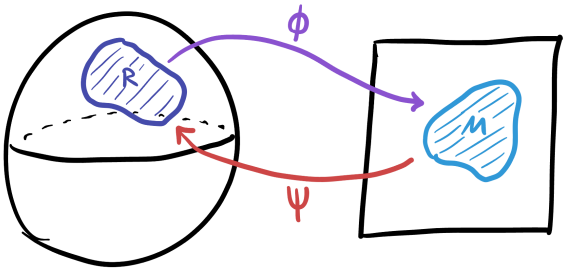


Figure 21.15.: A map M, its chart ϕ and parameterization ψ.

So, charts are functions that take pieces of the sphere and flatten them out onto pieces of the plane. But charts are (by definition) invertible, and so their inverse is a function which takes a region in the plane and presses it onto the sphere. Going this direction is also quite useful, so we'll give these a name, and call them *parameterizations*.

Definition 21.2 (Parameterization)**.** A parameterization of a region $R \subset {}^2$ is an invertible function $\psi : M \to {}^2$ from some map $M \subset \mathbb{E}^2$ onto $R \subset {}^2$.

The inverse of any chart is a parameterization, and the inverse of any parameterization is a chart. For this course, we will restrict our study to maps for which both the chart and parameterization are continuous and differentiable, so we can do geometry with infinitesimal vectors! In mathematics more generally, such maps are called *diffeomorphisms*.

Because the earth is a sphere and the plane is...not a sphere, its actually impossible to make a map which depicts every single point of the earth continuously. (This is readily believable, if you try to imagine continuously flattening a sphere onto the plane so no two points overlap, but its formal proof requires the subject of *topology*) Oftentimes, we are able to depict *most* of the earth (say, except for single lines or points where we cut it), and this causes no major issues. But if you insist on having a map of every point on the sphere, you need more than one map. Mathematicians call such a collection an *atlas*.

Definition 21.3 (Atlas)**.** An *atlas* is a collection of maps such that each point $p \in {}^2$ is in the domain of the chart of at least one of the maps.

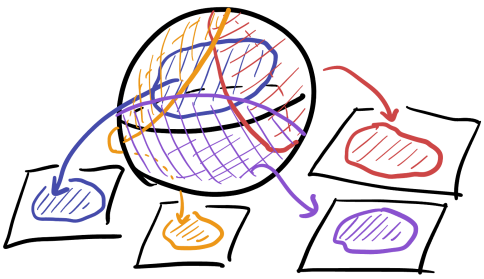


Figure 21.16.: An atlas of charts covers the entire sphere.

We will not have a need to consider atlases in this brief chapter, but they play a large role in the theoretical foundations of mathematical map-making: the subject known as *Riemannian geometry*.

Here, our goals are straightforward: given a map M with chart ϕ and parameterization ϕ, we want to be able to compute true things about the sphere, using *only the two dimensional map* M. To do so, we'll treat ϕ and ψ as *translation devices* taking us between the map and the actual sphere, and calculus to get quantitative results.

21.3. Getting Quantitative

The goal of a map is to *compute in M* but *learn about* [2]. We want to take a curve in M, and figure out how long the curve it represents on the sphere is. If two curves in in our map intersect (say, the path of two roads) we want to figure out what angle they intersect at on the sphere, while only measuring things using angles and vectors in M. And so on...

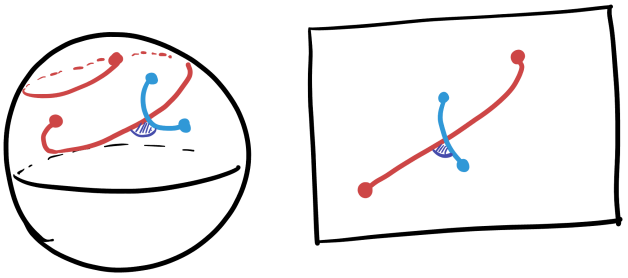


Figure 21.17.: True geometric quantities are on the sphere (left), but we want to compute them using a map (right).

21.3.1. Lengths

Probably unsurprisingly at this point in the course, the solution to all these problems is to *zoom in*, and use calculus to answer these things. This lets us replace the problem of curve lengths with *infinitesimal lengths*.

Proposition 21.1. *If $\gamma : [a,b] \to M$ is a curve drawn in a map M, then its map-length (the length of the curve it represents on the sphere) can be computed as*

$$\text{maplength}(\gamma) = \int_a^b \|D\psi_{\gamma(t)}\gamma'(t)\|dt$$

Proof. This is just a direct computation with the chain rule! If γ is a curve in M, then we can use the parameterization ψ to move it onto [2], so we get the true curve $\psi(\gamma(t))$ on the sphere. Now, we can compute the length of this - which is what we actually want!

$$\begin{aligned}\text{maplength}(\gamma) &:= \text{length}(\psi \circ \gamma) \\ &= \int_a^b \left\| (\psi(\gamma(t)))' \right\|, dt \\ &= \int_a^b \left\| D\psi_{\gamma(t)}\gamma'(t) \right\| dt\end{aligned}$$

□

Thus, the fundamental quantity we need to be able to compute is the *map-length* of an individual vector: given a vector $v \in T_pM$, find out how long of a vector it really represents on the sphere.

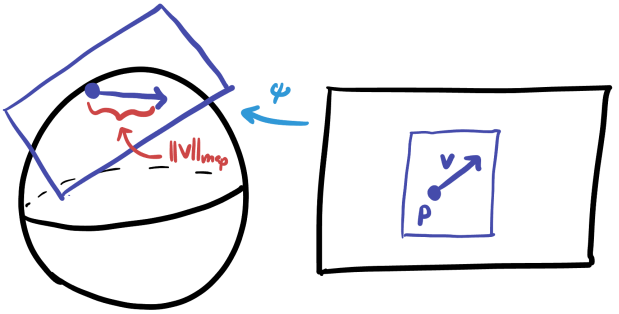


Figure 21.18.: The *map-length* of a vector is defined as the length of the vector it represents on the sphere.

Definition 21.4. Let M be a map of a region of the sphere with parameterization $\psi : M \to {}^2$. If $v \in T_pM$ then the *map-length* of v is given by

$$\|v\|_{\text{map}} = \|D\psi_p(v)\|_{\mathbb{E}^3}$$

Once we can compute infinitesimal lengths like this, not only have we solved the problem of finding the length of a curve on a map, but we can also draw the Tissot Indicatrices! Tissot imagined these as small disks at each point, and we can model them precisely as *infinitesimal disk* in the tangent space.

Definition 21.5 (Map Disk)**.** At each point $p \in M$, the *map disk* is the set of all tangent vectors v of maplength less than or equal to 1. We will denote this disk $\mathbb{D}^2_{\text{map}}(p)$:

$$\mathbb{D}^2_{\text{map}}(p) = \left\{ v \in T_pM \mid \|v\|_{\text{map}} \leq 1 \right\}$$

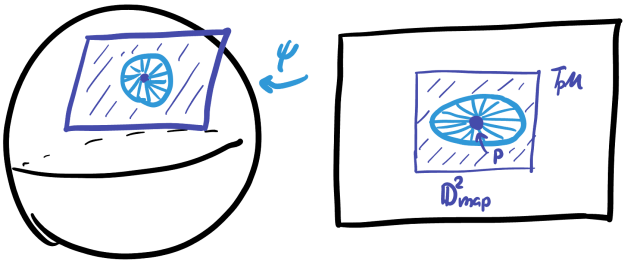


Figure 21.19.: The map-disk is the region in the tangent space T_pM which is mapped by the parameterization to a unit (infinitesimal) disk on the sphere.

21.3.2. Angles

We can similarly compute angles in a map, first use the parameterization to take us back to the sphere, and then measure the 'true' angle value there.

Proposition 21.2 (Map Angle)**.** *Let M be a map of some region of the sphere, with parameterization ψ. If v, w are two tangent vectors based at $p \in M$, their map angle is given by*

$$\theta_{\text{map}} = \arccos\left(\frac{(D\psi_p v)\cdot(D\psi_p w)}{\|D\psi_p v\|\|D\psi_p w\|}\right)$$

Proof. This is just the Euclidean formula for angles,

$$\theta = \arccos\left(\frac{V\cdot W}{\|V\|\|W\|}\right)$$

on the (EUclidean) tangent space to ^{2} in $\mathbb{E}^3$, applied to the vectors $V = D\psi_p v$ and $W = D\psi_p w$, which are the result of moving our vectors from the map to the sphere with ψ. □

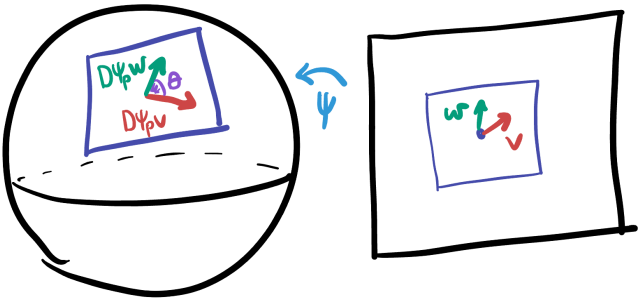


Figure 21.20.: The map-angle between two vectors in M is the angle between the vectors they represent on the sphere.

21.3.3. Areas

Finally, how do we calculate map area? Again - we can think infinitesimally, and ask what happens to *infinitesimal areas*, and then integrate the result. An infinitesmial aread dA in the Euclidean plane can be thought of as an infinitesimal unit square with sides $dx = \langle 1, 0\rangle \in T_pM$ and $dy = \langle 0, 1\rangle \in T_pM$ along the x and y directions respectively.

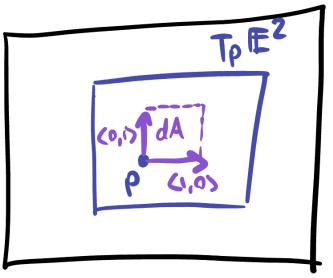


Figure 21.21.: Infinitesimal area on the plane is easiest to measure when using the orthonormal basis vectors in the x and y directions: in this case $dA_{\mathbb{E}^2} = dxdy$.

But such a unit square might not represent a unit area on the sphere! Think about the mercator projection: if this little square were near the north or south pole, mapping it onto the sphere would *shrink it a lot* (as projecting from the earth to the map *increases* size near the poles), so this square would actually represent a small area. (Indeed, we will see later that the area of an infinitesimal unit square in the Mercator projection at latitude θ is actually $\cos^2\theta$, which is very small near the poles $\theta = \pm\frac{\pi}{2}$)

To calculate the infinitesimal map area $(dA)_{\text{map}}$, we need to push the unit basis vectors $e_1 = \langle 1, 0\rangle$ and $e_2 = \langle 0, 1\rangle$ in T_pM to the sphere using ψ, and then measure the area that their images span in the tangent space $T_{\psi(p)}{}^2$.

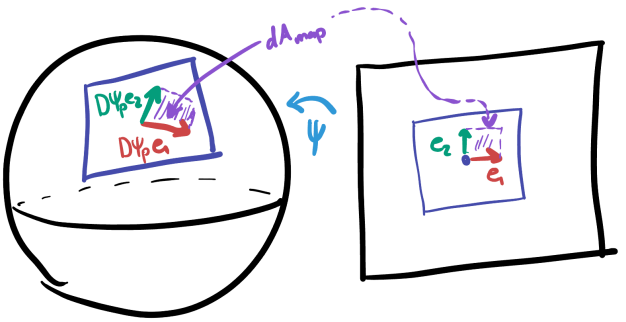


Figure 21.22.: The infinitesimal area dA_{map} on the plane is measured by taking an infinitesimal square, sending it to the sphere via the parameterization ϕ, and then measuring the area of the resulting parallelogram.

Such an area is rather annoying to measure in general, as the vectors $D\psi_p e_1$ and $D\psi_p e_2$ are both vectors with three coordinates. We know how to find the area spanned by two vectors in the plane, via the *determinant* (which we derived on homework way back in the calculus chapter!)

Definition 21.6. Let M be a map of some region of the sphere with parameterization ψ, and $e_1 = \langle 1, 0\rangle$, $e_2 = \langle 0, 1\rangle$ the unit basis vectors in $\mathbb{E}^2$. The map-area at a point $p \in T_pM$ is

$$(dA)_{\text{map}} = \|(D\psi_p e_1) \times (D\psi_p e_2)\| \, dxdy$$

Luckily, we will not often need to utilize the full generality of this definition. We will call a parameterization *rectangular* if it sends the x and y directions on M to pairs of orthogonal directions on ^{2}. For these maps, we can avoid the use of the cross product all together:

Proposition 21.3. *Let M be a map with rectangular parameterization ψ. Then the infinitesimal area is given by*

$$(dA)_{\text{map}} = \|D\psi_p e_1\| \|D\psi_p e_2\| \, dxdy$$

Proof. For maps with rectangular parameterizations, $D\psi_p e_1$ and $D\psi_p e_2$ are orthogonal to each other, and span a small rectangle in $T_{\psi(p)}{}^2$.

Because the area of a rectangle is hust base times height - we can simply transfer e_1 and e_2 to the sphere via ψ, measure their lengths, and take the product:

$$e_1 \mapsto D\psi_p e_1 \qquad e_2 \mapsto D\psi_p e_2$$
$$(dA)_{\text{map}} = \|D\psi_p e_1\| \|D\psi_p e_2\| \, dxdy$$

□

22. Examples

In this chapter we will apply some of the theory we developed to work with some well-known map projections used for depicting the earth. This is a slight digression from the logical flow of our text, as none of this work is strictly needed for anything that follows (and those interested in the purely mathematical story can move immediately to the next chapter, *stereographic projection*, where we apply these same techniques to the map we will use the most).

However, taking a brief look at examples serves two purposes: one, it will help us become more comfortable with the theory of maps, as we will do several explicit computations:

- We will calculate the map-length of a curve in the Orthographic projection.
- We will calculate the map-area of regions in Archimedes map, and show it is area preserving.
- We will compute angles in the Mercator map, and show it is angle preserving

And two; the study of maps is a beautiful application of mathematics to the wider human world - we might as well take a look - just for cultural reasons - while we are so nearby.

22.1. Orthographic Projection

To write down a map we need to give its chart: a map ϕ from some region in ^{2} onto a region of the plane. And perhaps the simplest formula taking points in $\mathbb{E}^3$ (where the sphere lives) to points of $\mathbb{E}^2$ is just deletion of a coordinate:

$$(x, y, z) \mapsto (x, y)$$

We can picture such a map as the *vertical orthogonal projection* of space onto the xy plane, and the result as the *shadow of an object under a vertical light source*

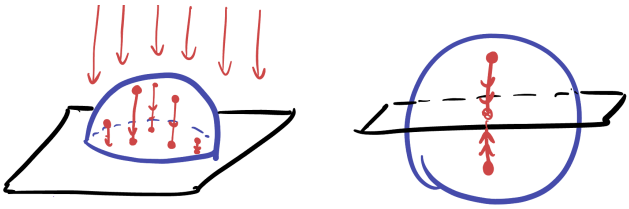


Figure 22.1.: Orthographic projection maps one hemisphere of the sphere onto the unit disk in the plane.

This cannot give a map of the entire earth at once, as each vertical line that intersects the sphere off the equator hits it in two points (x, y, z) and $(x, y, -z)$. However, if we restrict ourselves to one hemisphere, vertical projection *does* define a bijection between that hemisphere and the unit disk in the plane.

Figure 22.2.: Images of earth from far away in space are close to orthographic projections: like this one from the Discovr spacecraft 1 million miles away. True orthographic projection would be the view "from infinitely far away" where lines of sight would be parallel.

Definition 22.1 (Orthographic Map)**.** Let the region $R \subset {}^2$ be the northern hemisphere $R = \{(x, y, z) \in {}^2 \mid z \geq 0\}$, and M be the unit disk in the Euclidean plane $M = \{(x, y) \in \mathbb{E}^2 \mid x^2 + y^2 \leq 1\}$. Then the *orthographic projection* of R onto M is given by the chart ϕ and its inverse parameterization ψ,

$$\phi(x, y, z) = (x, y)$$

$$\psi(x, y) = \left(x, y, \sqrt{1 - x^2 - y^2}\right)$$

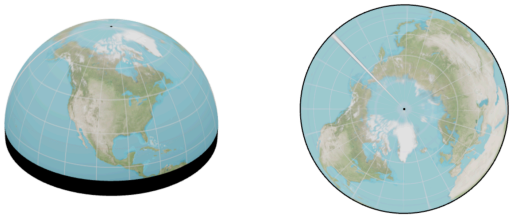


Figure 22.3.: The orthographic map projection of a hemisphere onto a disk.

Before delving into the quantitative calculus, let's try to develop a bit of a qualitative understanding of this map. Here's two facts we can see directly from its definition:

- Geodesics through the north pole N on ^{2} are mapped to straight lines through the origin O in the map.
- Circles about the north pole N are sent to Euclidean circles about the origin O in the map.

To see the first point, note that (1) geodesics on the sphere are great circles, which are the intersection of ^{2} with planes through the origin. Thus (2), geodesics containing the north pole N correspond to *vertical planes* (containing the z-axis), and so (3) the projection of a vertical plane onto the xy plane is just a line.

To see the second point, recall that the circle of radius r about N is described a Euclidean circle in the horizontal plane $z = \cos r$. The vertical projection deletes the z coordinate but leaves the x and y unchanged, so these circles map directly to circles in the xy plane.

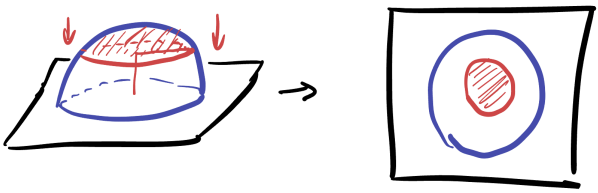


Figure 22.4.: Circles on the sphere about N are circles in horizontal Euclidean planes. These project horizontally to circles in the map.

This has the consequence that the *equator* of the sphere also maps to a circle on the plane - its the unit circle bounding our map M. So, in M we have some geodesics represented by straight lines, and one geodesic represented by a circle!

Remark 22.1. This implies that other geodesics of the sphere necessarily are represented by some curves that interpolate between straight lines and circles: not all geodesics are going to look like easy-to-understand curves in our map! That's one of the distortions we will have to learn to live with.

To get any quantitative understanding of this map, the first step is to take the derivative of the parameterization ψ:

$$\psi(x,y) = \begin{pmatrix} x \\ y \\ \sqrt{1-x^2-y^2} \end{pmatrix}$$

$$D\psi = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \frac{-x}{\sqrt{1-x^2-y^2}} & \frac{-y}{\sqrt{1-x^2-y^2}} \end{pmatrix}$$

Remark 22.2. When working with such a map, its often easiest to recall that $z = \sqrt{1-x^2-y^2}$ and just write 'z' anytime this expression occurs, to save mental space. Thus, we would write

$$D\psi = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \frac{-x}{z} & \frac{-y}{z} \end{pmatrix}$$

This derivative matrix is the key to all further calculation. From it we can directly compute map-lengths of vectors and curves, map-angles, and map-areas following the general theory.

Example 22.1 (Orthographic Map-Length of $\langle 1,0\rangle$)**.** If $e_1 = \langle 1,0\rangle$ is based at $p = (x,y)$ in the orthographic map M, its map-length is

$$\|\langle 1,0\rangle\|_{\text{map}} = \sqrt{\frac{1-y^2}{1-x^2-y^2}}$$

To see this, notice that $D\psi_p\langle 1,0\rangle$ is simply the first column of the derivative matrix, and then we need only compute its length in $\mathbb{E}^3$:

$$\begin{aligned}\left\|\left\langle 1, 0, \frac{-x}{z}\right\rangle\right\| &= \sqrt{1 + \frac{x^2}{z^2}} \\ &= \sqrt{\frac{z^2 + x^2}{z^2}} \\ &= \sqrt{\frac{1 - x^2 - y^2 + x^2}{1 - x^2 - y^2}} \\ &= \sqrt{\frac{1 - y^2}{1 - x^2 - y^2}}\end{aligned}$$

Knowing the infinitesimal lengths in the x direction lets us compute the total length of the horizontal curve $\gamma(t) = (t, 0)$ in our map. In M this is just the diameter of the disk, so its length appears to be 2: but we know this isn't right! The diameter of the disk represents *half of a great circle* going from the equator to the north pole and back, so its length should be π. Let's do the calculation to confirm:

Exercise 22.1. The map-length of the curve $\gamma(t) = (t, 0)$ from $t = -1$ to $t = 1$ in the orthographic projection is π.

To see this, recall that the map length of γ is given by integrating the infinitesimal map-lengths of its tangent vectors:

$$\text{length}_{\text{map}}(\gamma) = \int_{-1}^{1} \|\gamma'\|_{\text{map}}\, dt$$

Because $\gamma(t) = (t, 0)$ its immediate to see $\gamma'(t) = \langle 1, 0\rangle$ for all time, and using the above exercise then we see at the point $(x, y) = \gamma(t) = (t, 0)$, the map length of $\langle 1, 0\rangle$ is

$$\sqrt{\frac{1 - 0^2}{1 - t^2 - 0^2}} = \frac{1}{\sqrt{1 - t^2}}$$

Integrating this gives a familiar expression: we saw this exact integral in the definition of the arc cosine function (Proposition 14.2)! Since $\arccos(x) = \int_x^1 \frac{1}{\sqrt{1-t^2}} dt$, we see the expression we have come to is exactly $\arccos(-1)$, which by definition is the arclength along the top half of the unit circle (ie all the way from $x = 1$ to $x = -1$). Thus

$$\text{length}_{\text{map}}(\gamma) = \int_{-1}^{1} \frac{dt}{\sqrt{1 - t^2}} = \arccos(-1) = \pi$$

Beyond applying this just to the basis vector $\langle 1, 0\rangle$ we can us the same technique to find the length of any vector in T_pM:

Example 22.2 (Orthographic Map Infinitesimal Lengths)**.** If $v = \langle a, b\rangle$ is based at $p = (x, y)$ in the unit disk M, its map-length in the orthogonal projection is

$$\|v\|_{\text{map}} = \sqrt{}$$

We calculate this by applyign $D\psi_p$ to v, and then finding the length in $\mathbb{E}^3$:

$$D\psi_p\langle a, b\rangle = a\begin{pmatrix} 1 \\ 0 \\ \frac{-x}{z} \end{pmatrix} + b\begin{pmatrix} 0 \\ 1 \\ \frac{-y}{z} \end{pmatrix} = \begin{pmatrix} a \\ b \\ \frac{-ax-by}{z} \end{pmatrix}$$

Computing the length, we get

$$\begin{aligned} \|D\psi_p v\| &= \sqrt{a^2 + b^2 + \frac{(-ax - by)^2}{z^2}} \\ &= \sqrt{a^2 + b^2 + \left(\frac{ax + by}{1 - x^2 - y^2}\right)^2} \end{aligned}$$

Knowing the map-length of an arbitrary vector on T_pM lets us precisely describe the map-disks:

Example 22.3 (Orthographic Map Disks)**.** At the point $p = (x, y)$, the map-disk is the set of all vectors $\langle a, b\rangle \in T_pM$ where

$$a^2 + b^2 + \left(\frac{ax + by}{1 - x^2 - y^2}\right)^2 \leq 1$$

At the center of our map where $(x, y) = (0, 0)$, this equation for the map-disk reduces to the equation for the standard round unit disk $a^2 + b^2 = 1$. This means at the origin, we expect to see *essentially no distortion* in either size or angle! However, as soon as x or y are nonzero, things quickly change. Consider the point $(x, y) = (0, 1/2)$. What's the map-disk here? Just plugging in gives

$$a^2 + b^2 + \left(\frac{0 + \frac{1}{2}b}{1 - 0^2 - \frac{1}{4}}\right)^2 \leq 1$$

and after a bit of simplification, we find

$$a^2 + \frac{13}{9}b^2 \leq 1$$

This is an ellipse! Thus, we see at $(0, 1/2)$ the map is not distorting distance in the x direction (the coefficient of a is still 1) but it *is* distorting distance in the y direction.

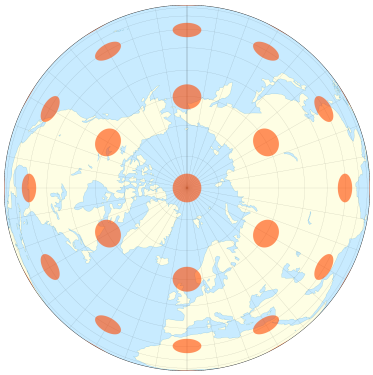


Figure 22.5.: Map Disks of the orthographic projection

Since it's affecting the two directions unequally, we expect that it will not be preserving angles very well, so let's confirm.

Example 22.4 (Orthographic Map Angle)**.** The vectors $\langle 1, 0\rangle$ and $\langle 0, 1\rangle$ are not map-orthogonal, even though the *look* orthogonal in the Euclidean plane where we drew the map!

To see this, all we need to do is compute the map-dot-product between $\langle 1, 0\rangle$ and $\langle 0, 1\rangle$ and see that its nonzero. With the derivative of our parameterization already in hand, this is quick work:

$$D\psi_p\langle 1, 0\rangle = \langle 1, 0, -x/z\rangle$$
$$D\psi_p\langle 0, 1\rangle = \langle 0, 1, -y/z\rangle$$

These are the two true vectors that $\langle 1, 0\rangle$ and $\langle 0, 1\rangle$ on our map represent, so the map dot product is equal to their actual dot product on [2]:

$$\langle 1, 0, -x/z\rangle \cdot \langle 0, 1, -y/z\rangle = \frac{xy}{z^2} = \frac{xy}{1 - x^2 - y^2}$$

Thus, whenever x and y are both nonzero, the vetors $\langle 1, 0\rangle$ and $\langle 0, 1\rangle$ don't actually point in orthogonal directions on our map!

Exercise 22.2. Can you find the coordinates (x, y) of a point on the map where the vectors $\langle 1, 0\rangle$ and $\langle 0, 1\rangle$ only make a 45-degree angle with one another?

Hint: can you make the problem easier for yourself by restricting x and y to lie on some line, so the problem ends up having one variable instead of two?

You can see how this would make such a map difficult to use for navigation: it would *look* like the map is telling you to turn 90 degrees but in reality you should only turn half that!

22.2. Archimedes' Map

In Archimedes' most cherished work *The Sphere and the Cylinder*, he proved that the surface area of the sphere and the cylinder agreed by showing that horizontally projecting the surface of the sphere onto the cylinder preserved infinitesimal areas, and thus (via integration) the total area. This suggests a means of creating a map of the earth which displays the true areas for each region: first project horizontally onto the surrounding cylinder, then unroll the cylinder onto the plane.

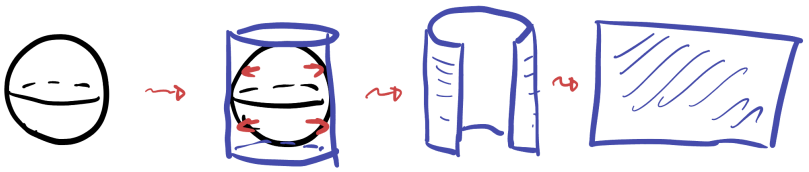


Figure 22.6.: Defining Archimedes' Map

For step 1, what happens when we horizontally map a point $(x, y, z) \in {}^2$ to the cylinder? Well, its height or z coordinate does not change, and the xy coordinates do not chnage direction, only length. That means that there must simply be some scalar s such that

$$(x, y, z) \mapsto (sx, sy, z)$$

How do we find this scaling factor s? Well, we know at the end we want the point to lie on the cylinder, so that its x and y coordinates lie on the unit circle. This means we need

$$(sx)^2 + (sy)^2 = 1 \implies s = \frac{1}{\sqrt{x^2 + y^2}}$$

We can then use the fact that (x, y, z) originally lies on the sphere, so that $x^2+y^2+z^2 = 1$ to see we can replace this with $\frac{1}{1-z^2}$ if we wish, to get

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

Now, we just need to unroll the cylinder onto the plane: this means we continue to leave the height, or z direction alone, but we wish to find the angle θ which the x and y coordinates make on the unit circle. Because the tangent of this angle is opposite over adjacent, we can get an explicit formula:

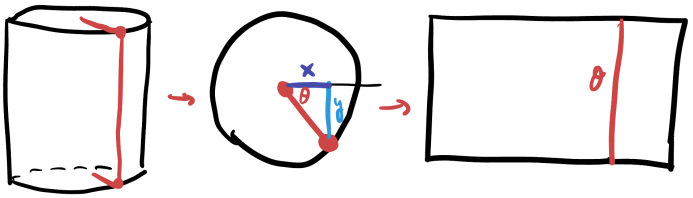


Figure 22.7.: The angle θ is defined by looking down on the cylinder with respect to the x axis: its tangent is y/x.

$$\tan\theta = \frac{\frac{y}{\sqrt{1-z^2}}}{\frac{x}{\sqrt{1-z^2}}} = \frac{y}{x}$$

Definition 22.2 (Archimedes' Map)**.** Let $R \subset {}^2$ be everything except the north and south poles, and let $M \subset \mathbb{E}^2$ be the rectangle $M = \{(\theta, h) \mid -\pi < \theta \leq \pi,\ -1 < h < 1\}$. We define Archimedes' chart as

$$\phi(x, y, z) = (\theta, h) = \left(\arctan \frac{y}{x}, z \right)$$

and its inverse, Archimedes' parameterization

$$\psi(\theta, h) = \left(\sqrt{1-h^2}\cos\theta, \sqrt{1-h^2}\sin\theta, h \right)$$

In cartography, this map is not named after Archimedes but rather the 17th century mapmaker Lambert, and called the Lambert Cylindrical projection, or the Lambert Equal-Area projection.

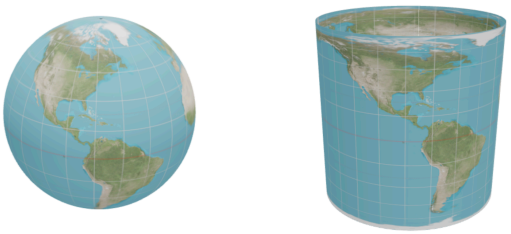


Figure 22.8.: Archimedes map, horizontally projecting the sphere onto the cylinder

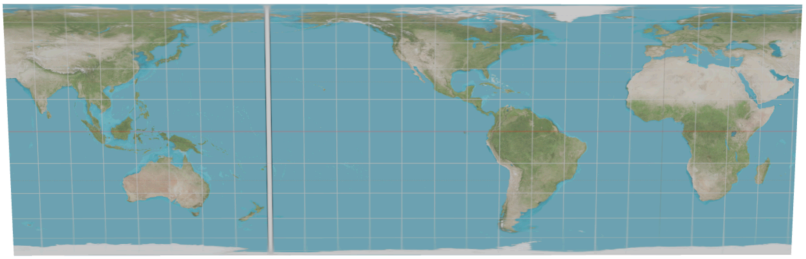


Figure 22.9.: Archimedes map unrolled onto a rectangle in the plane.

While it is natural to us earthlings to project the earth onto a cylinder whose axis passes through the north and south poles, it is by no means necessary: the sphere is homogeneous after all! So there are many *unfamiliar maps* that can be produced by this technique, sharing all the same mathematical properties. Here we illustrate just one option, unrolling along an axis through the equator.

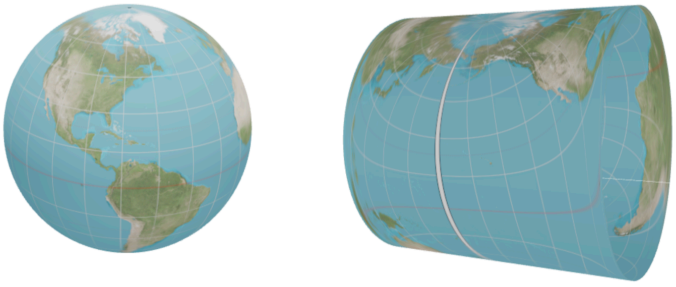


Figure 22.10.: Projecting onto a different bounding cylinder.

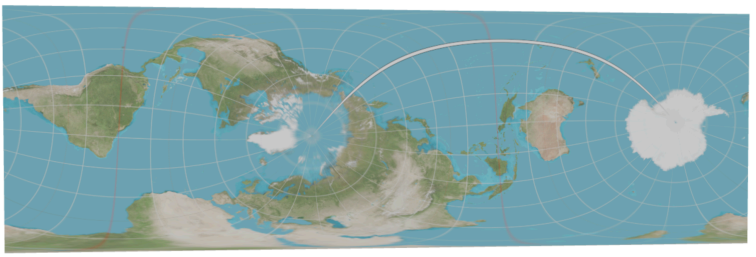


Figure 22.11.: The resulting map of the earth!

To understand what Archimedes map does to regions of the sphere, a useful spot to start is to calculate its map-disks (Tissot's Indicatrices) and see what shape they are!

Theorem 22.1 (Archimedes Map Disks). *At the point $p = (\theta, h)$ on the Archimedes map, the map disk of unit radius is given by the set of all vectors $\langle a, b\rangle \in T_pM$ with*

$$a^2(1-h^2) + \frac{b^2}{1-h^2} \leq 1$$

Proof. We calculate this by just applying $D\psi_p$ to the vector $\langle a, b\rangle$, then finding the resulting length-squared in $\mathbb{E}^3$, and simplifying a lot. □

Exercise 22.3. Do this calculation!

Because essentially all calculations require us to know infinitesimal information about the parameterization (translating vectors on the map to their true counterparts on the sphere), we begin with a calculation of $D\psi$:

$$D\psi = \begin{pmatrix} \partial_\theta\sqrt{1-h^2}\cos\theta & \partial_h\sqrt{1-h^2}\cos\theta \\ \partial_\theta\sqrt{1-h^2}\sin\theta & \partial_h\sqrt{1-h^2}\sin\theta \\ \partial_\theta h & \partial_h h \end{pmatrix}$$

$$= \begin{pmatrix} -\sin\theta\sqrt{1-h^2} & \frac{-h\cos\theta}{\sqrt{1-h^2}} \\ \cos\theta\sqrt{1-h^2} & \frac{-h\sin\theta}{\sqrt{1-h^2}} \\ 0 & 1 \end{pmatrix}$$

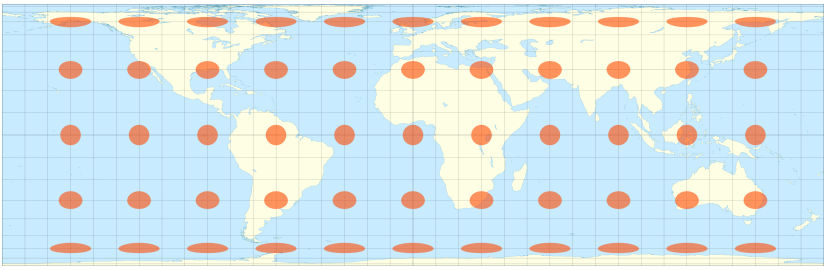


Figure 22.12.: Map Disks on Archimedes' Map

The map-disks are ellipses, meaning that angles in general are not preserved. However, we can calculate the area of these map-disks to understand better the area distortion (or lack thereof) on the map. The ellipses we found turn out to be lined up nicely along the two axes of $\mathbb{E}^2$, much like the ellipses whose formulas we first uncovered in Definition 13.4. Thus, their areas are computable as in Exercise 15.4: its equal to πab where a and b are the 'radii' of the ellipse along the x and y direction.

Theorem 22.2 (Archimedes Map is Area Preserving)**.** *At every point $p \in M$ the map-disk has area π. Since the map disk represents the infinitesimal vector which map to the unit disk tangent to the sphere (which also has area π), the parameterization does not distort infinitesimal area.*

Proof. If an ellipse has radii r_1 along the x axis and r_2 along the y axis, we saw its area is $\pi r_1 r_2$, which we could compute by stretching the unit circle in Exercise 15.4. The formula for such an ellipse is

$$\frac{x^2}{r_1^2} + \frac{y^2}{r_2^2} \leq 1$$

So in our case we have $r_1 = \frac{1}{\sqrt{1-h^2}}$ and $r_2 = \sqrt{1-h^2}$. These are reciprocals of one another, so $r_1 r_2 = 1$ and

$$A = \pi r_1 r_2 = \pi$$

But this is the area of the unit disk! So, at each point p of the map, the mapdisks (Euclidean) area accurately represents its true area on the sphere. The map does not distort infinitesimal areas. □

Because we are still learning how to compute effectively with maps, we'll give a second proof of this fact, where we do not bother working out the details of our map disks, but rather just directly look at infinitesimal lengths are areas, figuring out what happens to an infinitesimal unit square.

Exercise 22.4. Give a second proof that Archimedes map is area-preserving, that looks at infinitesimal squares instead of ellipses. Show that at each point $p \in M$ the vectors $\langle 1, 0\rangle$ and $\langle 0, 1\rangle$ are sent by ψ to orthogonal vectors on the sphere. Find their lengths on the sphere (ie the map-lengths), and use this data to find the area of the infinitesimal rectangle they form.

Now only does this immediately imply archimedes overall result that the two areas are equal (each area is *by definition* the integral of its infinitesimal areas, and we just showed all the infinitesimal areas are equal), but it also shows that the area of any region on the map accurately portrays the true area of the region it represents on the sphere.

Theorem 22.3. *Let $R \subset {}^2$ be a region on the sphere, and $M = \phi(R) \subset \mathbb{E}^2$ its map under Archimedes chart. Then*

$$\text{area}(R) = \text{area}(M)$$

Proof. Because the chart and parameterization are inverses, we could just as well call $\phi(R) = M$ the map, and then the original region is $\psi(M) = R$. We compute the area of R as an integral, and use ψ to write it as an integral over M:

$$\text{area}(R) = \iint_R dA_2 = \iint_{\psi(M)} dA_2 = \iint_M dA_{\text{map}}$$

But now we know that $dA_{\text{map}} = dA_{\mathbb{E}^2}$, that's what we've calculated! So we can sub this out, and then realize the resulting integral is just the *definition* of the Euclidean area of M in the plane:

$$= \iint_M dA_{\mathbb{E}^2} = \text{area}(M)$$

□

However, its important not to forget what we learned along the way: the map-disks for Archimedes map form extremely distorted ellipses as one approaches the poles: with horizontal length stretching near infinite and vertical height crushing to zero. This map *massively* distorts the shapes of regions, distances between points and angles between curves in its attempt to preserve area. Like the orthographic map before it, this makes Archimedes' map unsuitable for navigational tasks, where figuring out accurately what *direction* you must go to reach your desired destination is of utmost importance.

22.3. Equirectangular Projection

There is an entire collection of maps which are defined as modifications of Archimedes' original idea, these days called *cylindrical projections* as they start by projecting onto a cylinder. Perhaps the two most common of these are the Mercator projection (discussed in the next chapter as a potential *final project* opportunity), and the Equirectangular projection, which I will only briefly mention here (for anyone who is doing a final project on Maps and would like another, easy-to-compute-with and yet still real-world example).

The problem the Equirectangular projection tries to solve is the *vertical distortion* of Archimedes' map. Archimedes made the vertical height on the map equal the *vertical height of the sphere at that point*: this clever move ensured area was preserved, but what if we wanted the vertical height on the map to actually be related to the *north-south distance*? Archimedes' map fails badly at this, as we see in the picture above.

What we want from our new map projection is that it

- Continues to be a cylindrical projection
- Distance along the (vertical) h axis of the map accurately reflects the actual *geodesic distance* along lines of longitude on the sphere.

Because we are still projecting onto a cylinder, the chart for such a map is still going to have $\theta = \arctan(y/x)$. But as arclength is angle, the height (distance from the equator) will need to be the angle φ that a point on the sphere makes with the xy plane:

Definition 22.3. The chart for the equirectangular projection is defined on the region $R \subset {}^2$ of the sphere containing all points except the north and south poles, and maps onto the rectangle

$$M = \{(\theta, \varphi) \in \mathbb{E}^2 \mid -\pi \leq \theta \leq \pi,\ -\pi/2 \leq \varphi \leq \pi/2\}$$

by the chart function

$$\phi(x, y, z) = \left(\arctan \frac{y}{x}, \arcsin z\right)$$

Exercise 22.5. Derive the *parameterization* for the Equirectangular projection.

Figure 22.13.: The equirectangular projection maps the earth onto a cylinder of height π: the distance from the north to south pole.

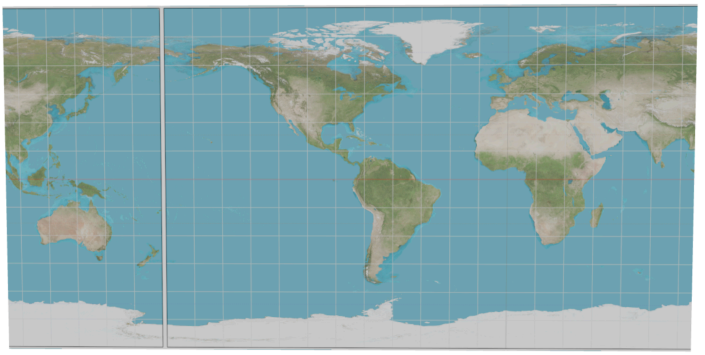


Figure 22.14.: Unrolled into the plane, the Equirectangular map is a rectangle twice as wide as tall.

Exercise 22.6. Spherical coordinates in mathematics and physics are almost the same as the equirectangular projection: the only difference being a convention on *where* to measure the angle φ from. Here we've measured it from the equator, so that it accurately captures *latitude* on the earth. But in spherical coordinates it is usually measured from the north pole.

Write down the chart and parameterization for spherical coordinates, and see that it is what you are taught in multivariable calculus!

Because this map captures distance accurately along the equator, as well as north-south distance along lines of longitude, it is an easy map to work with, and has become

the *default* map in many contexts.

Exercise 22.7. Find the important map-quantities in the equirectangular projection:

- Find the map-length of a vector $v = \langle a, b \rangle$ based at $(\theta, \varphi) \in M$.
- Find the equation for the map-disk at (θ, φ). Show that it's an ellipse: what do its *vertical radius* and *horizontal radius* tell you about the map?
- Does this map preserve angle?
- Find the *map-area*: since horizontal and vertical lines in M map to orthogonal curves on ^{2} (latitude and longitude), infinitesimal squares on M are taken to infinitesimal rectangles on ^{2}. Does this map preserve area?

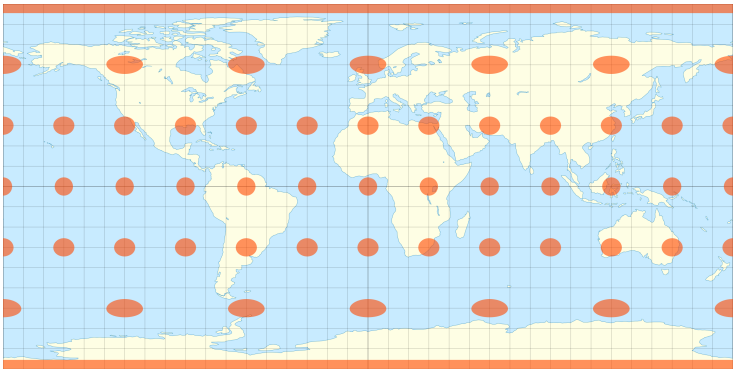


Figure 22.15.: Map Disks on the equirectangular projection.

23. Mercator

The Mercator projection is the classic map

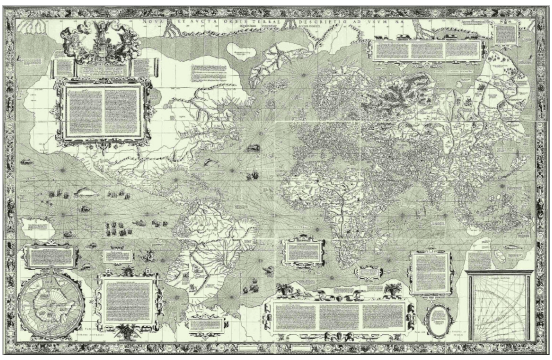


Figure 23.1.: The original Mercator map: check out California!

One means of building Mercator's map is to begin with Archimedes', and and perform some modifications. We will follow this, and will attempt to change as little from the previous map as possible: indeed we will attempt to construct this new map also by first projecting the sphere onto its bounding cylinder, and then unrolling that cylinder onto the plane.

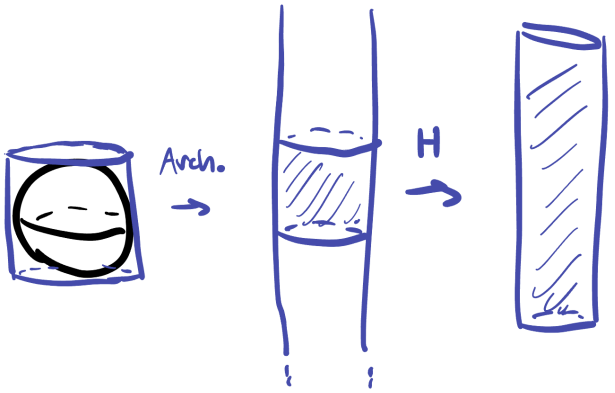


Figure 23.2.: The Mercator projection (an all cylindrical projections) are Archimedes map followed by some sort of vertical stretch.

The only choice we made in the above derivation of Archimedes's map was that the projection was *horizontal*, or that the height h on the map was equal to the original z coordinate. Here, we must do something else (lest we end up with the same map!) so we let h be a function of z: this is equivalent to first doing Archimedes map, then stretching the vertical axis of the cylinder by a function H. Different choices of $H(z)$ will describe different *cylindrical projections*, and our goal here is to find a good choice for H.

Definition 23.1 (A General Cylindrical Projection)**.** Let $H(z)$ be any height function taking the latitudes of the unit sphere $z \in [-1, 1]$ to some height h along the cylinder. Then the cylindrical projection corresponding to H is given by

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

There are all sorts of maps you can make by choosing different functions H here (and then unrolling the resulting cylinder onto the plane). But most of them will distort angles: they'll take infinitesimal squares on the map to infinitesimal rectangles on the sphere, and vice versa (like Archimedes' example). Only one will take squares to squares *at every single point* - and this is what Mercator was after!

Here was his idea: we know how much the circle starting at height z has to get stretched *horizontally* - because we are projecting it onto the cylinder of radius 1. At height z, the radius of the circular slice of the sphere $x^2 + y^2 + z^2 = 1$ must be

$$r = \sqrt{x^2 + y^2} = \sqrt{1 - z^2}$$

Thus its circumference is $2\pi\sqrt{1-z^2}$, and it is going to get stretched to the unit circle, with circumference 2π. So, we know that our map is scaling by a factor of $\frac{1}{\sqrt{1-z^2}}$ horizontally. This means that vertically, we must ensure it is *also* scaling by $\frac{1}{\sqrt{1-z^2}}$!

But the vertical direction involves two stretches: first we need to think about the effect of *horizontal projection* onto the cylinder, and then second, we need to tack on the *vertical stretch* induced by H.

The first of these is something we can already compute using our knowledge of Archimedes map! At a point p on Archimedes' map, the vertical vector $\langle 0,1\rangle$ points along this cylinder. Its map-length is

$$\|\langle 0,1\rangle\|_{\text{map}} = \|D\psi_p\langle 0,1\rangle\|_2 = \left\| \begin{pmatrix} \frac{-h\cos\theta}{\sqrt{1-h^2}} \\ \frac{-h\sin\theta}{\sqrt{1-h^2}} \\ 1 \end{pmatrix} \right\|_2$$

Where we found the vector as the second column of $D\psi$ for Archimedes map. Computing this length with the 3D pythagorean theorem (which measures the true length on the sphere, as the tangent spaces to ^{2} use the Euclidean dot product), we see

$$\|\langle 0,1\rangle\|_{\text{map}} = \sqrt{1+\frac{h^2}{1-h^2}} = \frac{1}{\sqrt{1-h^2}}$$

This means a vector of *Euclidean Length 1* on the map gets sent to a vector of length $1/\sqrt{1-h^2}$ on the sphere, so when going from the map to the sphere the length of a vector is divided by $\sqrt{1-h^2}$. This means inversely, projecting from the sphere to the map *multiplies* the true length of a vector by $\sqrt{1-z^2}$ (where we use z on the sphere and h on the plane/cylinder, just to keep things separate).

Exercise 23.1. Check all this!

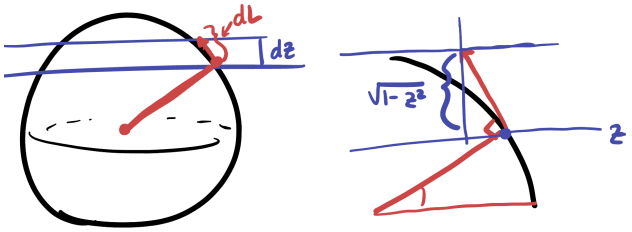


Figure 23.3.: At height z, the horizontal projection of a vector changes its length by a factor of $\sqrt{1-z^2}$.

Now we're ready to think about the stretch H we need. Going from the sphere to the cylinder stretched the vertical direction by $\sqrt{1-h^2}$, and we want the end result to be that the map gets stretched instead by $1/\sqrt{1-z^2}^2$

That is, our function $H(z)$ must have the property that it stretches by $\frac{1}{\sqrt{1-z^2}}$ to *undo the horizontal projection*, and then stretches again by $1/\sqrt{1-z^2}$ to get where we want. This means

$$H'(z) = \frac{1}{\sqrt{1-z^2}}\frac{1}{\sqrt{1-z^2}} = \frac{1}{1-z^2}$$

This is a differential equation for our function H which we can solve via integration! (Hey, remember me? Integration by Partial Fractions?)

$$H(z) = \int_0^z \frac{1}{1-z^2}dz = \log\sqrt{\frac{1+z}{1-z}}$$

Definition 23.2 (Mercator's Map)**.** Let $R \subset {}^2$ be all the points of the sphere except the north and south poles, and let $M \subset \mathbb{E}^2$ be the entire vertical strip $M = \{(\theta, h) \mid -\pi < \theta \leq \pi,\ -\infty < h > \infty\}$. Then Mercator's map has chart

$$\phi(x,y,z) = \left(\tan\frac{y}{x}, \log\sqrt{\frac{1+z}{1-z}}\right)$$

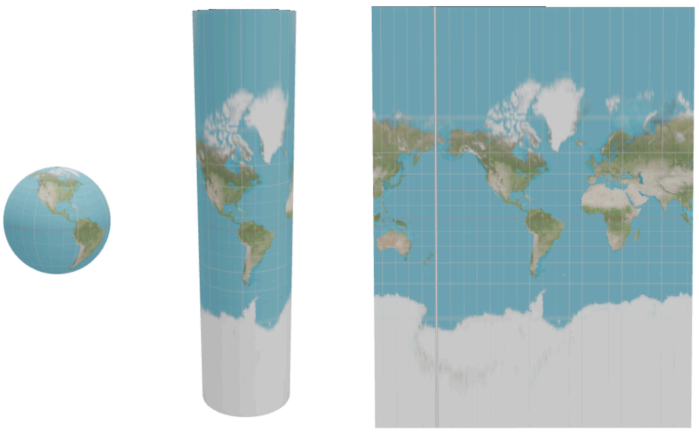


Figure 23.4.: Mercator's map projection projecting the Earth onto a cylinder and then unrolling onto a sphere.

Exercise 23.2. Find the parameterization for the mercator map. *Hint: first calculate the inverse function of* $h = \log(\sqrt{(1+z)/(1-z)})$ *and show it is*

$$z = \frac{e^{2h} - e^{-2h}}{e^{2h} + e^{-2h}}$$

Now we know the z component of the parameterization, and we know the x, y components together are going to be some multiple of $(\cos\theta, \sin\theta)$. But which multiple? Well, we *do know* that (x, y, z) must lie on the sphere! And that determines everything:

Exercise 23.3. Show that if the point

$$(x, y, z) = \left(k\cos\theta, k\sin\theta, \frac{e^{2h} - e^{-2h}}{e^{2h} + e^{-2h}}\right)$$

lies on the unit sphere, then

$$k = \frac{2}{e^{2h} + e^{-2h}}$$

Putting these two exercises together, we have successfully computed the parameterization to the Mercator projection!

Theorem 23.1. *The parameterization for the mercator projection is the map* $\psi : (-\pi, \pi) \times (-\infty, \infty) \to {}^2$

$$\begin{aligned}\psi(\theta, h) &= \left(\frac{2\cos\theta}{e^{2h} + e^{-2h}}, \frac{2\sin\theta}{e^{2h} + e^{-2h}}, \frac{e^{2h} - e^{-2h}}{e^{2h} + e^{-2h}}\right) \\ &= \left(\frac{\cos\theta}{\cosh h}, \frac{\sin\theta}{\cos h}, \tanh h\right)\end{aligned}$$

Where in the second line I have written these combinations of exponentials in their equivalent form using *hyperbolic trigonometric functions*. We will meet these functions in a different context very soon!

Figure 23.5.: Different unfamiliar views of the Mercator projection, arising from choosing points other than the north and south pole to become the axis of the cylinder.

The fact that Mercator's map sends preserves angles is a huge advantage not only for navigation, but also for calculation. Since it sends infinitesimal squares to infinitesimal squares, it scales *all lengths* by the same scaling factor, which we can find by

Exercise 23.4. Show that at any point (θ, h) in the mercator map, the map length of a vector $v = \langle a, b\rangle$ is just its Euclidean length divided by $\cosh(h)$:

$$\|v\|_{\text{map}} = \frac{1}{\cosh h}\|v\|_{\mathbb{E}^2} = \frac{\sqrt{a^2+b^2}}{\cosh h}$$

Hint: since all vectors are scaled the same, can you find the map length of $\langle 1, 0\rangle$*?*

Being able to measure the infinitesimal length of any vector lets us write down the map-disks for the mercator projection, and also lets us compute the infinitesimal area element:

Exercise 23.5. Explain why the map-area is can be calculated by

$$dA_{\text{map}} = \frac{1}{\cosh^2(h)}dxdy$$

Hint: think about what it does to an infinitesimal unit square

Exercise 23.6. At a point $p = (\theta, h)$, write down an equation that determines when a tangent vector $v = \langle a, b\rangle \in T_p\mathbb{E}^2$ is in the map-disk at that point. Explain why the map-disk is a circle: what's its Euclidean radius?

23.0.1. Application: Geodesics

There is *one important result* about the sphere that has eluded us this entire time. In the plane, we saw that there were three different notions of line that defined the same curves: distance minimizing, straight, and fixed by isometries. For the sphere, we wrote down the same three conditions, and said we would prove them equivalent here as well. But what did we *actually do*?

We first discovered great circles as they were fixed by isometries, and then we proved that these curves were also straight, using the correct definition of acceleration on the sphere. Ever since, we've been using them to define our distance - but we never actually *proved* they are distance minimizing! I promised we would do that at a future time (in a non-circular way) when we had developed more tools to help us, so we could avoid some nasty integrals in 3D space.

And now is that time! One of the superpowers of using *maps* is it lets us take the sphere which was originally a curved surface in three dimensions, and accurately represent it by *regions in the 2-dimensional plane.* And calculus on the plane is much easier than calculus on a surface in three dimensions. This lets us mimic quite closely the original proof we gave in Euclidean geometry that lines were distance-minimizing (Theorem 12.1).

Here using the Mercator map, we will focus on a line of longitude, which is a vertical line on the map. We know these great circles are both straight and fixed by symmetries, so our goal now is to show they are length minimizing (at least, when they go less than half way around the sphere)

Theorem 23.2. *Let $L \in \mathbb{R}$. Then the curve $\gamma(t) = (0, t)$ for $t \in [0, L]$ in $\mathbb{E}^2$ is Mercator map-length minimizing: it represents a curve of shortest length between its endpoints on the sphere.*

Proof. Let $\alpha(t) = (x(t), t)$ be a curve between $(0, 0)$ and $(0, L)$ in the Mercator map, for $t \in [0, L]$. We will show that the map-length of α is greater than or equal to the map-length of the straight line $\gamma(t) = (0, t)$.

PICTURE

First, let's write down the infinitesimal length of α: the tangent vector is $\alpha' = \langle x', 1 \rangle$ so

$$\|\alpha'\|_{\text{map}} = \frac{1}{\cosh(t)} \sqrt{(x')^2 + 1}$$

The integral of these infinitesimal lengths gives the overall length:

$$\text{length}_{\text{map}}(\alpha) = \int_a^b \frac{1}{\cosh t} \sqrt{(x')^2 + 1}\, dt$$

Now lets do the same for $\gamma' = \langle 0, 1\rangle$ at the point $\gamma(t) = (0, t)$: its infinitesimal length is

$$\|\gamma'\|_{\text{map}} = \frac{1}{\cosh t}\sqrt{0^2 + 1^2} = \frac{1}{\cosh t}$$

And so its total length is

$$\text{length}_{\text{map}}(\gamma) = \int_0^L \frac{1}{\cosh t}\, dt$$

Remembering that $\cosh(t) = (e^t + e^{-t})/2$, its possible to actually do this integral! But we will not need its value here. Instead, all we need to show is that our arbitrary curve α is *longer* than this.

And this is clearly true! Since $(x')^2$ is a nonnegative number, we know that for all t

$$(x')^2 + 1 \geq 1$$

The same equality remains true after taking the square root, and after dividing by $\cosh(t)$, so at each point of the map

$$\|\alpha'\|_{\text{map}} \geq \|\gamma'\|_{\text{map}}$$

Integrating this we see that

$$\text{length}_{\text{map}}(\alpha) \geq \text{length}_{\text{map}}(\gamma)$$

so γ, the great circle, is the shortest among all curves of this form! □

The careful reader will notice that this proof is not *quite* technically complete: we showed that the great circle is the shortest of all curves of the form $(x(t), t)$: but what about all general curves $(x(t), y(t))$? Can you extend the argument to this case? *Hint: U-sub!*

Exercise 23.7. The careful reader will notice that this proof is not *quite* technically complete: we showed that the great circle is the shortest of all curves of the form $(x(t), t)$: but what about all general curves? Can you extend the argument to this case, and show if $\alpha(t) = (x(t), y(t))$ for $t \in [a, b]$ with $\alpha(a) = (0, 0)$ and $\alpha(b) = (0, L)$, then $\text{length}_{\text{map}}(\alpha) \geq \text{length}_{\text{map}}(\gamma)$?

Hint: look back at the Euclidean proof where we did this: Theorem 12.1. Can you prove that

$$\text{length}_{\text{map}}(\alpha) \geq \int_a^b \frac{y'}{\cosh y} dt$$

and then perform a u-sub to relate this to the length of the great circle γ?

This finishes off the final fact we needed to complete our study of the geometry of the sphere. Congratulations!

Exercise 23.8. Show that

$$\int \frac{1}{\cosh x} dx = 2\arctan(e^x) + C$$

Hint: write $\cosh x$ *in terms of its definition in exponentials, multiply the top and bottom of the resulting fraction by* e^x *and do a u-sub to get an integral related to arctan.*

23.1. The Mapmaker's Dilemma

We've now gotten rather comfortable computing true quantities about spherical geometry using a map and calculus. Since all of our maps have distorted the sphere in some pretty serious ways, its pretty important to have these abilities as you cant just trust your eyes!

Of course some maps did better than others: orthographic projection messed up basically every quantity we could think of, whereas Archimedes map managed to accurately portray area and Mercator's accurately represented angles. But none of our maps accurately represented *both* area and angle at the same time.

Indeed - while we did not check it, all the maps in the Cartography chapter have this property: some of them preserve area, some of them preserve angle, but none of them do so simultaneously. But this doesn't mean its *impossible* to make such a map - there's an infinite variety of things that we haven't tried (and an infinite number of possible maps that *no human has ever drawn*) - perhaps one of them is able to preserve two quantities of the sphere at once? After all, some of the maps we did see in the previous chapter did a pretty good job of approximately preserving both (and shifting some of the complexity to the shape of the mapping region M). Who is to say that someday a supercomputer running AI mapping software wont discover an absolutely absurdly complicated domain M in the plane, and a map of ^{2} drawn in M which manages to accurately represent both?

Math, that's who says this will never happen.

Theorem 23.3 (The Mapmaker's Dilemma)**.** *It is impossible to make a map which simultaneously accurately represents both angles and areas.*

Proof. Assume for the sake of contradiction that there is such a map M, defined on some region R of the sphere, and let ϕ be its parameterization. If this map preserves angles then ψ sends infinitesimal squares of M to infinitesimal squares on the sphere. But if it also preserves *area* it must send a square with side length s (and thus area s^2) to another infinitesimal square of area s^2 (and thus side length s).

This means that our map must preserve all infinitesimal lengths! Choose any point $p \in M$ and any vector $v \in T_pM$, and build a square with v as one side (if $v = \langle a, b \rangle$ we can use the orthogonal vector $\langle -b, a \rangle$ of the same length as the other side defining the square). Now $D\psi_p$ maps this to another square whose side lengths are the same, so $\|D\psi_p v\| = \|v\|$!

But a map that preserves infinitesimal distances is an isometry - and this is going to spell trouble. In particular, we know that isometries send geodesics to geodesics, circles to circles, and preserve the length of all curves. Because of this, as we saw in the chapter on curvature, isometries preserve the value of all the terms showing up in the limit that defines curvature: and any two points related by an isometry must have the same curvature.

But ψ relates points of the plane to points of the sphere! So this implies that the sphere and the plane have the same curvature, which is a contradiction: we know the plane's curvature is 0 and the sphere's curvature is 1. □

During the proof of this we noticed another, easier dilemma: its impossible to make a map that preserves distances!

Theorem 23.4 (The Mapmaker's Dilemma, Distance)**.** *It is impossible to make a map which accurately shows the distance between any pair of points on the sphere.*

Proof. Such a map would then preserve infinitesimal distances, and thus be an isometry. But this would again preserve the curvature, which implies a contradiction: that the sphere and the plane have the same curvature! □

This is a pretty amazing result: proving nonexistence theorems are hard, as you have to somehow rule out all of the possible examples, even the ones you can't imagine. Proving such a theorem often requires finding some deep mathematical property that can tell things apart, some sort of *invariant*. And for us, that invariant is *curvature*.

You can see the mapmaker's dilemma as in some sense a capstone of this entire section of the course: if you dig deep enough almost everything we have done since the introduction of calculus goes into its proof in some way or another.

From one perspective, it essentially finishes off the entire theory of mapmaking, answering the fundamental question. But from another, it tells us useful pragmatic information about how to move on: *don't worry about making your map look accurate*, our theorem warns us *you don't need to look at it to compute, anyway! That's what calculus is for. Just make it easy to work with. There's no best map, but there may be a good map for your specific desires or purpose, just build that one.*

And build such a map, we will!

24. Stereographic

In the wake of our proof of the mapmaker's dilemma, we rise once more to build a map: this time no longer worried about trying to make it optimal in every regard, but just mathematically simple, and easy to interpret.

One very natural contender for such a map is *stereographic projection* first invented by the Greeks to make a star chart, representing the spherical sky on a flat piece of paper. As we've come to expect of Greek mathematics, this map has a *geometric definition*

Definition 24.1 (Stereographic Projection)**.** Given the unit sphere ^{2} in $\mathbb{E}^3$, the *stereographic projection* of a point $p \in {}^2$ is the point $\sigma(p) \in \mathbb{E}^2$ such that the straight line in $\mathbb{E}^3$ connecting p to the north pole $N = (0, 0, 1)$ intersects the xy plane in the point $(\sigma(p), 0)$.

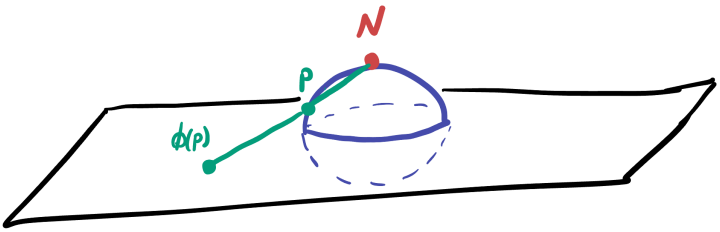


Figure 24.1.: Stereographic projection maps the sphere to the plane as though there were a light source at the north pole, casting shadows.

This is much easier to see in three dimensions with an animation than a drawing-by-hand, so here's one to help (though, in both of these animations I have moved the sphere *above* the plane: this doesn't change the math in any essential way but makes things easier to see what is going on)

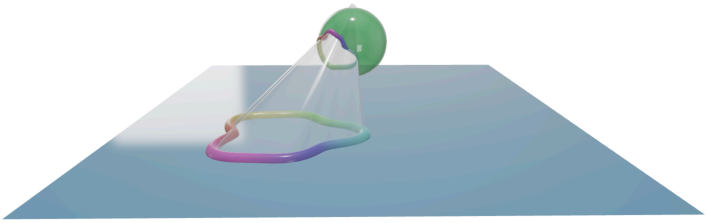


Figure 24.2.: Stereographic projection as a light source at the north pole casting a shadow onto the plane.

Stereographic projection acts like finding a shadow, from a light source at the top of the sphere.

From this picture, we can derive an algebraic formula describing this projection (note that while we have visualized it above as though the sphere is above the plane, the algebra is a bit easier if you assume the plane intersects the sphere in the equator, as in the hand-drawn image above).

Proposition 24.1 (Stereographic Projection Formula)**.** *Stereographic projection provides a map of the sphere (except for the north pole), onto the entire Euclidean plane. Its chart is*

$$\phi(x, y, z) = (X, Y) = \left(\frac{x}{1-z}, \frac{y}{1-z}\right)$$

and the parameterization

$$\psi(X, Y) = (x, y, z) = \left(\frac{2X}{X^2+Y^2+1}, \frac{2Y}{X^2+Y^2+1}, \frac{X^2+Y^2-1}{X^2+Y^2+1}\right)$$

Proof. We compute the chart, and leave the process of finding its inverse as an exercise. In fact, we can simplify things further by noting that the result must be some multiple of (x, y), as the line connecting (x, y, z) to $(0, 0, 1)$ in $\mathbb{E}^3$ is $(0, 0, 1)+t(x, y, z-1)$ which projects to (tx, ty) in the plane.

Thus, we can look at the 1D version of the problem, which is the projection of a circle onto the real line through its center, to figure out the scaling factor.

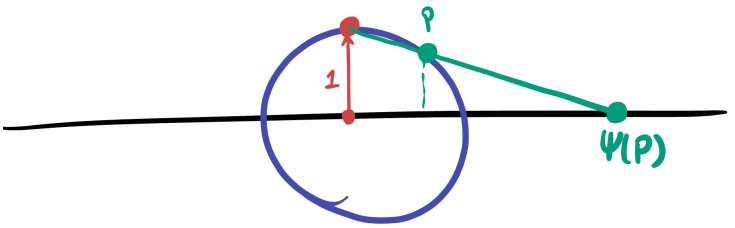


Figure 24.3.: A 1D slice of stereographic projection.

And now we have a problem purely in *Euclidean plane geometry*, where two similar triangles make an appearance. The result of the mapping takes our point (x, z) to a point distance L along the real line, so the right triangle with sides L and 1 is similar to the triangle with sides $L - x$ and z:

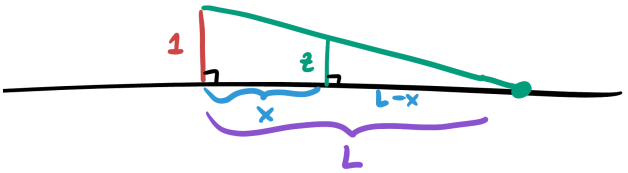


Figure 24.4.: Finding the projection point amounts to a calculation with similar triangles.

Equating the ratios of the sides gives

$$\frac{L-x}{z} = \frac{L}{1}$$

which simplifies to $L = \frac{x}{1-z}$. Thus, in general we have

$$\phi(x, y, z) = \left(\frac{x}{1-z}, \frac{y}{1-z}\right)$$

□

Exercise 24.1. Derive the formula for the parameterization associated to stereographic projection, by (1) like above, first focusing on 1 dimension, and then (2) starting with a point X along the line, solving for the intersection of the line connecting it with N and the sphere.

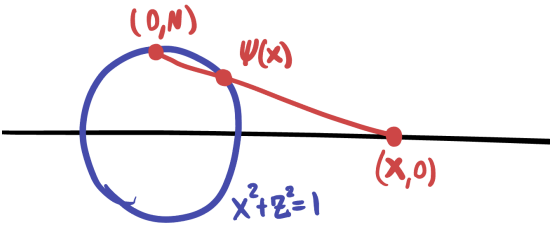


Figure 24.5.: The parameterization of stereographic projection, in a slice.

Figure 24.6.: Stereographic projection of the earth onto the plane.

This map is very simple algebraically: both the chart and parameterization are given by rational functions (quotients of polynomials). But its also simple *geometrically* in several particularly nice way, which we explore in the section below.

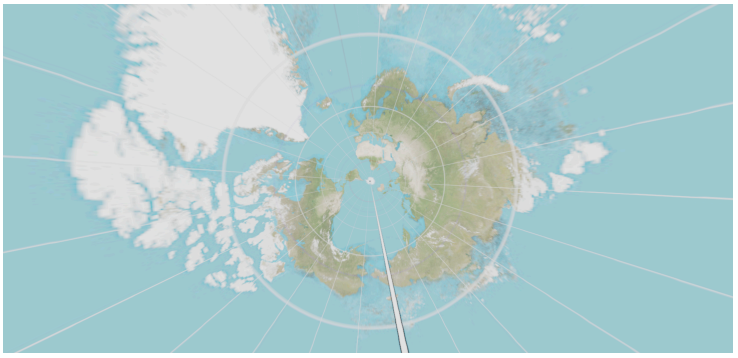


Figure 24.7.: Stereographic projection from a point near the north pole.

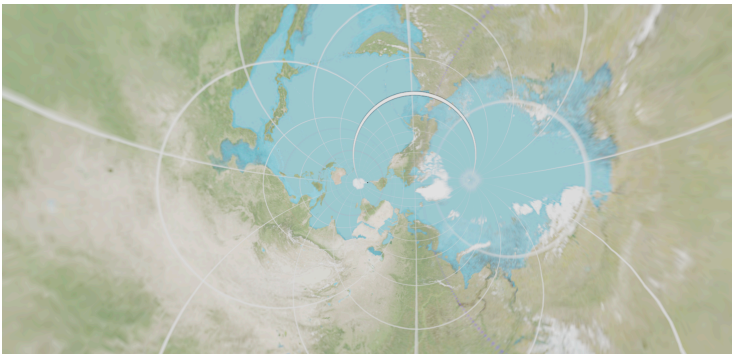


Figure 24.8.: Stereographic projection from a point in eurasia.

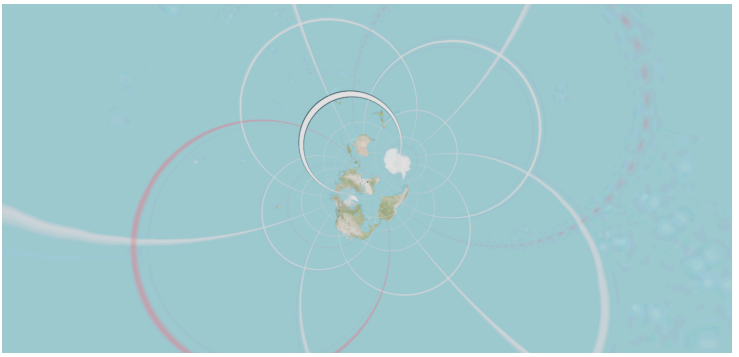


Figure 24.9.: Stereographic projection from a point in the pacific ocean far from any land.

24.1. Geometry of the Map

Example 24.1 (Equator sent to Unit Circle)**.** Stereographic projection sends the equator of ^{2} to the unit circle of $\mathbb{E}^2$.

We can see this geometrically, as the unit circle already lies in the plane $(x, y, 0)$ so *by definition* stereographic projection doesn't do anything to it! And, we can see it algebraically, by noting that the z coordinate of points on the equator is already zero, so

$$\phi(x, y, 0) = \left(\frac{x}{1-0}, \frac{y}{1-0}\right) = (x, y)$$

But this behavior extends beyond the equator to all lines of latitude of the sphere: they are all mapped to circles about the origin in $\mathbb{E}^2$!

Example 24.2 (Latitudes sent to Circles)**.** Let C be a circle on ^{2}, centered on the south pole S. Then stereographic projection maps C to a circle in $\mathbb{E}^2$ centered at the origin O.

To see this, note that the circles of ^{2} centered at one of the poles are contained within a horizontal plane, so the z-coordinate of all points in C is constant. Thus after applying ϕ we get

$$\phi(x, y, z) = \left(\frac{x}{1-z}, \frac{y}{1-z}\right) = (kx, ky)$$

Where $k = 1/(1-z)$ is a constant for all points x, y. Thus the result is just a *similarity* applied to the original curve C, which was a circle - and similarities take circles to circles!

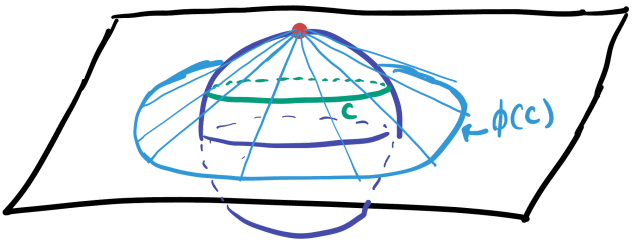


Figure 24.10.: Stereographic projection sends circles about S on ^{2} to circles about O in $\mathbb{E}^2$.

If we want to be even more precise, we could figure out exactly *which* circles in the plane they map to.

Exercise 24.2. Let C be the circle of radius r about the south pole S of ^{2}. Show that under stereographic projection the circle this maps to in $\mathbb{E}^2$ has Euclidean radius

$$\frac{\sin r}{1 + \cos r}$$

Hint: Show the circle centered at S of radius r lies in the plane $z = -\cos r$...then apply stereographic projection.

The other curves we can understand well are the great circles through the poles.

Example 24.3 (Great Circles through Projection Map to Lines)**.** Each great circle through the poles of ^{2} projects to a line through the origin.

To see this, we recall that great circles are all contained in a plane through the origin, and so a great circle through the poles is contained in a *vertical plane*. But the definition of stereographic projection involves drawing lines between N and points, and for points in a vertical plane, these lines also lie in that same vertical plane! Thus, the projection of this vertical plane is just its intersection with the horizontal plane, which is a straight line through the origin.

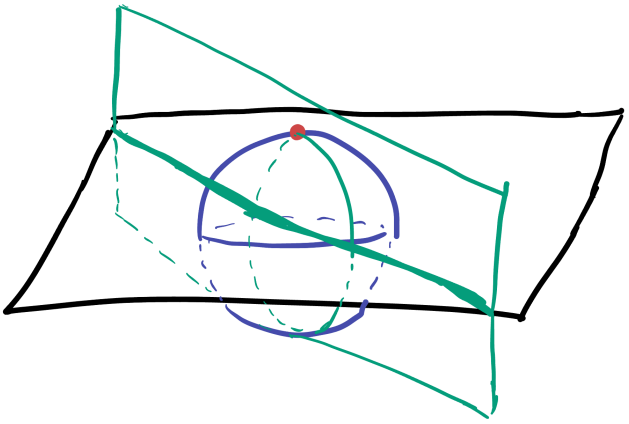


Figure 24.11.: Stereographic projection sends geodesics through the poles in ^{2} to geodesics (lines) through O in $\mathbb{E}^2$

This gives a nice grid on the plane

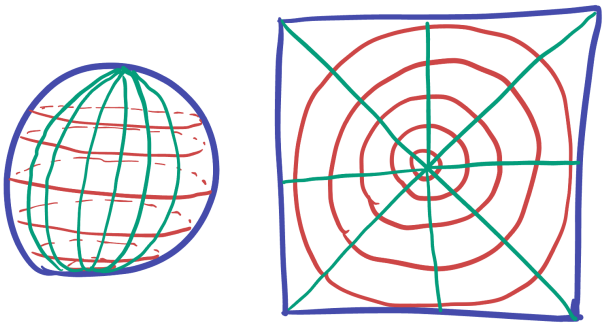


Figure 24.12.: The latitude/longitude grid on the sphere, stereographically projected onto the plane.

24.1.1. Generalized Circles

We saw already that great circles through the north pole get mapped to straight lines through the origin int the plane. But this does not mean that all geodesics map to lines, as the equator maps to the unit circle!

But its just just geodesics that map to circles either, we saw that circles around the north and south pole also map to circles in the map. It seems that circles and lines (geodesics) on the sphere are sent to circles and lines on the plane, but they might get *mixed up*. What a weird property! And one that's hard to state. So let's introduce a nice piece of terminology.

Figure 24.13.: Some circles project to circles, others to lines.

Definition 24.2 (Generalized Circle)**.** A *generalized circle* on the plane is a curve that is either (1) a Euclidean circle, or (2) a Euclidean straight line.

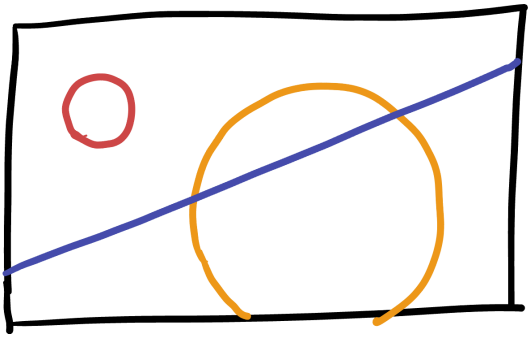


Figure 24.14.: Generalized circles in the Euclidean plane are the collection of all circles and lines.

Theorem 24.1 (Stereographic Projection Preserves Generalized Circles)**.** *Stereographic projection sends any circle on the sphere to a generalized circle on the plane.*

Proof. A circle on ^{2} is the intersection of the sphere with a plane (when this plane is through the origin, its a great circle, which is also a geodesic). So we are really interested in showing that stereographic projection maps the spots where a plane intersects the sphere to a circle in the plane.

This can be done geometrically, or by an algebraic computation. Here I'll give the algebra, and below I'll link to a beautiful visualization of the geometric proof. A plane in $\mathbb{E}^3$ is described by an equation of the form

$$ax + by + cz = d$$

But we can use ψ to express these x, y, z on the sphere in terms of (X, Y) on the plane:

$$\psi(X, Y) = (x, y, z) = \left(\frac{2X}{X^2 + Y^2 + 1}, \frac{2Y}{X^2 + Y^2 + 1}, \frac{X^2 + Y^2 - 1}{X^2 + Y^2 + 1} \right)$$

Plugging these in, we see

$$a\frac{2X}{X^2 + Y^2 + 1} + b\frac{2Y}{X^2 + Y^2 + 1} + c\frac{X^2 + Y^2 - 1}{X^2 + Y^2 + 1} = d$$

This looks bad with all the fractions, but we can clear denominators to get

$$2aX + 2bY + c(X^2 + Y^2 - 1) = d(X^2 + Y^2 + 1)$$

This still looks pretty bad, but its not really! Let's collect all the terms with x's and y's on one side, and group things with similar constants together.

$$(c - d)(X^2 + Y^2) + 2aX + 2bY = d + c$$

This is a quadratic equation where X^2 and Y^2 have the same coefficient! That means, this is a circle! (We just need to complete the square if we want to know its center and radius....)

Except.....when that coefficient is equal to zero (when $c = d$). Then there are no X^2 or Y^2 terms, and its a *linear equation*! So intersections of planes with the sphere either map to circles in the plane or to lines in the plane, as required. □

24.2. Infinitesimal Geometry

Our main goal of infinitesimal geometry is to show that stereographic projection is *conformal*: that angles are preserved, and all infinitesimal quantities are controlled by a single scaling factor.

Of course, one means of doing this is brute force calculus: just differentiate the parametrization and compute its action on infinitesimal squares. But with a map as nice as stereographic projection, one can avoid getting so messy with formulas and instead reason more *geometrically* as well. We shall pursue the geometric approach in this section, though I recommend you work out the calculus-only argument for practice.

The first thing we notice, from our dealings with lines of latitude and longitude above is that these originally orthogonal curves on the sphere are sent to two families of orthogonal curves in the plane. This implies that infinitesimal squares lined up with latitude and longitude to infinitesimal rectangles lined up with circles about 0 and lines through 0, and vice versa.

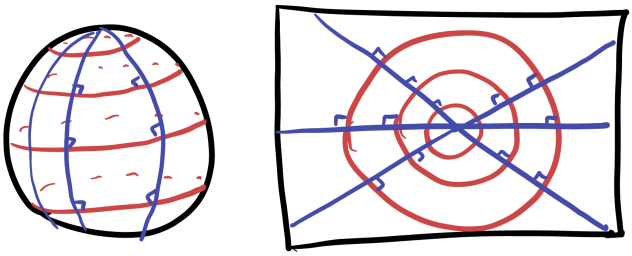


Figure 24.15.: Lines of longitude and latitude project to orthogonal curves on the plane.

To show that the overall map is conformal then, all we need to show is that such an infinitesimal square is stretched the same amount in each of these directions.

Proposition 24.2 (Infinitesimal Angle Length)**.** *At a point distance r from the south pole on ^{2}, the infinitesimal stretch along the circle of latitude containing this point is*

$$\frac{1}{1+\cos r}$$

Proof. A circle of radius r on the sphere has circumference $2\pi \sin r$. This projects to a circle on the plane of Euclidean radius $\frac{\sin r}{1+\cos r}$, and hence circumference $2\pi \frac{\sin r}{1+\cos r}$.

The ratio of lengths is the scaling factor: how much the length of the circle was increased or decreased by projection:

$$\frac{2\pi \frac{\sin r}{1+\cos r}}{2\pi \sin r} = \frac{1}{1+\cos r}$$

□

Proposition 24.3 (Infinitesimal Radial Length)**.** *At a point distance r from the south pole on ^{2}, the infinitesimal stretch along the line of longitude containing this point is*

$$\frac{1}{1+\cos r}$$

Proof. Here we need only take the derivative along any geodesic through S. One such geodesic is the great circle in the xz plane $\gamma(t) = (\sin t, 0, -\cos t)$ which passes through $(0, 0, -1) = S$ at $t = 0$. This maps under stereographic projection to

$$\phi(\gamma(t)) = \frac{\sin t}{1+\cos t}$$

Whose derivative measures the expansion rate of the geodesic as it is mapped onto the plane:

$$\begin{aligned} D(\phi \circ \gamma)_t &= \frac{\cos t(1+\cos t) - \sin t(-\sin t)}{(1+\cos t)^2} \\ &= \frac{\cos t + \cos^2 t + \sin^2 t}{(1+\cos t)^2} \\ &= \frac{1+\cos t}{(1+\cos t)^2} \\ &= \frac{1}{1+\cos t} \end{aligned}$$

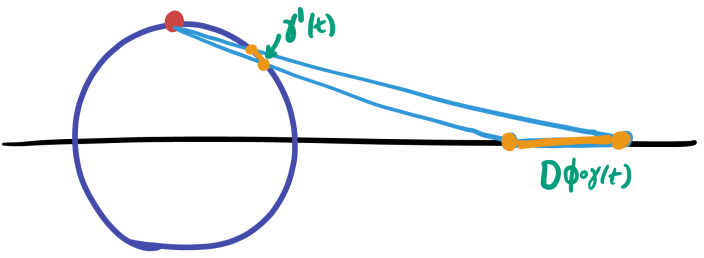


Figure 24.16.: Derivative in the radial direction.

Thus at distance r away from the center (so $t = r$ as the unit circle is a unit-speed curve), the rate of stretching rate is

$$\frac{1}{1+\cos r}$$

□

The two thereoms above tell us that at any point of the sphere, the latitude and longitude directions are *both stretched by the same factor*! This means that infinitesimal squares in ${T_p}^2$ are mapped to infinitesimal squares

Theorem 24.2 (Stereographic Projection is Conformal)**.** *Stereographic projection preserves angles: it sends infinitesimal squares to infinitesimal squares.*

Proof. At any point p, the stereographic projection chart the curve of longitude and latitude through p to orthogonal curves on the plane. It stretches each of these curves by the same factor, $1/(1+\cos r)$, meaning that it takes a unit square in ${T_p}^2$ to a square of side length $1/(1+\cos r)$ in $T_{\phi(p)}M$.

Thus, ϕ takes infinitesimal squares to other infinitesimal squares! This means (by the discussion in the angle chapter) that ϕ preserves all angles, so ϕ is a conformal map. □

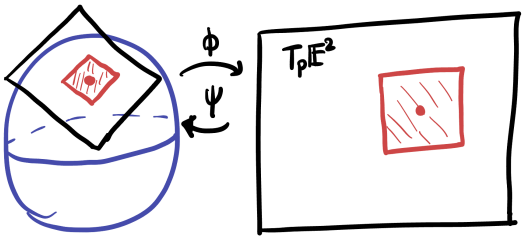


Figure 24.17.: Stereographic projection takes infinitesimal squares in the tangent space to other infinitesimal squares: that is, it preserves angle.

Now that we know that stereographic projection is conformal, we know it stretches all vectors by the same amount at a given point. Our calculations above confirmed this fact using the *chart*, but most interesting calculations we will want to do need the *parameterization*.

Exercise 24.3 (Stereographic Map-Coordinates)**.** The fact that stereographic projection is conformal means that at a given point $p = (X, Y)$ in $\mathbb{E}^2$, the parameterization

ψ must stretch all vectors by the same amount. By applying $D\psi$ to a vector, calculate this amount and show it to be

$$\frac{2}{1+X^2+Y^2}$$

Because all vectors are stretched in the same way, we can write down the *map dot product* easily: after a little calculation we see it is just a multiple of the Euclidean dot product on the plane!

Theorem 24.3 (Stereographic Dot Product)**.** *Let $p = (X, Y)$ be a point in the plane, and $v = \langle v_1, v_2\rangle$, and $w = \langle w_1, w_2\rangle$ be two vectors in $T_p\mathbb{E}^2$. Their map dot product is*

$$(v \cdot w)_{\text{map}} = (D\psi_p v) \cdot (D\psi_p w) = \frac{4}{(1+X^2+Y^2)^2}(v \cdot w)$$

Proof. Let $e_1 = \langle 1, 0\rangle$ and $e_2 = \langle 0, 1\rangle$ to help with notation. Then we can write v as a linear combination of this basis:

$$v = \langle v_1, v_2\rangle = v_1\langle 1, 0\rangle + v_2\langle 0, 1\rangle = v_1 e_2 + v_2 e_2$$

and similarly for w. Next, using the fact that the derivative is a *linear map* (its a matrix, after all) we can distribute, and pull out the scalars v_i and w_i:

$$\begin{aligned} D\psi_p(v) &= D\psi_p(v_1 e_2 + v_2 e_2) \\ &= D\psi_p(v_1 e_) + D\psi_p(v_2 e_2) \\ &= v_1 D\psi_p(e_1) + v_2 D\psi_p(e_2) \end{aligned}$$

and again, similarly for w. Now we wish to compute the dot product, so we multiply it all out:

$$\begin{aligned} (D\psi_p v) \cdot (D\psi_p w) &= (v_1 D\psi_p e_1 + v_2 D\psi_p e_2) \cdot (w_1 D\psi_p e_1 + w_2 D\psi_p e_2) \\ &= v_1 w_1 (D\psi_p e_1) \cdot (D\psi_p e_1) + v_1 w_2 (D\psi_p e_1) \cdot (D\psi_p e_2) \\ &+ v_2 w_1 (D\psi_p e_2) \cdot (D\psi_p e_1) + w_1 w_2 (D\psi_p e_2) \cdot (D\psi_p e_2) \end{aligned}$$

Now we have to think a bit about what we know! Since ψ does not change angles, and e_1 and e_2 are orthogonal, we know that $D\psi_p e_1$ and $D\psi_p e_2$ are orthogonal to one another. Thus, their dot product is equal to zero! This gets rid completely of the two middle terms in our big expression, and so we are only left with the first term and the last term.

But both of these are involve the dot product of a vector with itself: like $(D\psi_p e_1) \cdot (D\psi_p e_1)$. This is the definition of the *squared length* of that vector, but we know exactly how stereographic projection changes lengths! Thus for both e_1 and e_2 we know

$$(D\psi_p e_i) \cdot (D\psi_p e_i) = \left(\frac{2}{1 + X^2 + Y^2}\right)^2 = \frac{4}{(1 + X^2 + Y^2)^2}$$

Now we putting this all together, we find

$$\begin{aligned}(v \cdot w)_{\text{map}} &= \frac{4}{(1 + X^2 + Y^2)^2} v_1 w_2 + 0 v_1 w_2 + 0 v_2 w_1 + \frac{4}{(1 + x^2 + y^2)^2} v_2 w_2 \\ &= \frac{4}{(1 + X^2 + Y^2)^2} (v_1 w_2 + v_2 w_2) \\ &= \frac{4}{(1 + X^2 + Y^2)^2} (v \cdot w)\end{aligned}$$

□

Definition 24.3 (Stereographic Metric)**.** The dot product on $\mathbb{E}^2$ given at the point $p = (X, Y)$ by

$$(v \cdot w)_{\text{map}} = \frac{4}{(1 + X^2 + Y^2)^2} (v_1 w_1 + v_2 w_2)$$

Is called the *stereographic metric*.

Using this, we can compute any quantity we may care about on the sphere, using only coordinates of the plane. For instance, the spherical length of a vector is just

$$\|v\|_{\text{map}} = \sqrt{(v \cdot v)_{\text{map}}} = \frac{2}{1 + X^2 + Y^2} \|v\|$$

The area of an infinitesimal piece of the sphere is

$$dA = (\langle 1, 0 \rangle \cdot \langle 0, 1 \rangle)_{\text{map}} \, dX dY = \frac{4}{(1 + X^2 + Y^2)} dX dY$$

etc.

24.2.1. The Disk and Half Plane

One use of stereographic projection is to write down a map of the sphere, as we've seen above. But it is also used a lot in mathematics as a *tool* to help create new and useful functions that would otherwise be difficult to guess. It shows up in this context in applications across geometry, complex analysis, and other fields of math because its *conformal*, and so we know when building things with stereographic projection as one of the components, it is not going to mess up any angle measures.

Here, we will look at a fundamental example of this, and will use stereographic projection to write down a *conformal map* which takes points in the unit disk $x^2 + y^2 < 1$ to the upper half plane $y \geq 0$ in $\mathbb{E}^2$. This map becomes very important in the study of hyperbolic geometry, where we can use it to help us relate two different maps of the mysterious hyperbolic plane.

Here's the idea: starting with the unit disk in the plane centered at O, we can use the *parameterization* of stereographic projection to map this onto the sphere. Doing so moves the region onto the entire southern hemisphere of ^{2} (since the unit circle maps to the equator):

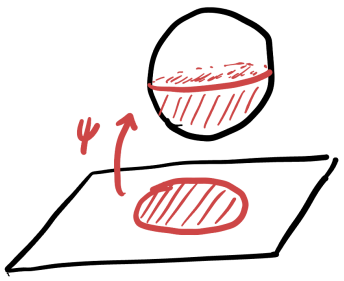


Figure 24.18.: Mapping the unit disk to the lower hemisphere of ^{2} via the parameterization ψ.

Now, we can rotate the sphere a quarter of a turn about the x-axis, so that the equator becomes a line of longitude, now passing through the north and south pole, and what was the southern hemisphere is now the *positive y hemisphere*: the south pole has been moved to $(0, 1, 0)$. (How do we write down a rotation about the x axis? Well, its going to *fix* the x direction so we know the first column. Then, the second and third column will be where the y and z basis vectors go under a quarter turn)

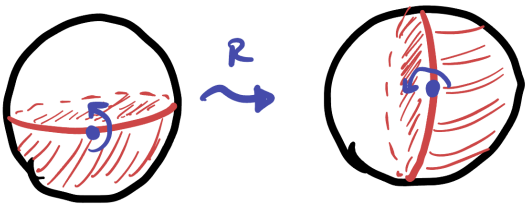


Figure 24.19.: Rotating the sphere about the x axis by a quarter turn takes the lower hemisphere to the hemisphere of positive y.

And finally, we can use the *chart* of stereographic projection to re-project this down onto the plane. Now, the great circle bounding it passes through the north pole, so it projects to a line: the x-axis! This divided the sphere in two, and so its image divides the plane in two, with our *positive y hemisphere* becoming the *positive y half-plane.*

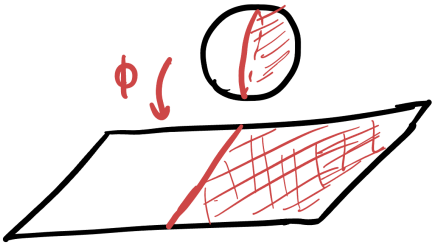


Figure 24.20.: Projecting the hemisphere of positive y to the plane with ϕ gives the half plane with positive y.

Exercise 24.4 (Disk and Half Plane: Construction)**.** Let $\mathbb{D}$ be the unit disk $\mathbb{D} = \{(x, y) \mid x^2 + y^2 < 1\}$ and let $\mathbb{U}$ be the upper half plane $\mathbb{U} = \{(x, y) \mid y > 0\}$. Let $T : \mathbb{D} \to \mathbb{U}$ be the map described above. Prove that $T can be expressed as

$$T(x, y) = \left(\frac{2x}{1 + x^2 + y^2 - 2y}, \frac{1 - x^2 - y^2}{1 + x^2 + y^2 - 2y} \right)$$

By building it step by step: applying ψ to get the disk onto the sphere, rotating by the correct quarter turn about the x-axis, and then applying ϕ to return to the plane.

This map is *conformal* - meaning that it preserves all angles! And even more than that, it takes *generalized circles to generalized circles.*

Exercise 24.5 (Disk and Half Plane: Understanding)**.** Prove that these claims are in fact true: that our new function is conformal, and sends generalized circles to

generalized circles. *Hint: what kinds of maps is it built out of? What do each of these maps to do angles, or to generalized circles (on the plane) / circles (on the sphere)?*

Use this to "transfer" this picture of polar coordinates in the unit disk onto the plane, via our new map.

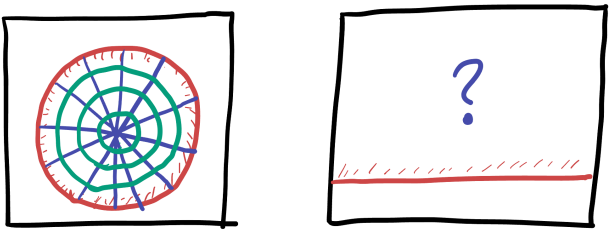


Figure 24.21.: What do these generalized circles look like when mapped to the half plane?

24.3. The Sphere of Radius R

Throughout this chapter we have studied stereographic projection in detail, but on the *unit sphere*. It is not too hard to generalize what we have done to spheres of other radii, and while this may not sound super exciting at first, it actually turns out to be absolutely fundamental to how we are going to discover hyperbolic space! So, it is a rather important exercise to work this all out for yourself.

The good news is you have this entire chapter as a guide, where I've worked out many of the details for the case of the unit sphere. The formulas will be quite similar, but there'll be R's inserted in various places: so the second piece of good news is that I'll give you the formulas that you need to derive! That way, you can check your work.

Definition 24.4. Let $\overset{2}{R}$ be the sphere of radius R in $\mathbb{E}^3$. Then the chart ϕ for stereographic projection of this sphere is defined geometrically exactly as in the original version: given a point $p \in \overset{2}{R}$, $\phi(p)$ is where the line connecting p to the north pole $N = (0, 0, R)$ intersects the xy plane.

Exercise 24.6. Show that the formulas for both the chart and the parameterization of stereographic projection here are as follows:

$$\phi(x, y, z) = (X, Y) = \left(\frac{Rx}{R - z}, \frac{Ry}{R - z}\right)$$

$$\psi(X, Y) = (x, y, z) = \left(\frac{2R^2X}{X^2 + Y^2 + R^2}, \frac{2R^2Y}{X^2 + Y^2 + R^2}, R\frac{X^2 + Y^2 - R^2}{X^2 + Y^2 + R^2}\right)$$

(It might help to look back at Proposition 24.1, and attempt Exercise 24.1).

Running through the same arguments as in the chapter above (which you don't have to write down), its straightforward to check that this new map is a conformal map between $\overset{2}{R}$ minus N, and the plane. This means its parameterization ψ both preserves angles and stretches all vectors by a uniform length: we can use this fact to compute the dot product for this map.

Exercise 24.7. At a point $p = (X, Y)$ on the plane, what is the factor by which a vector $v \in T_p\mathbb{E}^2$ is stretched when mapped onto $\overset{2}{R}$ by the parameterization of stereographic projection? *Hint: we know the factor is the same for all vectors: so pick an easy vector to calculate with and find its length*!

Once you know this, follow the argument style of Theorem 24.3 to compute the map-dot product on the plane, and show that it is equal to

$$(v \cdot w)_{\text{map}} = \frac{4R^4}{(R^2 + X^2 + Y^2)^2}(v \cdot w)$$

Part VI.

Hyperbolic Space

25. Discovery

In this chapter we will embark on our final journey in the *Foundations of Geometry*, and construct the third flavor of geometry: hyperbolic space.
To do so, we will draw on all of our knowledge from the course, from the Greeks, through calculus, to spherical geometry and its expression in maps.

The discovery of hyperbolic geometry was one of the big mathematical achievements of the previous millennium, as not only did it open our minds to the much richer geometric world mathematicians now work in, but it also definitively answered the most important outstanding question of the Greeks. So, it is only fitting, that in this final chapter we look back to where we began.

25.1. Prelude: The Legacy of the Greeks

The ancient Greeks were right to feel proud of their progress in geometry: after a couple of centuries of deep thought they mastered the axiomatic method, and went forth to prove hundreds upon hundreds of theorems answering almost every question they had about the geometry of 2 and 3 dimensional space.

Almost.

The golden age of Greek mathematics closed with a few important open problems remaining, that they posed to the generations of the future. Three of these were questions about the specifics of greek method of *constructing* geometric figures using a ruler and compass:

- **Doubling the cube:** given a cube, use a ruler and compass to construct a new cube with double the volume.
- **Squaring the Circle:** Given a circle, use a ruler and compass to construct a square with the same area as the circle.
- **Trisecting angles:** Given an arbitrary angle, use a ruler and compass to draw two new lines which divide it evenly in thirds.

The deep knowledge of greek mathematicians becomes even more clear as we look back at these problems as guiding lights for mathematics over the last two thousand years. While today all have been solved, it took until the 1800's, the answers required an entirely new branch of mathematics, and they were all answers the greeks and their successors never imagined:

- Doubling the cube with a ruler and compass is *impossible*.
- Squaring the circle with a ruler and compass is *impossible*.
- Trisecting an arbitrary angle with a ruler and compass is *impossible*.

Unfortunately: we will not answer these three in this course. It turns out the new idea needed to answer all three of these questions came from *abstract algebra* - specifically, from the theory of fields, and field extensions! So while the *questions* are pure geometry, their *resolutions* are arguments in algebra! If you are currently in abstract algebra - please feel free to come ask me about them!

We will instead focus our attention on the much larger problem the greeks left open: not a problem about some particular method of drawing geometric figures, but about the nature of geometry itself.

- Prove the fifth postulate from the remaining four (or show that this is impossible).

This question was worked on for approximately two thousand years by the greatest mathematical minds, with essentially no success until its eventual resolution 200 years ago (around 1823). It is one of the great joys of undergraduate mathematics to be able to seriously engage with the mathematics of the intervening two thousand years, and be able to fully understand the answers to these lasting questions, and this course has essentially been designed to exposit this problem's solution. So, lets begin!

25.1.1. Disproof by Counterexample

Perhaps, like the other questions of antiquity, the reason the greeks failed to prove the fifth postulate from the other four is that this is also *impossible*. One way to show a claim is impossible is by finding a counterexample, and that is the approach that we will take here. But a counterexample to what? If the question is to "prove Postulate V from postulates I-IV", then a counterexample would be a geometry which satisfies the first four axioms, but for which the fifth is false. The existence of such a geometry would spell doom to any hope of proving the fifth in the same way that the number 4 spells doom to any hope to proving that all numbers greater than 1 are prime. It would once and for all settle the question, and in the same style as the other greek resolutions: it seems that when it comes to geometry, the only things the greeks *didn't* succeed at were impossible!

But while the dream of producing a counterexample is clear, the actual process for doing so is not. While its easy to write down a geometry in our modern definition (you just need to give me a set of points, and rules for doing calculus with infinitesimal arc lengths and angles at each point), most everything you will write down will *not* satisfy all four of the first axioms of the greeks.

The first and third axioms basically state that "space doesn't have any holes, tears, rips or edges in it": you can always connect two points with a line segment and once you have a line segment you can always sweep it around in a circle.

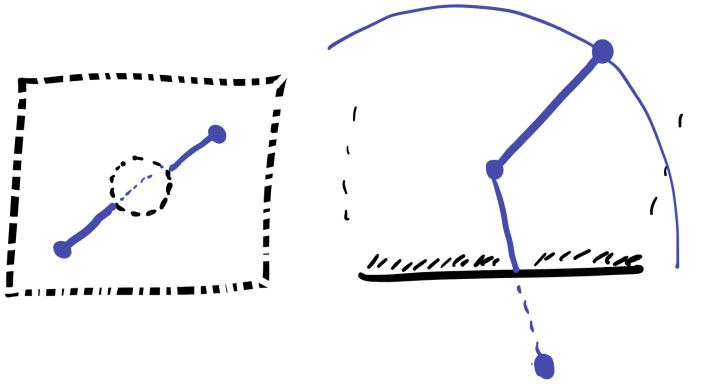


Figure 25.1.: Spaces with holes or edges violate postulates 1 or 3.

The second axioms says something about space being *infinite*: we already saw that this fails on finite spaces like the sphere. But its more particular than that - it can also fail on some infinite spaces like the cylinder, where geodesics can be extended indefinitely in one direction, but not in others. For axioms 2 to be satisfied in the way the Greeks originally stated it, space needs to be "infinite in all directions".

A more modern reading of this axiom by Bernhard Riemann in the 1800s replaces it with a weaker statement that only says geodesics can't come to an "edge" you have to be able to continue forwards forever, but its OK if you retrace your steps. This allows the sphere to be worked with more naturally in the context of axiomatic geometry.

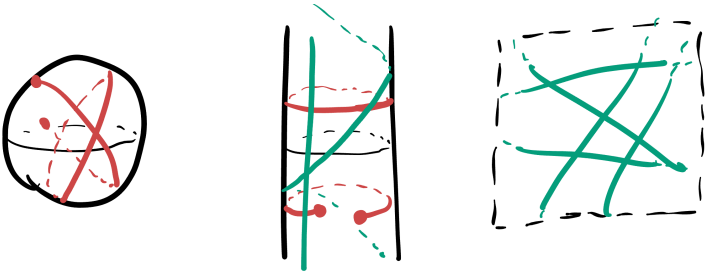


Figure 25.2.: Spaces with geodesics of finite total length violate Postulate 2, as originally written.

The axiom that is *hard to satisfy* is the fourth: *all right angles are equal.* Remember that when making this precise, we realized it implies that space is both *homogeneous*

and isotropic: you can move any point to any other, and you can rotate about any point, in any direction. This axiom is the key to a lot of what we do when proving things: we always move something to the origin, or to the north pole to simplify our calculations, and justify it by our proofs that the sphere and plane are homogeneous. But any little change whatsoever to the sphere or plane generally produces something that *is not homogeneous*:

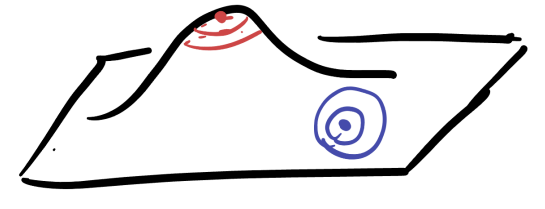


Figure 25.3.: Spaces that are not homogeneous and isotropic violate Axiom 4. This is the hardest one to satisfy: even a single little bump on the infinite plane can mess it all up.

Thus, its relatively easy to produce a geometry which satisfies axioms 1,2 and 3 but fails 4, and its also possible (but harder) to produce a geometry which satisfies 1,3, and 4, but fails 2 (the sphere). Unfortunately neither of these help us in our quest for a counterexample. Even though we proved on the sphere that the parallel postulate is *false* (we found a triangle with three 90 degree angles: so the angles on one side added up to 180 yet the lines still intersected!), this does nothing towards our goal, as it didn't satisfy the other four to begin with!

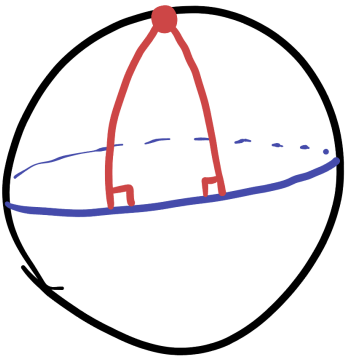


Figure 25.4.: The sphere violates the parallel postulate, but this does not solve the greeks problem as it also violates postulate 2.

To build our counterexample, we want to take as much inspiration from the sphere

and the plane as possible - since these are the only two geometries that we know of that satisfy Postulate 4. But we want our new creation to be infinite like Euclidean space, and yet fail the parallel postulate like the sphere.

25.2. A Radical Idea

While we often speak as though we have studied two geometries so far - Euclidean and spherical - this actually isn't quite right. Indeed, we proved along the way that there is actually a *different* sphere geometry for each radius! More precisely, we've studied the Euclidean plane, and a whole 1-parameter family of spherical geometries

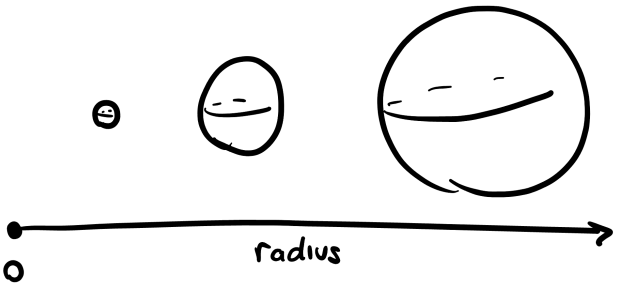


Figure 25.5.: A spherical geometry for each radius $R \in (0, \infty)$.

The sphere geometries are all closely related to one another, (by a similarity that is not an isometry) and so their geometric formulae are all closely related as well. This allowed us to start with careful study of the unit sphere and then expand our knowledge to spheres of any other radius: for example, deriving the formulas for circumference and area of circles on a sphere of radius R:

$$C(r) = 2\pi R \sin\left(\frac{r}{R}\right)$$

$$A(r) = 2\pi R\left(1 - R\cos\left(\frac{r}{R}\right)\right) = 4\pi R^2 \sin^2\left(\frac{r}{2R}\right)$$

Exercise 25.1. Derive this formula for the area of a circle on the sphere of radius R: use the fact that you found the circumference on a previous homework, and that area is teh *integral of circumference*

$$A(r) = \int_0^r C(r)\, dr$$

Then apply a trigonometric identity to simplify things and get the second expression.

The same ideas apply to other important formulas in geometry like the spherical pythagorean theorem: you found this on the homework to be $\cos(c) = \cos(a)\cos(b)$ on the unit sphere, which generalizes on the sphere of radius R to

$$\cos\left(\frac{c}{R}\right) = \cos\left(\frac{a}{R}\right)\cos\left(\frac{b}{R}\right)$$

And the area of a spherical triangle, which is $A = \alpha + \beta + \gamma - \pi$ on the unit sphere, and

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

on the sphere of radius R. This continues for the spherical trigonometry of right triangles. Consider the triangle with angles α, β, opposite sides a, b and hypotenuse c. On the unit sphere we had the relations $\sin\alpha = \frac{\sin a}{\sin c}$ and $\cos\alpha = \frac{\tan b}{\tan c}$. Thus, on the sphere of radius R we have

$$\sin\alpha = \frac{\sin\frac{a}{R}}{\sin\frac{c}{R}} \qquad \cos\alpha = \frac{\tan\frac{b}{R}}{\tan\frac{c}{R}}$$

Then rather separately, we had the geometry of the Euclidean plane, where we spent the first half of the semester carefully confirming all of the results we knew from earlier education:

$$C_{\mathbb{E}^2}(r) = 2\pi r \qquad A_{\mathbb{E}^2}(r) = 2\pi r^2$$
$$c^2 = a^2 + b^2$$
$$\sin\alpha = \frac{a}{c} \qquad \cos\alpha = \frac{b}{c}$$

These formulas are different than those of the sphere, but not wholly so. Indeed - we saw earlier a hint at the connection, by recovering the Euclidean pythagorean theorem from the spherical one, as the radius of the sphere goes to infinity. But this wasn't an accident: if you take $\lim_{R\to\infty}$ in any of the spherical formulas above, you'll recover the Euclidean counterpart!

:::{#rem-euc-triang} Here of course there's no analog of the formula giving a triangles area in terms of its angle sum, as in Euclidean space the angles do not determine the size of a triangle at all!:::

Exercise 25.2 (Circle Area and the Euclidean Limit:)**.** Let $C_R(r) = 2\pi(1 - R\cos(r/R))$ be the area of a circle of radius r on $\frac{2}{R}$. Prove that as $R \to \infty$ this converges to the Euclidean formula

$$\lim_{R\to\infty} C_R(r) = \pi r^2$$

Exercise 25.3 (Trigonometry and the Euclidean Limit)**.** Show that in the limit $R \to \infty$, the trigonometric formulas for spherical right triangles converge to their Euclidean counterparts.

$$\lim_{R\to\infty} \frac{\sin\frac{a}{R}}{\sin\frac{c}{R}} = \frac{a}{c}$$

$$\lim_{R\to\infty} \frac{\tan\frac{b}{R}}{\tan\frac{c}{R}} = \frac{b}{c}$$

This is even true for the maps we've studied! For example you're calculating the scaling factor for stereographic projection on the homework, and finding the map-dot product on the plane:

$$(v \cdot w)_{\text{map}} = \frac{4R^4}{(R^2 + X^2 + Y^2)^2}(v \cdot w)$$

Thus, this tells us how to rescale the Euclidean dot product $v \cdot w$ depending on our location $(x, y) \in \mathbb{E}^2$, so that the result accurately reflects teh *true dot product* on the sphere. But when $R \to \infty$, this true dot product becomes the Euclidean plane!

Exercise 25.4 (Maps and the Euclidean Limit)**.** Show that as $R \to \infty$, the scaling factor of stereographic projection becomes a *constant.* Thus, the map now treats vectors based at different points all the same - just like Euclidean geometry! (In fact, the result *is* Euclidean geometry).

Because this all fits together so nicely, we might picture the geometries we have discovered so far as forming a line (the spheres) with teh Euclidean plane off at infinity.

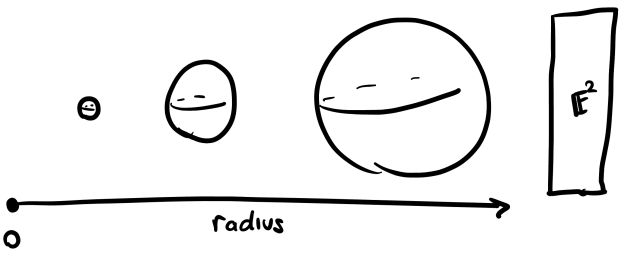


Figure 25.6.: Euclidean geometry is the limit of spherical geometry as $R \to \infty$.

This picture suggests a radical idea: as we increase the radius of the sphere, we keep our geometry nice and homogeneous but make it larger: getting closer to satisfying

postulate 2. Once we go all the way to $R = \infty$ we reach Euclidean space, where Postulate 2 is satisfied! What if we *kept going*? Can we go beyond Euclidean space, pushing the radius past ∞, and uncover new spaces, where Postulate 2 is true, but - like the spheres - the parallel postulate is false?

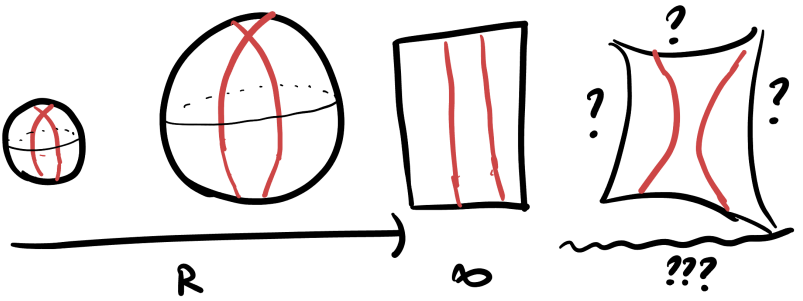


Figure 25.7.: If we can find a way to *keep going*, can we find a new geometry?

25.2.1. To Infinity and Beyond

Theres one obvious problem with our grand plan however: what could it possibly mean to go beyond infinity? We can't even go *to* infinity rigorously - everything must be done in terms of limits. We want to use the formulas we've worked hard to develop this semester as a guide for whatever new geometry lies on the other side, but the formuls are no good if the first step is "plug in a number larger than ∞" when in fact there are no such things.

What we need is a change in perspective. We would like to replace our discussion of $\lim_{R\to\infty}$ with something we can evaluate at a *finite* number. And, curvature provides the key. When studying spherical geometry, we proved there was an inverse relationship between radius and curvature: the bigger a sphere the less it was curved, and the smaller the sphere the larger its curvature. More precisely we showed that

$$\kappa = \frac{1}{R^2}$$

Thus, we can go to any of our formulas above, and directly replace any occurrences of R^2 with κ without changing any of the math. For example, the circumference $C_R(r) = 2\pi R \sin(r/R)$ becomes

$$C_\kappa(r) = \frac{2\pi}{\sqrt{\kappa}} \sin(\sqrt{\kappa} r)$$

where we've re-arranged the relationship of radius and curvature above to be able to substitute $R = \frac{1}{\sqrt{\kappa}}$.

Exercise 25.5. Give the analogs of other spherical geometry formulas in terms of curvature.

The euclidean limit now is no longer $R \to \infty$, but rather $\kappa \to 0$: and it now makes perfect conceptual sense to ask *what happens if we keep going*? We should just get $\kappa = 0$, and then - pushing onwards - get *negative values* of κ!

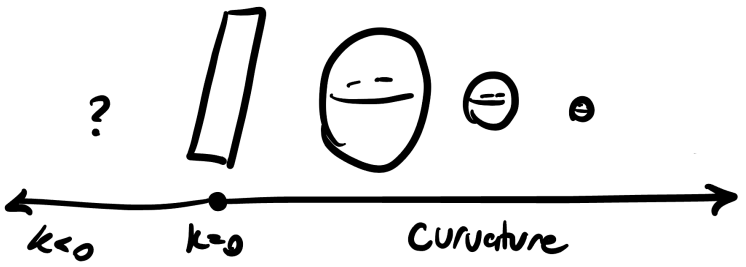


Figure 25.8.: Expressing things in terms of curvature takes the $R = \infty$ limit to $\kappa = 0$, and the mysteries beyond infinity to the familiar world of negative numbers.

There's just one problem with this: it doesn't seem to work when we look at our formula. Indeed, when $\kappa = 0$ the circumference is an indeterminate form $0/0$ (since $\sin(0) = 0$), and when $\kappa < 0$ we are trying to plug a negative number into a square root. But this is not as serious of a problem as it seems at first. We've expressed this function in one way, using the function sin which we invented for the study of circles in Euclidean geometry. But mathematics doesn't care how we write things down in symbols - those are human inventions. And we can write the exact same function using a *series expansion*, since right after our definition of sin we derived its series $\$\sin x = x - x^3/3! + x^5/5! - \cdots$

$$C_\kappa(r) = \frac{2\pi}{\sqrt{\kappa}}\left(\sqrt{\kappa}r - \frac{1}{3!}(\sqrt{\kappa}r)^3 + \frac{1}{5!}(\sqrt{\kappa}r)^5 - \cdots\right)$$
$$= \frac{2\pi}{\sqrt{\kappa}}\left(\sqrt{\kappa}r - \frac{1}{3!}\sqrt{\kappa}^3 r^3 + \frac{1}{5!}\sqrt{\kappa}^5 r^5 - \cdots\right)$$

We can simplify this by noting that every single term in the parentheses contains a $\sqrt{\kappa}$, and so we can safely factor one out, cancelling its counterpart in the denominator of 2π:

$$= 2\pi\left(r - \frac{1}{3!}\kappa r^+ \frac{1}{5!}\kappa^2 r^5 - \cdots\right)$$

This formula is *exactly equal to* the formula we started with, just expressed in another notation. BUt this other notation is much more suggestive when we are looking to be bold, and think about what happens if we explore beyond the original $\kappa > 0$ regime

in which we derived it. Indeed - this formula is completely well-defined for all κ! So we may take κ zero, and we recover directly $C_0(r) = 2\pi r$ - Euclidean geometry!

What happens when κ is negative? Well, if we look at the formula, we see that each term is multiplied by an

Half of the terms (those with odd powers of κ) change sign! This makes ALL the signs positive!! Here we've written $\kappa = -|\kappa|$ and used $|\kappa|$ in the formula to cancel these signs and emphasize that everything is positive

$$C_\kappa(r) = 2\pi(r + \frac{1}{3!}|\kappa|r^3 + \frac{1}{5!}|\kappa|^2 r^5 + \cdots)$$

The case we will be most interested in is when we push κ all the way to negative 1. Here the circumference function is just

$$C(r) = 2\pi(r + \frac{1}{3!}r^3 + \frac{1}{5!}r^5 + \cdots)$$

What is this new function like, qualitatively? In making all the terms the same sign, we prevent all the cancellation that happens in the series expansion for sin and cos that keeps them bounded in size. Instead, this function gets *larger* with each term we add, in the end growing faster than any polynomial! This means *if* there really is some space for which this is the circumference formula, circles here would be very very large. Since smaller-than-Euclidean circles signifies positive curvature, bigger-than-Euclidean circles hints that the space will have negative curvature. But we can do better than that! We have an exact, quantitative formula for curvature involving circle's circumference, after all.

Exercise 25.6. Let $k < 0$. Prove that a space with the circumference function

$$2\pi(r + \frac{1}{3!}|\kappa|r^3 + \frac{1}{5!}|\kappa|^2 r^5 + \cdots)$$

has curvature κ, by computing the limit

$$\lim_{r \to 0^+} \frac{3}{\pi} \frac{2\pi r - C(r)}{r^3}$$

This tells us that if we construct a space whose circumference-to-radius-function is as above, then this space necessarily has curvature κ. Doing so will create for us a space of every possible value of negative curvature, just as varying the radius of a sphere produces a space of every possible positive curvature. Such spaces are called *hyperbolic*

Definition 25.1 (Hyperbolic Geometry)**.** Given a negative number κ, *hyperbolic geometry of curvature κ is denoted $\mathbb{H}^2_\kappa$, and is the space where at every point, the circumference of circles grows as $C_\kappa(r)$. When $\kappa = -1$, we often shorten this to just* hyperbolic geometry* and denote it by $\mathbb{H}^2$.

Remark 25.1. It's reasonable to wonder if this uniquely defines a space: after all, all we have done is say how the circles (and thus, the curvature) behaves. This is a good thing to think about, and it was resolved in 1839 by *Minding's Theorem* which states that if two spaces have the same constant curvature, then they are locally isometric to one another.

To learn about this new *hyperbolic space*, we will push this technique as far as we possibly can: starting with a formula on the sphere, and pushing it to its limit (infinite radius, zero curvature) and beyond to the realm of negative curvature. In doing this we'll see the function we met above - like the sine but with all the terms positive, is actually just the tip of the iceberg. There is a whole collection of *parallel universe trigonometry functions* out there, defined analogously. And since these functions will play a large role in our further development, now is a good time to pause and get ourselves acquainted.

25.3. Interlude: Hyperbolic Functions

Definition 25.2 (Hyperbolic Sine and Cosine)**.** The hyperbolic sine and cosine function are defined by their series expansions, which are identical to those for sin and cos, except that the sign of all coefficients have been made positive:

$$\sinh(x) = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} \cdots$$

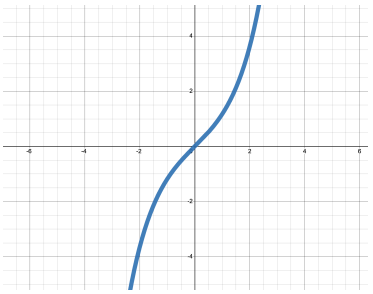
$$\cosh(x) = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} \cdots$$

Remark 25.2. The function name cosh is pronounced as it is spelled, but sinh is pronounced *sinch* in the US and *shine* in the UK.

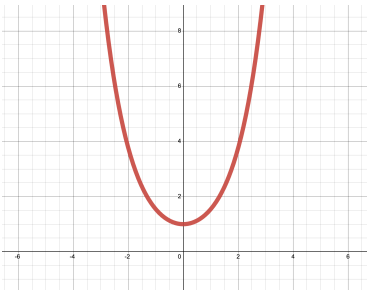
To begin to develop intuition for these functions, we should take a look at their graphs. The first major revelation is that unlike the usual trigonometric functions, neither are periodic!

Also, they seem to grow *extremely fast.* How fast? We can actually quantify it by relating these functions to exponentials. Since sinh contains all the odd terms of the series for e^x and cosh contains all the even terms, we can see that

$$\sinh(x) + \cosh(x) = e^x$$



(a) The Hyperbolic Sine



(b) The Hyperbolic Cosine

Exercise 25.7 (Hyperbolic Functions in terms of Exponentials)**.** Show that we can actually express the hyperbolic trigonometric functions in terms of exponentials:

$$\sinh(x) = \frac{e^x - e^{-x}}{2} \qquad \cosh(x) = \frac{e^x + e^{-x}}{2}$$

Because e^{-x} quickly becomes small, this gives us a very precise understanding of just how quickly sinh and cosh grow. For $x >> 1$ we have

$$\cosh(x) \approx \sinh(x) \approx \frac{e^x}{2}$$

In direct analogy with regular trigonometry, starting from sinh and cosh we can build up a family of other hyperbolic trigonometric functions: (The first two of these are often pronounced *tanch* and *coth*. I am too scared to attempt to pronounce the second two)

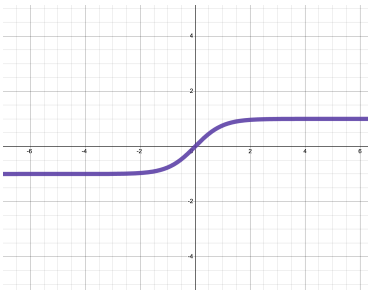
Definition 25.3 (Other Hyperbolic Functions)**.**

$$\tanh x = s\frac{\sinh x}{\cosh x} \qquad \coth x = \frac{\cosh x}{\sinh x}$$

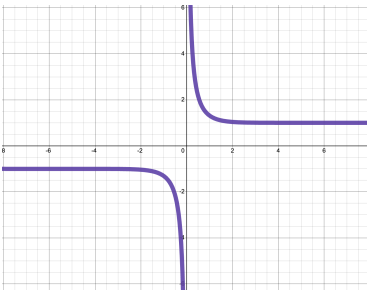
$$\operatorname{sech} x = \frac{1}{\cosh x} \qquad \operatorname{cschs} x = \frac{1}{\sinh x}$$

Knowing what some of these look like as well will be important, so let's take a minute to think about their graphs. How should tanh behave? Well, both sinh and cosh are growing exponentially at the same rate, so their ratio should be approximately 1 for large inputs. And, sinh is odd whereas cosh is even, so the quotient is an odd function: passing through 0 with horizontal asymptotes at ± 1!

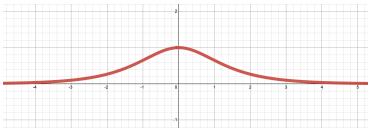
Thus, $\coth(x) = 1/\tanh(x)$ should also asymptote to ± 1, but always be *larger* in absolute value, diverging to $\pm\infty$ at the origin.



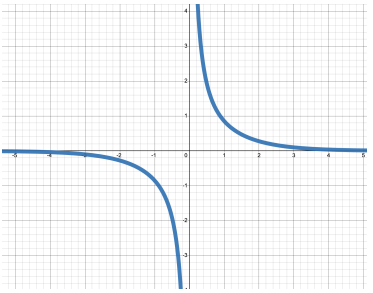
(a) The Hyperbolic Tangent



(b) The Hyperbolic Cotangent



(a) The Hyperbolic Secant



(b) The Hyperbolic Cosecant

The reciprocals sech and csch of cosh and sinh share similar behavior, both asymptoting to 0, while the hyperbolic secant stays bounded in $[0, 1]$ and the hyperbolic cosecant diverges at the origin.

These functions satisfy many analogous identities to the standard trigonometric ones as well, which can be proved directly from their definition:

Exercise 25.8 (Hyperbolic Trigonometric Identities)**.** Prove that

$$\cosh^2(x) - \sinh^2(x) = 1$$

Use this to deduce that

$$\operatorname{sech}^2(x) + \tanh^2(x) = 1$$

These identities are the same as the Euclidean versions but with the sign switched! *Hint: the formula relating to exponentials!*

Exercise 25.9. Derive this trigonometric identity:

$$\sinh(2x) = 2\sinh(x)\cosh(x)$$

For the hyperbolic functions it turns out that calculus is even simpler than in the Euclidean case:

Exercise 25.10 (Hyperboloic Trigonometric Derivatives)**.** Prove that

$$\frac{d}{dx}\cosh(x) = \sinh(x) \qquad \frac{d}{dx}\sinh(x) = \cosh(x)$$

No minus signs to remember in hyperbolic trigonometric calculus! Use these to find the derivative of $\tanh(x)$.

We will find these hyperbolic trigonometric functions to be extremely useful in giving us a shortened way to write results, without carrying around infinite series. For example, we can already simplify the circumference function:

Exercise 25.11 (Simplifying Circumference)**.** If $k < 0$, show that the circumference formula $C_\kappa(r)$ derived above can be rewritten in terms of hyperbolic trigonometry as

$$C_\kappa(r) = \frac{2\pi}{\sqrt{|\kappa|}} \sinh\left(\sqrt{|\kappa|}r\right)$$

25.4. Geometry with Curvature -1

To daydream of a world with curvature -1, we now have a rather concrete way forward to make initial guesses about its properties:

- Derive a formula for spherical geometry on the unit sphere
- Generalize that formula to spheres of radius R
- Replace the radius with the curvature κ
- Make sense of this formula for negative values of κ: if necessary, use a series expansion.
- Plug in $\kappa = -1$ to get our conjecture for how negatively curved geometry should behave.
- Convert back from a series to a hyperbolic trigonometric function, for a compact way to write things.

We saw this process play out in full for the circumference of circles above, where after all this work we ended up just going from $2\pi\sin(r)$ to $2\pi\sinh(r)$. Often, this process does indeed just boil down to replacing the usual trigonometric functions with their hyperbolic counterparts,

$$\sin r \mapsto \sinh r$$
$$\cos r \mapsto \cosh r$$
$$\tan r \mapsto \tanh r$$

But there can be other sign changes that occur as well, from factors of κ, so its important to actually run the arguments when you could be unsure.

25.4.1. Properties

We've already seen what happens to circumference of a circle in negative curvature: it grows *exponentially* with radius. What else can we learn about this mysterious geometry?

Exercise 25.12. Show that when $\kappa = -1$ we expect a circle's area to be related to the radius by

$$A(r) = 2\pi(\cosh r - 1) = 4\pi \sinh^2(r/2)$$

(Here we had a switch in the first term from $1 - \cos r$ to $\cosh r - 1$, which you may have missed if you just tried to replace cos with cosh) This formula - like the one for circumference before it - tells us that circles areas also increase *exponentially with radius.* There really is a lot of space in negative curvature!

Exercise 25.13. What are the areas of circles of radius $1, 10$ and 50?

A 50 meter radius (so, 100 meter diameter) circle on earth is just large enough to fit a football field inside: a football field is approximately 100m by 50m, so we can fit

$$\frac{\pi \cdot 50^2}{100 \cdot 50} = \frac{\pi}{2} = 1.57$$

football fields in our circle, in Euclidean space. How many football fields area fit inside the same radius circle in negatively curved space?

Exercise 25.14 (Hyperbolic Pizza)**.** One way to try and develop intuition for the strange behavior of circles is to think about the type of circles we see in daily life: pizzas! One major factor determining how good a pizza is is its *crust percentage* which we will define as

$$\text{CrustPercent} = \frac{\text{area(Crust)}}{\text{area(Pizza)}}$$

In this problem we will consider pizzas which have 1 inch crusts: meaning a 10 inch (radius) pizza has a 9inch radius center of toppings, surrounded by a 1 inch thick circle of crust.

- Show the CrustPercent for Euclidean pizza is $$$$. From this we see that as $r \to \infty$ the crust percent drops to zero: this makes sense, if you imagine an extremely large pizza with only a 1inch thick crust, it's totally reasonable that *most of the pizza is not crust*!

- WHat is the CrustPercent for a hyperbolic pizza of radius r? Show that when r is large, this limits to the constant

$$\text{CrustPercent} \to 1 - \frac{1}{e} \approx 63$$

  Thus crust is an inevitable part of life in hyperbolic space: even if you try to make the pizza huge it will still be well over half crust!

Exercise 25.15 (Hyperbolic Pizza II)**.** In this problem, we will imagine our unit to be inches (so, the radius appearing in formulas for space of curvature -1 is measured in inches).

You are at a pizzeria and are trying to decide if the 5 inch radius pizza they sell is large enough for you and your friends. They also sell a six inch (radius) pizza, but it costs twice as much. You think this is crazy! Is this a good deal, or not?

Your is feeling very hungry, and jokingly asks the pizza-maker how large of a pizza he would need to order so that its areas is the same as an american football field (100×50 yards). The pizza-maker says "I think I have room for that in my oven, coming right up!" How big of a pizza is he going to make?

Its OK to approach this question totally numerically: the goal is just to showcase how truly strange this new world is.

One thing we've already learned here - as this area formula goes to ∞ as the radius does, we see that space is infinite! So, we have a good reason to believe that Postulate 2 will end up being true about this geometry, unlike what we saw for the sphere.

We can learn more about this space by trying to take our other formulas across the divide.

Example 25.1. For the area of a triangle, starting from

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

we could scale lengths by a factor of R to get onto a sphere of radius R. THis scales areas by R^2 so the result is

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

Rewriting this in terms of curvature, we see

$$A = \frac{\alpha + \beta + \gamma - \pi}{\kappa}$$

And now, setting $\kappa = -1$, we get

$$A = \frac{\alpha + \beta + \gamma - \pi}{-1} = -(\alpha + \beta + \gamma - \pi)$$
$$= \pi - (\alpha + \beta + \gamma)$$

This formula is very striking, and quite confusing at first. Whereas circles get extremely large, here the area of a triangle (which must be a positive number) is equal to π minus some other number. This means it is *at most* π! So, triangles in this space seem to not be allowed to ever have area greater than 3.14. How can this be? Once we actually build the geometry, we will come to see.

Remark 25.3. Remember - we don't actually know this crazy space exists yet! We are just playing around with formulas to see what we should expect, if there really is a geometry on the other side of $\mathbb{E}^2$. If I were working 200+ years ago on the problem of parallels and I came across these two facts, I would think surely I should be able to quickly find a contradiction: how can circles areas grow exponentially but all triangles areas be less than 4?

Beyond this, we can continue by looking at the pythagorean theorem, and the trigonometry of right triangles.

Exercise 25.16. If $\kappa < 0$, show that the pythagorean theorem for a right triangle with legs a, b and hypotenuse c should be

$$\cosh\left(c\sqrt{|\kappa|}\right) = \cosh\left(a\sqrt{|\kappa|}\right)\cosh\left(b\sqrt{|\kappa|}\right)$$

Thus, when $\kappa = -1$ we get

$$\cosh c = \cosh a \cosh b$$

Example 25.2. For a right triangle with angles α, β, legs a, b and hypotenuse c continuing the formula for spherical geometry to $\kappa = -1$ gives

$$\sin\alpha = \frac{\sinh a}{\sinh c} \qquad \cos\alpha = \frac{\tanh b}{\tanh c}$$

and analogously for $\sin\beta, \cos\beta$

Exercise 25.17. Check this.

Exercise 25.18. Use these trigonometric rules and the pythagorean theorem for $\kappa = -1$ to discover a relationship between the tangent of an angle α, its opposite side a, and adjacent side b:

$$\tan\alpha = \frac{\tanh a}{\sinh b}$$

25.4.2. Parallels

Now that we've seen glimpses of the possible negatively curved world beyond the Euclidean plane, we should spend some time thinking through their *implications*. In particular, we are most interested in whether or not such a space of negative curvature will truly be the counterexample we seek: will it fail the fifth postulate?

The 5th postulate gives a condition on when two geodesics will intersect: it says if you can measure the angle of intersection they make with a third line, and its less than two right angles, then the two lines intersect one another at some distance along that direction.

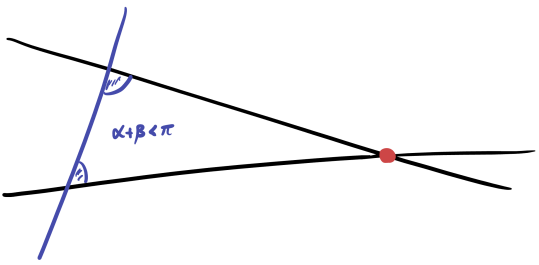


Figure 25.12.: Postulate 5 gives a condition on when two lines will intersect, in terms of the angles of intersection with a third line.

Our tool to analyze this situation will be *trigonometry*, where we'll reason as follows. Choose some line and pick a point at distance d away from that line. Through this point, any line we draw makes some angle with the vertical segment of length d, which we can call θ. Our goal will be to understand how to relate this angle to the point whre the two lines intersect (if they do at all).

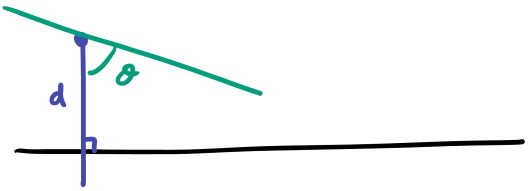


Figure 25.13.: A setup for testing the truth or falsity of the parallel postulate in a special case.

Let's review first what happens in the Euclidean case. Here, we can use trigonometry to express the relationship between the height d, the angle θ, and the distance L to the point of intersection

$$\tan\theta = \frac{L}{d}$$

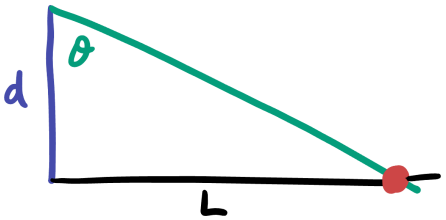


Figure 25.14.: Verifying the parallel postulate is satisfied in this special case in Euclidean geometry, using trigonometry.

We can then solve this for L, getting $L = d\tan\theta$. This formula gives us a value for L any time that $\tan\theta$ is defined, and so we get a point of intersection whenever $\theta \neq \pi/2$. When $\theta = \pi/2$ there is no solution for L, and so there is no point of intersection. Thus, our new line intersects the first except exactly when the sum of the two angles is π (each being $\pi/2$), as the parallel postulate states.

What happens to this reasoning if we want to translate it to negatively curved geometry? Well, we will need to figure out what the analogous relationship between θ, d and L should be. While we don't have any ready-made trigonometric relationship sitting around (yet), we can derive one from what we already know, much like you have done in spherical geometry before:

Exercise 25.19. Using the analogs of right triangle trigonometry in negative curvature, show that

$$\tan\theta = \frac{\tanh L}{\sinh d}$$

Hint: write $\tan\theta = \sin\theta/cos\theta$*, use the fact that we* do* know what the trigonometric formulas for sine and cosine should look like (from Example 25.2), and then we also know the pythagorean theorem in negative curvature (Exercise 25.16) to simplify.*

Like in the Euclidean case, we can solve this for L: but perhaps its easier to start by just solving for $\tanh L$:

$$\tanh L = \sinh d\tan\theta$$

What does this equation tell us? Well, whenever we can find an L that makes it true that means we have found a point of intersection between our two lines. And,

whenever there is *no* L which works, we know the two lines do not intersect - they must be *parallel*! So, analyzing the solutions to this equation will tell us exactly when lines in negative curvature would and would not intersect.

To analyze the possible solutions here, we need to think a bit about the behavior of the hyperbolic-trigonometric that arise. Most importantly, the function$ $\tanh L$ has horizontal asymptotes at ± 1, so it is *impossible* to have any solution to $\tanh L = x$ if $|x| \geq 1$.

But wait: it seems pretty easy to make the other side of the equation bigger than 1: both $\tan\theta$ and $\sinh d$ are functions that can grow unboundedly! Let's do an explicit example: if we look at a point $d = 1$ unit of distance away, our equation becomes

$$\sinh(1)\tan\theta \approx 1.175\tan\theta > 1$$

This happens anytime $\tan\theta > 1/1.175 \approx 0.85$, which is pretty easy to arrange - if $\theta = 0.73$ radians then $\tan\theta = 0.9$ which will do the trick.

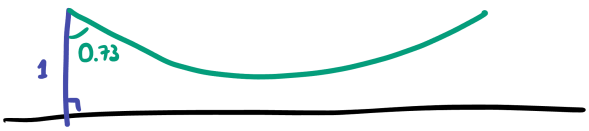


Figure 25.15.: The parallel postulate will be **false** in hyperbolic geometry.

But what does this mean? This means that when $d = 1$, the line making angle 0.73 with the vertical *never manages to intersect the original horizontal line*! We have an explicit pair of lines making total angle less than two right angles, whcih nonetheless never intersect one another! Thus, Euclid's fifth postulate *must be false* in hyperbolic space!

We can use this same sort of reasoning quickly to see that the equivalent Playfairs axiom is also violated here. Considering the same situation as above, we can take a line, and a point on that line, and find many lines through that point which do not intersect the original: any line making original angle greater than $\arctan(1/\sinh(1))$ will do!

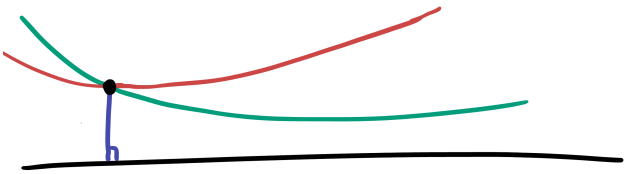


Figure 25.16.: Playfair's axiom will be false in hyperbolic geometry.

Thus, our daydreams of pushing the radius past infinity - if they can be formalized - will indeed solve the most important problem of the greeks.

26. Models

At this point we have a lot of ideas of what negatively curved, or *hyperbolic space* is like, but we do not have a concrete way to work with it. This causes two major problems:

- **We don't have any means of discovering new facts about hyperbolic space** everything we've learned as come from our understanding of spherical geometry, passed along the transition as curvature goes to -1. This is a great approach for learning facts about hyperbolic geometry that are analogous to the sphere, but it leaves us blind to any potential *differences*. And we know there will be differences! As a concrete case where this will be important: we know on the sphere there's just one type of (orientation preserving) isometry: rotations. But in the Euclidean plane there are *two types*: translations and rotations! How many types of isometries does hyperbolic space have?
- **How do we really know that hyperbolic space even exists?** For both the Euclidean plane and the sphere, we started with a *definition* we specified the points of the geometry, its tangent vectors, and its infinitesimal length / inner product. Then we built up all other results as *theorems*. But so far we have no analogous construction of hyperbolic space. We have a bunch of theorems about how distance *should* work, or area *should* work, but we don't even have a clear notion of what the points are we are measuring distance between! Because the implications of this discovery are *enormous* (namely, it resolves the most important problem of the greeks, which was open for more than two thousand years), we should not be satisfied with this state of affairs. Its like the discovery of a new species deep in the jungle: its useful to have a detailed description of the creature, but this isn't enough. If you claim to have discovered Bigfoot, you'd better present a specimen!

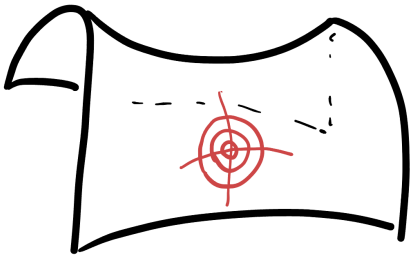
In this chapter we aim to simultaneously remedy these two concerns, and produce an explicit model of hyperbolic space.

26.1. Conceptual Troubles

Upon sitting down to try and rigorously *define* the hyperbolic geometry we've glimpsed through the transition, we immediately hit a conceptual hurdle: how are we supposed to find the right set of points and definition of infinitesimal length? For

both the Euclidean plane and the sphere, we didn't have to confront this problem because we essentially already knew the right definitions to write down: we've all seen many planes and spheres before in daily life. But negatively curved space? We do see around us spaces that have some negative curvature, like a Pringle chip

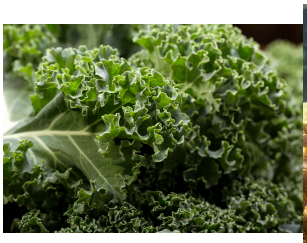
(a) A pringle chip has negative curvature



(b) But the curvature of a hyperboloid is not *constant*: its much less curved far away from the center.

But this is not hyperbolic space. The curvature of a pringle chip is more negative at the center and less negative (closer to flat) the farther out you go: here's a zoomed out view of the *pringle chip function* $x^2 + y^2$

But hyperbolic geometry is supposed to have curvature -1 *everywhere*: indeed, if we really reach this space along a path including spheres and the Euclidean plane, it should be both *homogeneous* and *isotropic*. A pringle chip is neither! Other shapes familiar from the world around us are closer to having constant negative curvature, such as the surface of a kale leaf, certain types of coral, or the wavy back of a nudibranch.



(a) A kale leaf is very negatively curved.

(b) Many corals also exhibit negative curvature.

(c) So does this slug.

Anytime evolution tries to *increase the area of something* beyond Euclidean constraints, it resorts to negative curvature.

However these surfaces are all rather small: hyperbolic space is supposed to be infinite! And its not quite clear upon looking at one of these surfaces how to continue

it, and make it bigger. As the surface grows exponentially fast it wrinkles more and more - it seems hard to believe one could continue this infinitely without the surface self-intersecting, or worse - having a crease or points of non-differentiability (rendering all of our calculus based tools useless).

In fact, this is a very reasonable worry to have. This was though deeply about at the turn of the 20th century, and it was actually proved that there is *no way to do this*.

Theorem 26.1 (Hilbert's Theorem)**.** *There is no surface in* $\mathbb{E}^3$ *which has the geometry of the hyperbolic plane.*

Remark 26.1. In the technical statement of the theorem, it is important that the surface is described by at least *second differentiable functions* for the proof to go through. But this is no serious restriction as we are already working in a world where calculus forms the foundations of everything we do, and you need second derivatives to even *define* curvature (meaning, if our surface were not described by such functions, then the limit defining curvature would not exist at some points).

Thus, we must abandon hope of being able to work exactly as we did for the sphere, where we found a subset $^2 \subset \mathbb{E}^3$ of points, found the tangent spaces, and then defined all geometry using the *euclidean dot product* on these tangent spaces. Instead, we are going to have to make use of our study of *maps*, and the flexibility they provide.

26.2. Making a Map

So, there is no surface in $\mathbb{E}^3$ which has the same geometry as the hyperbolic plane. But that's alright - there isn't any region in $\mathbb{E}^2$ that has the same geometry as ^{2} (via *The Mapmaker's Dilemma*) but we are able to completely accurately do computations about ^{2} using regions of $\mathbb{E}^2$ via a *map*.

In the previous part, our construction of a map went like this: we started with the 'actual' sphere, and wrote down a chart onto a region of the plane, and its inverse, a parameterization ψ taking that region to the sphere. We then used this parameterization ψ to make all geometric properties of the sphere computable on the map directly (map-lengths, map-angles, map-dot-products, map-areas, etc). At the end of this procedure, we are left with a region on the plane, and a new rule on how to compute geometric quantities in that region (different from the original Euclidean way). But, if we follow these new map-formulas correctly, it is precisely as good as if we worked directly on the sphere itself.

Taking this observation to its logical (but shocking) conclusion was the subject of Riemann's *Habilitationsschrift*, or post-doctorate research assigned by Gauss. Riemann realized that to do geometry, maps are all you need!

> **The Fundamental Insight of Riemannian Geometry:** Any region of the plane, together with a rule for how to perform the dot product at each point (and thus, to compute infinitesimal lengths and angles) is a map of some geometry - just perhaps a geometry that we have never before seen.

Example 26.1. From Riemann's perspective, if I were to write down the strip in the plane $|y| < \pi$ and explain that at the point (x, y) in this strip, the dot product of two vectors $v = \langle, v_1, v_2\rangle$ and $w = \langle w_1, w_2\rangle$ is

$$\frac{v_1 w_2 + v_2 w_2}{\cosh^2 y}$$

Then I am describing a true geometry. No matter what questions you asked me about the geometry, with enough work I'd be able to do the calculations and answer them, so this is truly *indistinguishable* from having the 'real thing' in front of me - whatever that might mean! Indeed - Riemann's whole point is that *there is no real thing above and beyond the map*: the map is just as real as anything else!

Of course in this particular case if I sat down and started doing lots of calculations, I would realize that this geometry has some familiar properties: all of its geodesics have length 2π, triangles angles are determined by their angle sum, and so on. This is just the dot product that we found for the Mercator projection of the sphere!

But Riemann's genius is to go further, and take something like the below seriously

$$\frac{v_1 w_2 + v_2 w_2}{\cos^2 y}$$

or even

$$(x^2 + 1)v_1 w_1 + e^{x+\sin y} v_2 w_2$$

These dot products describe some geometric spaces we have not yet encountered. And its up to us to do math, and figure it out!

In taking this bold step, Riemann reduced the set of things needed to do geometry down to one: just a dot product on a region in the plane, or a region in some higher dimensional (euclidean) space. Because this dot product is the main tool we use to find infinitesimal lengths, then lengths, then distances (and hence produce a *metric*), it's called the Riemannian metric.

Definition 26.1 (Riemannian Metric)**.** Given a region $R \subset \mathbb{E}^2$ in the plane, a *Riemannian metric* on R is a choice of formula for the dot product at each point $(x, y) \in R$. Any such choice is specified by three functions, $G(x, y)$, $E(x, y)$ and $F(x, y)$ where we write

$$(v \cdot w)_{(x,y)} = G(x, y)v_1w_1 + F(x, y)(v_1w_2 + v_2w_1) + F(x, y)v_2w_2$$

Remark 26.2. The two middle terms both have the same coefficient $F(x, y)$ because the order we take the dot product in shouldn't matter: we want $v \cdot w = w \cdot v$.

Thus, as we move around to different points (x, y) the definition of the dot product is allowed to vary, representing that our map might be distorting the underlying geometry by different amounts at different points.

Exercise 26.1. Write the Riemannian metric (as a matrix) for the stereographic and mercator projections of the sphere.

Remark 26.3. Riemann actually went further than this: recalling that it is not even possible to cover the entire earth wtih a single chart, it would be shortsighted to define *geometry* in general in terms of a single map. Instead, a space is called a *manifold* if there is some *atlas of charts* covering it, and a Riemannian manifold if each of these charts has a Riemannian metric.

Alright, so how do we go about finding a map, or in mathematical terminology, a *Riemannian metric* for hyperbolic space? We continue pushing spherical things beyond infinity, of course! We can take any map we like for the sphere and run our *hyperbolization* procedure on it: write it for radius R, convert to curvature, then push to $\kappa = -1$.

Because out of all the maps we studied stereographic projection certainly had the (mathematically) nicest properties, we will start wtih this. We found the dot product for a sphere of radius R to be, at the point $p = (x, y)$

$$(v \cdot w)_{\text{map}} = \frac{4R^4}{(R^2 + |p|^2)^2}(v \cdot w)_{\mathbb{E}^2}$$

Writing this in terms of curvature is straightforward, as only even powers of R show up already

$$(v \cdot w)_{\text{map}} = \frac{4\frac{1}{\kappa^2}}{(\frac{1}{\kappa}^2 + |p|^2)^2}(v \cdot w)_{\mathbb{E}^2} \tag{26.1}$$

$$= \frac{4}{\kappa^2(\frac{1}{\kappa} + |p|^2)^2}(v \cdot w)_{\mathbb{E}^2} \tag{26.2}$$

$$= \frac{4}{(\frac{\kappa}{\kappa} + \kappa|p|^2)^2}(v \cdot w)_{\mathbb{E}^2} \tag{26.3}$$

$$= \frac{4}{\left(1 + \kappa|p|^2\right)^2}(v \cdot w)_{\mathbb{E}^2} \tag{26.4}$$

This formula is exactly equivalent to what we already knew for the sphere when $\kappa > 0$, but continues to make sense when $\kappa \leq 0$. In the case that κ is negative, its easiest to write things in terms of $|\kappa|$ as usual, where we get

$$(v \cdot w)_{\text{map}} = \frac{4}{\left(1 - |\kappa||p|^2\right)^2}(v \cdot w)_{\mathbb{E}^2}$$

This map has some interesting behavior! Its not defined on the whole plane, but only on a subset of it: once $|p|$ reaches a size of $1/\sqrt{|\kappa|}$ then we get division by zero. In particular, for the "unit sized" space $\kappa = -1$ the map is only defined inside the unit disk! Nonetheless as we will shortly seen, the geometry described by this map is infinite. This is the so-called *disk model* of the hyperbolic plane.

26.3. The Disk Model

Definition 26.2 (Disk Model of $\mathbb{H}^2$)**.** The disk model (also called the Poincare Disk or the Beltrami-Poincare Disk) is given by the unit disk in the plane

$$\{(x, y) \in \mathbb{E}^2 \mid x^2 + y^2 < 1\}$$

together with the Riemannian metric

$$(v \cdot w)_{PD} = \frac{4}{(1 - |p|^2)^2}(v_1 w_1 + v_2 w_2)$$

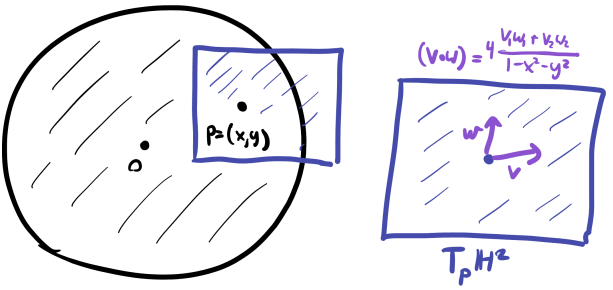


Figure 26.3.: The definition of the Disk Model

What does this model of the hyperbolic plane look like? Well, we see immediately that it is *conformal* (no surprise, since stereographic projection was as well) since the dot product is a multiple of its Euclidean counterpart. Thus, any angles we see in the disk will represent the true angles on the plane. But the scaling factor has completely the opposite behavior of stereographic projection. It approaches infinity as we head towards the boundary of the disk, which means that objects there are *much bigger than they appear*. Equivalently, if you move an object towards the boundary of the disk it should appear to *shrink* in size!

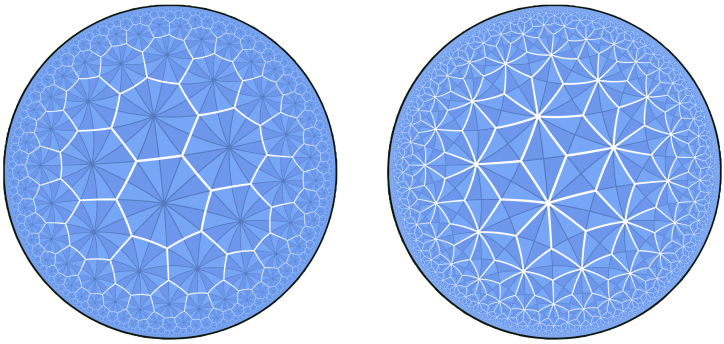


Figure 26.4.: The Disk Model

These strange distortions of size aside, there are many positive qualities about this model. In some ways, having built our model of hyperbolic space *inside* of the Euclidean plane is actually more useful than doing something else - as we can use things we know about $\mathbb{E}^2$ to help us deduce facts about $\mathbb{H}^2$: here's a useful example, with almost no knowledge we can deduce the circles in $\mathbb{H}^2$ about the center of the disk model.

First, Euclidean rotations of the plane about the origin restrict to isometries of the Disk Model of $\mathbb{H}^2$: Rotation isometries of the plane do not change the distance of a

point from the origin, and they do not affect the Euclidean dot product of two tangent vectors. Because the hyperbolic disk dot product is just the Euclidean one rescaled by a function of distance from the origin, it is also unchanged by a rotation, so these rotations are hyperbolic isometries!

Knowing this, we can see that hyperbolic circles about the origin of the disk model are Euclidean circles. Euclidean rotations of $\mathbb{H}^2$ about the origin are isometries of the model, so every point on the same Euclidean circle as p about O is the same hyperbolic distance from O: this is the definition of a hyperbolic circle!

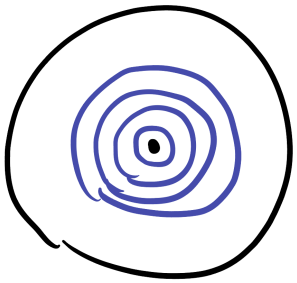


Figure 26.5.: Circles about O in the Disk model are Euclidean circles.

Note that this argument does not actually tell us the hyperbolic radius of this circle, because we don't yet know how to measure the hyperbolic distance in our model from O to p. We will figure out how to do this in the next chapter, and confirm this space really does have curvature -1.

The disk model is wonderful for depciting the hyperbolic plane, and for doing any sort of calculations that involve rotational symmetry, as we saw above. But serveral other computations are rather challenging in the disk model, due to the fact that the scaling factor in front of the dot product depends on the *euclidean distance from O. This makes some work with x and y coordinates cumbersome, as they naturally describe affine lines and not circles in the plane.

But there is nothing tying us down to *exclusively* use this disk we just discovered! Its just one map, after all. And, just like for the sphere, different maps are best suited for different purposes. In the next section below we derive the map perhaps most used for computations, the *Half Plane Model*.

26.4. The Half Plane Model

The half plane, or upper half plane, or Poincare half plane model of the hyperbolic plane is so named becasue it is a map which is drawn on the upper half space of the

Euclidean plane. In fact this upper half plane map is just a different perspective on the map we already drew above, using stereographic projection!

We saw previously that stereographic projection lets us write down a conformal map that takes the unit disk to the half plane by projecting to the sphere, rotating, and re-projecting to the plane:

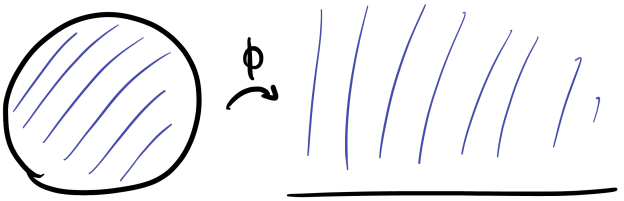


Figure 26.6.: Mapping the Disk to the Half Plane

Indeed, in Exercise 32.67 you actually derived the formula expressing this map, which after a little simplification is

$$\phi(x,y) = \left(\frac{2x}{x^2 + (y-1)^2}, \frac{1-(x^2+y^2)}{x^2+(y-1)^2} \right)$$

We can similarly derive its inverse (let's call it ψ), which maps the upper half plane to the disk.

$$\psi(x,y) = \left(\frac{x^2+y^2-1}{x^2+(y+1)^2}, \frac{-2x}{x^2+(y+1)^2} \right)$$

Remark 26.4. These maps can be described much shorter using complex arithmetic: if we write $z = x + iy$ for a point in the plane, then

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

and its inverse ψ is

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

We now use this to transfer the disk model itself onto the upper half plane. The result is going to be a conformal model of Hyperbolic space as well (the original model was conformal, and this new model is the result of applying a conformal map to it), so we know the dot product is just going to be some mutliple of the Euclidean version. And all that needs to be done is calculate that scaling factor!

Exercise 26.2 (The Scaling factor of the UHP)**.** At a point (x, y), show that the scaling factor for the dot product is simply $\frac{1}{y^2}$, following the steps below:

- Start with a unit vector, say $\langle 1, 0\rangle$ based at some point (x, y) in the upper half plane.
- Apply $D\psi$ to this vector, to move it to the disk model, based at the point $\psi(x, y)$.
- Find the scaling vactor at this point, and simplify the result!

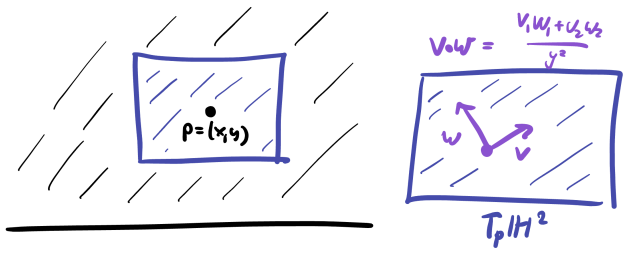


Figure 26.7.: The definition of the Half Plane Model

Definition 26.3 (Half Plane Model of $\mathbb{H}^2$)**.** The points of the half plane model of hyperbolic space are

$$\{(x, y) \mid y > 0\}$$

together with the Riemannian metric

$$(v \cdot w)_{HP} = \frac{v_1 w_1 + v_2 w_2}{y^2}$$

Thinking again about what this map *represents*, we know that since its conformal we can trust the angles we see to be accurate, but we should distrust lengths. Since the scaling factor is going to infinity as we approach the x-axis, we know that the true size of a shape is *much larger than it appears* down by the axis. This is analogous to the what happens near the unit circle in the disk model.

As $y \to \infty$ the scaling factor approaches zero, which says that the half plane model also distorts distances the *other way* as we move up, away from the line $y = 1$ where infinitesimal distances are correct. Objects high up in the half plane appear much larger than they actually are, like Greenland on the Mercator map of the sphere.

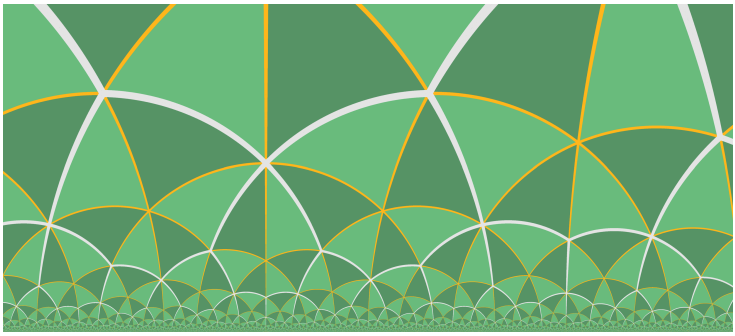


Figure 26.8.: The Half Plane Model

It will prove very useful to us to be able to go back and forth between the disk and half plane models of hyperbolic space. Pre-empting this, we had a homework exercise to think through this map when it was first derived (just in the context of stereographic projection), which we will now go through together. Say we start with this grid of polar coordinates in the disk:

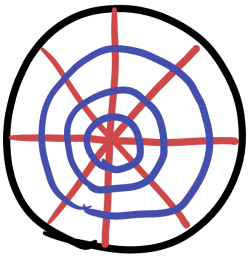


Figure 26.9.: Polar Coordinates drawn in the Disk Model

When we map to the upper half plane, we know a couple things: - By direct calculation, we can find the origin O maps to $(0, 1)$. - The map sends generalized circles to generalized circles, so circles about O go to generalized circles.
- But, a circle only maps to a line if it passes through the projection point. None of these do - they're all inside the lower hemisphere, and then rotated to be inside the rightmost hemisphere, and the projection point is the north pole. Thus, they project to circles:

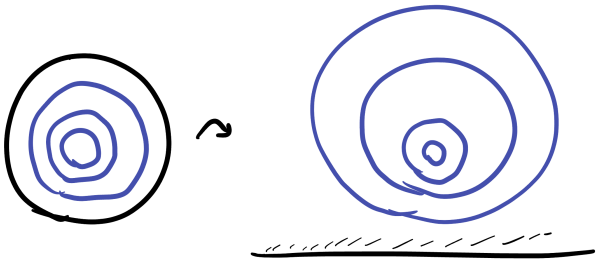


Figure 26.10.: Circles about O in the Disk and their images in the Half Plane

- Similarly for the lines through the origin - these go to generalized circles.
- The map is conformal, so it preserves angles. The two original families of curves intersected orthogonally, so the new curves must as well.
- In particular, the radial lines that intersected the unit circle orthogonally now intersect the x-axis orthogonally.

This tells us that the radial lines must go to lines/segments of circles passing through $(0, 1)$ and hitting the x axis orthogonally. The only line with both these properties is the y axis, so the rest must be half-circles, whose centers are on the real axis

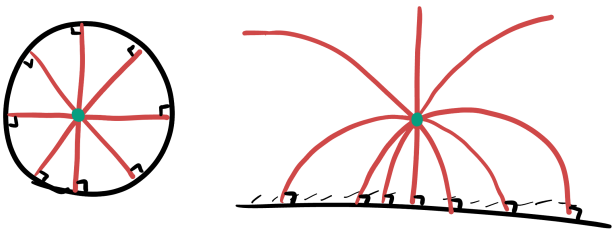


Figure 26.11.: Diameters of the Disk and their images in teh Half Plane

We can run a similar argument in reverse, to think about what the standard xy grid pattern on the half plane would look like in the disk. Here we find that both families of curves must go to arcs of circles in the disk, like below:

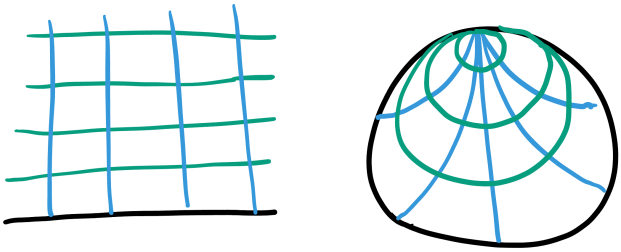


Figure 26.12.: A grid in the Half Plane and its image in the Disk

26.5. Other Maps

We discoverd the disk model, and later the half plane model starting with the stereographic projection map. And this was a good choice - the niceness of stereographic projection led to two particularly nice maps of hyperbolic space! But these are far from the only maps: its possible to apply this same procedure to many maps of the sphere, and arrive at a map of hyperbolic space.

26.5.1. Mercator

What happens if we take the Mercator projection and extrapolate to negative curvature? The Mercator projection's dot product is also conformal and given by a relatively simple formula:

One may (correctly) guess the form this will take if we were to (1) write it out for a sphere of radius R (2) convert this to curvature (3) push the curvature to a negative number, using the taylor series representation and (4) convert back to trigonometric form at $\kappa = -1$: it will switch the denominator $\cosh(y)^2$ for simply $\cos(y)^2$!

Exercise 26.3. Check this.

The result is a map onto the plane of geometry with curvature -1: hyperbolic geometry! This map's dot product is

$$(v \cdot w)_{(x,y)} = \frac{v_1 w_1 + v_2 w_2}{\cos^2 y}$$

Which becomes undefined when $\cos y = 0$, so the region where the map makes sense around the origin is in the strip

$$\left\{(x, y) : |y| < \frac{\pi}{2}\right\}$$

The dot product on this strip has scaling factor diverging to infinity towards its boundary. This tells us we should expect that objects appear very small near the boundary of the strip, compared to their actual size.

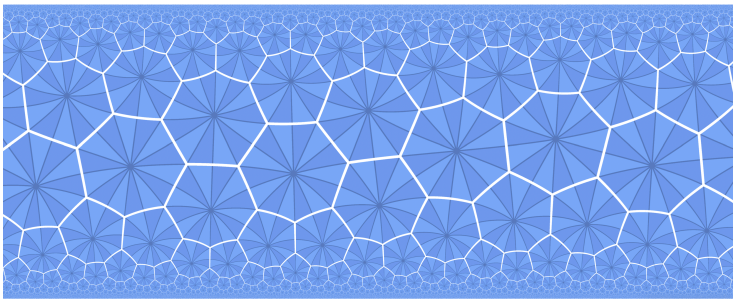


Figure 26.13.: The Band Model (Mercator)

This is often called the *band model*, as it is defined in the interior of a band, or horizontal strip in the Euclidean plane. You may recognize the band model from our book's front cover!

Exercise 26.4. Which isometries are easy to see must exist in the band model?

While we will wait until the next chapter for formal verification, we can already figure out what some of the geodesics in the band model must be. In the Mercator porjection, the equator geodesic was represented by the line $y = 0$ in all curvatures - and so once the curvature becomes negative this line is still a geodesic of our model! Other parallel lines to it on the sphere wer *not* geodesics (they were lines of latitude, or circles around the poles), and this also remains true in the Band model.

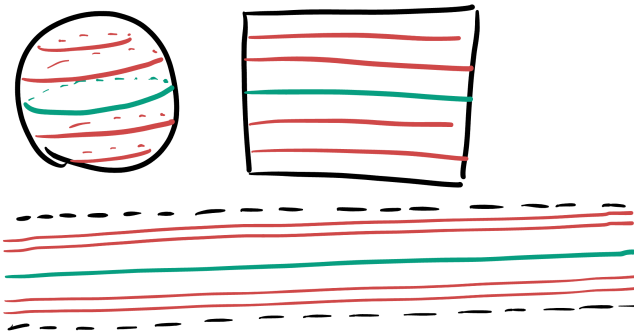


Figure 26.14.: The line $y = 0$ is the Equator in the mercator projection, which is a geodesic. All other horizontal lines are latitudes - *not* geodesics. The same remains true in the Band Model.

Vertical lines in the Mercator projection are lines of longitude, which are geodesics for spheres of every curvature. Thus, passing to negative curvature we find that all vertical lines in the Band model still to be geodesics:

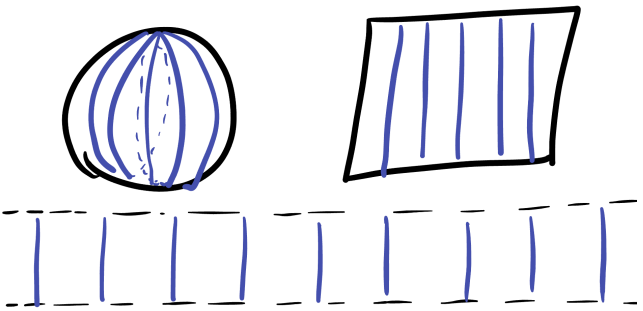


Figure 26.15.: Vertical lines in the Mercator projection represent lines of longitude, which are geodesics. The same remains true in the Band Model.

Its useful to pause here and think about what a vertical strip between two such curves on our map actually represents. On the sphere, the scaling factor $1/\cosh(y)$ shrank towards zero as we moved away from the equator, meaning that the lines were actually *closer* than they appear.

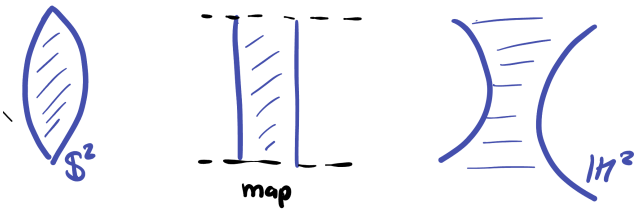


Figure 26.16.: Geodesics converge in positive curvature, and diverge in negative curvature.

But once $\kappa = -1$ and the scaling factor becomes $1/\cos(y)$ which gets extremely large as y grows, meaning that these geodesics actually become *much farther apart* than they appear to be. Quantitatively: the distance along a horizontal line is $1/\cos(y)$ times as long as it appears! This is the first quantitative measurement where we can confirm that geodesics indeed race away from one another in negative curvature.

Remark 26.5. Note this isn't precisely the *distance* between points, as we are measuring the length of a horizontal curve which is not a geodesic. To find distance we'll have to do some more computations in the next chapter.

26.5.2. Archimedes?

Archimedes map of the earth was an *equal area projection.* So far, all of our maps of hyperbolic space have been conformal but have massively distorted area. Can you find an analog of Archimedes map in negative curvature, to give an area-preserving option?

27. Geometry

Now that we have a concrete model of hyperbolic space (actually, many of them), its time to put them to work, and prove some things about hyperbolic space. Of course we already know a lot - we followed many geometric formulas directly through the limiting procedure! But these formulas speak of *geodesics* and *circles*, and we don't know what these things look like in our models, so it's hard to begin using these facts to do anything! We will remedy that gap in this chapter, and in the end, prove that hyperbolic space satisfies Euclid's postulates 1-4 while failing the fifth.

27.1. Homogenity & Isotropy

As we have for both the plane and the sphere, we begin our journey by investigating the symmetries of hyperbolic space. This will be our first foray into using the two models to *inform one another*, and switching back and forth whenever is convenient, whether than being tied down to a single way of calculating.

As we saw in defining the disk model, its easy to see that rotations about its center are isometries (though its hard to see that there are any other symmetries at all, at first!)

Proposition 27.1 (Rotations of Disk are Isometries)**.** *Euclidean rotations of the plane about the origin restrict to isometries of the Disk Model of* $\mathbb{H}^2$*.*

Proof. Rotation isometries of the plane do not change the distance of a point from the origin, and they do not affect the Euclidean dot product of two tangent vectors. Because the hyperbolic disk dot product is just the Euclidean one rescaled by a function of distance from the origin, it is also unchanged by a rotation, so these rotations are hyperbolic isometries! □

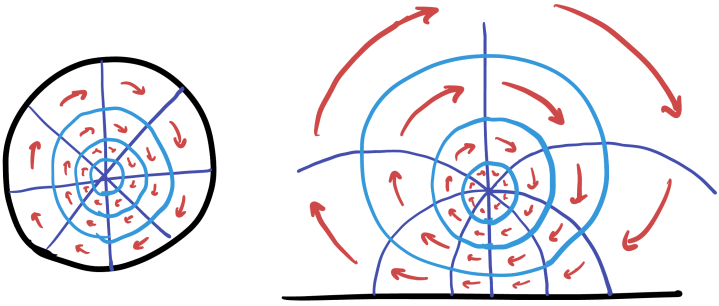


Figure 27.1.: Rotations about O in the Disk, and the corresponding rotations about $(0, 1)$ in the Half Plane.

To find more isometries, we can switch our viewpoint and take a look in the Half Plane model. Since isometries are maps which preserve the infinitesimal dot product, we can use the fact that the scaling factor here only depends on y to find some new isometries of hyperbolic space.

Proposition 27.2 (Horizontal Translation is an Isometry)**.** *Any euclidean horizontal translation* $(x, y) \mapsto (x + a, y)$ *is an isometry of the upper half plane model.*

Proof. Translations are Euclidean isometries, so they preserve the Euclidean dot product. But horizontal translations also preserve the y coordinate, and hence the scaling factor. Thus, they preserve the hyperbolic dot product as well. □

These isometries are *different* than those that we discovered before, as they move the point $(0, 1)$ (which is where the origin of the disk was sent), whereas our earlier-discovered isometries fixed the origin (and hence would would fix $(0, 1)$ here).

We can see what these look like in both models, using the pictures we developed above. Vertical lines in the Half Plane correspond to circles through $(0, 1)$ on the boundary in the Disk Model, and our new isometries move these curves to one another.

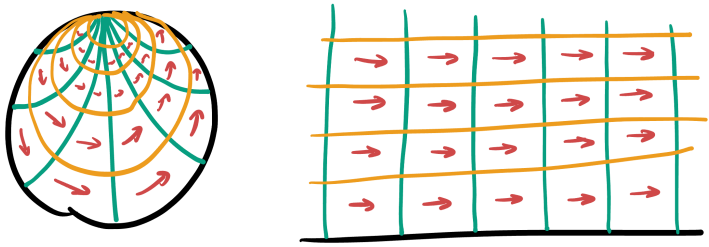


Figure 27.2.: Horizontal translation in the Half Plane is an isometry of hyperbolic space.

Other translations of the Half Plane are *not isometries*, as any map that changes the y coordinate is going to change the scaling factor applied by the dot product:

Example 27.1. The map $\tau(x, y) = (x, y + 1)$ is not an isometry of the Half Plane model.
Let v, w be two vectors based at (x, y) in the upper half plane: then their hyperbolic dot product is

$$(v \cdot w)_{\mathbb{H}^2} = \frac{v_1 w_1 + v_2 w_2}{y^2}$$

The derivative $D\tau$ of any translation is the identity, so $D\tau v = v$ and $D\tau w = w$, but now based at $(x, y + 1)$ which means their new dot product is

$$((D\tau v) \cdot (D\tau w))_{\mathbb{H}^2} = \frac{v_1 w_1 + v_2 w_2}{(y + 1)^2}$$

This is a *different number* so the dot product was not preserved, and the map was not an isometry.

We can see in this example what went wrong however: we moved vertically (increasing the denominator) without changing the vectors (leaving the numerator the same). Instead, when we

Proposition 27.3 (Homothety is an Isometry)**.** *Let $\lambda > 0$. Then the Euclidean similarity $s(x, y) = (\lambda x, \lambda y)$ is an isometry of the Half Plane model of hyperbolic space.*

Proof. Like the example, we just compute: here the derivative Ds is just scaling by λ, so our vectors v and w at (x, y) go to the vectors λv and λw at $(\lambda x, \lambda y)$. In the new dot product, the numerator is multiplied by λ^2 (one λ from each vector), and the denominator is multiplied by λ^2, so they cancel out:

$$\begin{aligned}((Dsv)\cdot(Dsw))_{\mathbb{H}^2} &= \frac{\lambda v_1 \lambda w_1 + \lambda v_2 \lambda w_2}{(\lambda y)^2} \\ &= \frac{\lambda^2(v_1 w_1 + v_2 w_2)}{\lambda^2 y^2} \\ &= (v\cdot w)_{\mathbb{H}^2}\end{aligned}$$

□

What does this kind of isometry look like in our two models? We understand well what it looks like through our *euclidean eyes* in the Half Plane: its just a similarity of the plane! But while it looks like points are getting farther apart here, they're not. This is an isometry after all, so it preserves hyperbolic distances! The fact that it looks to be expanding is just a consequence of the fact that distances are artificially short-looking near the bottom, and artificially long towards the top. But what about in the disk model?

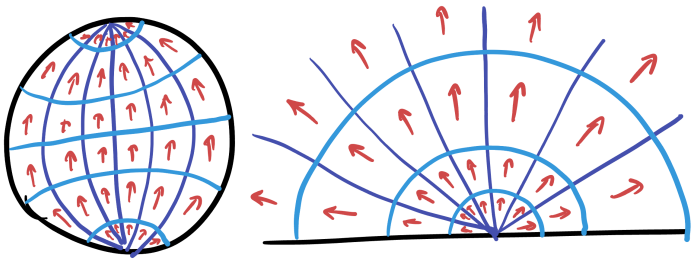


Figure 27.3.: Euclidean similarities $(x, y) \mapsto (\lambda x, \lambda y)$ are isometries of the Half Plane model. These act like translations, as is easier to see in the Disk.

We've now discovered enough isometries to prove rigorously that hyperbolic geometry is both homogeneous and isotropic:

Exercise 27.1 ($\mathbb{H}^2$ is Homogeneous)**.** There is an isometry taking any point of hyperbolic space to any other.

Next, we want to use this to see that space is isotropic: that about any point, there is a rotation by any amount. Here we may wish to switch over to the Disk Model, brining with us the fact that we just proved space is isotropic (even though we haven't bothered to write down what those isometries look like in the disk). We know we can rotate by any angle we like about O, and now we *also* know that its possible to move O to any arbitrary point p. So, what do we do? We translate p to O do any rotation we like, and then translate back of course! Conjugation saves the day.

Corollary 27.1 ($\mathbb{H}^2$ is Isotropic)**.**

This is enough to see that Euclid's Postulate 4 is true for hyperbolic space: remember that Euclid's phrasing *all right angles are equal* was the original way of saying *homogeneous and isotropic.*

27.2. Geodesics

We've been able to track down a couple curves in the various models that certainly must be geodesics - reasoning from watching what happens to great circles on the sphere under our transition. But we are still lacking in two things: (1) a rigorous proof, done in hyperbolic geometry itself, that these are in fact distance minimizing and more importantly (2) a classification of what *all* the geodesics are. We remedy this below, with computations that should feel very analogous to both the Euclidean and spherical cases.

Proposition 27.4 (Vertical Lines are Geodesics)**.** *In the Half Plane model, the vertical curve $\gamma(t) = (0, t)$ is minimizing between any two points.*

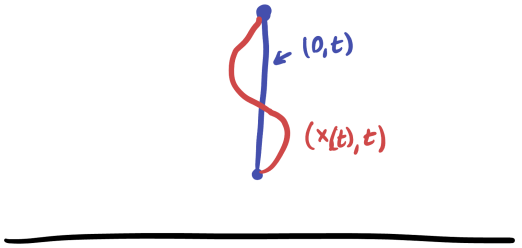


Figure 27.4.: Vertical segments are length-minimizing in the Half Plane.

Proof. Consider the endpoints $(0, a)$ and $(0, b)$, and let $\alpha(t) = (x(t), t)$ be any other curve between these for $t \in [a, b]$. Then $\alpha' = \langle x', 1\rangle$ based at (x, t) so its infinitesimal length is

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

We can do our usual trick, and notice that whatever x' is, we know its square is positive, so we certainly have

$$\|\alpha'\| \leq \frac{1}{t}$$

But this is exactly the infinitesimal lenght of γ, since $\gamma' = \langle 0, 1 \rangle$. Thus we can integrate this inequality to find

$$\|\gamma'\| \leq \|\alpha'\| \implies \int_a^b \|\gamma'\| dt \leq \int_a^b \|\alpha'\| dt$$
$$\implies \text{length}(\gamma) \leq \text{length}(\alpha)$$

□

Given that these curves are distance minimizers, we can directly find the distance function between vertically separated points:

Corollary 27.2 (Vertical Distance Formula)**.** *The distance between* $(0, a)$ *and* $(0, b)$ *is* $\log(b/a)$.

Proof. This is just an integral of the infinitesimal length of $\gamma(t) = (0, t)$ from $t = a$ to $t = b$:

$$\text{length}(\gamma) = \int_a^b \frac{1}{t} dt = \log(t)\Big|_a^b = \log(b) - \log(a) = \log\left(\frac{b}{a}\right)$$

□

This tells us that all vertical lines are geodesics in the half plane model. And we can transfer this information over to the Disk Model to learn about some of the geodesics there. The vertical line through $(0, 1)$ goes to a diameter of the unit disk as we can see by direct computation, plugging in $(0, t)$. Then, since we know the unit disk as rotations as isometries, we can see that *all diameters of the disk* are geodesics.

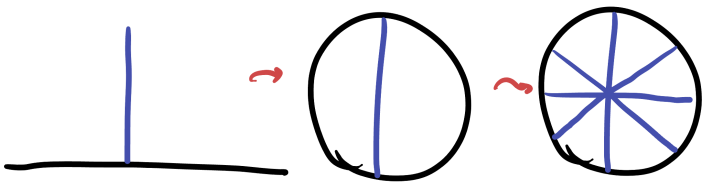


Figure 27.5.: Proving one vertical line in the Half Plane is a geodesic implies that all diameters of the Disk are geodesics.

Next, we can transfer this information *back* to the Half Plane. Consider just a horizontal diameter of the disk. After transferring back, we know this must go to a

generalized circle through $(0, 1)$, and that it must intersect the vertical line at a right angle. This doesn't leave us many options: in fact, it uniquely specifies it! It must be the top half of the unit circle.

So, this half circle is a geodesic. But as soon as we know that, we can use the fact that horizontal translations and similarities of the plane are isometries to get that *all half circles orthogonal to the real line* are geodesics, and all vertical lines as well.

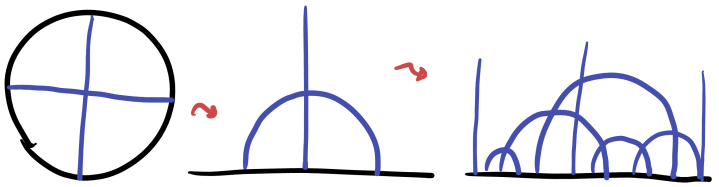


Figure 27.6.: Knowing that diameters of the Disk are geodesics implies all half circles are geodesics of the Half Plane.

Theorem 27.1 (Geodesics in the Half Plane Model)**.** *The geodesics in the disk model are arcs of generalized circles which are orthogonal to the x-axis boundary.*

Finally we perform one more transfer, and move all this knowledge back to the Disk Model. We know that the transfer map preserves generalized circles, and that it is conformal. Thus, it must take generalized circles orthogonal to the $x-$axis to generalized circles orthogonal to the unit circle. These are the geodesics of the Disk.

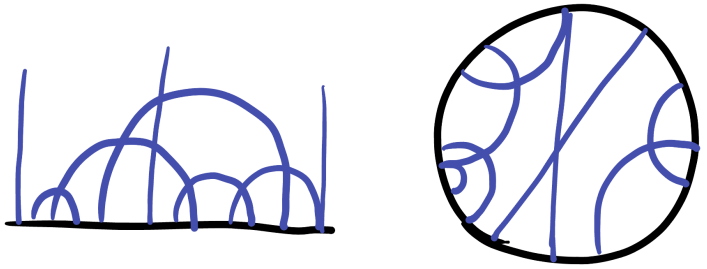


Figure 27.7.: Using a classification of geodesics of the Half Plane to find the geodesics of the Disk

Theorem 27.2 (Geodesics in the Disk Model)**.** *The geodesics in the disk model are arcs of generalized circles which are orthogonal to the unit circle boundary.*

Now we can transfer this information right back to the upper half plane: since the translation between the two models preserves angles and generalized circles, we immediately conclude

27.2.1. Parallels

Now that we explicitly know the geodesics, it is easy to see that this geometry fails the parallel postulate! While we can work with either version, its slightly simpler to consider Playfair's formulation, as it does not require us to actually measure any angles.

Theorem 27.3 (Playfairs Axiom is False in Hyperbolic Geometry)**.**

Proof. This is a proof by *counterexample* we just need to find a a line, where at least two other lines pass through a point not on it, and neither intersects the original line. This is quick work now that we know the geodesics.

First, lets start with the point $(0, 1)$. We are familiar with two geodesics through this point: the vertical line, and the top half of the unit circle. So now, all we need is a third geodesic that doesn't intersect either of these! There are tons of possibilities: let's just take the vertical line at $x = 2$. □

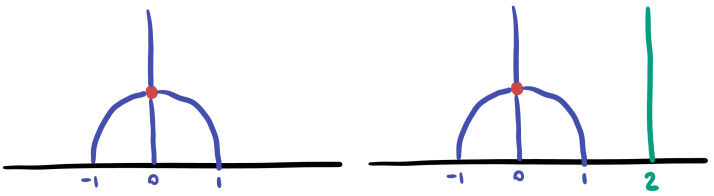


Figure 27.8.: Playfair's Axiom is False in Hyperbolic Geometry.

If we want to work with angles, we can also see that Euclid's original version is false rather immediately:

Proposition 27.5. *Euclid's parallel postulate is false for hyperbolic geometry.*

Proof. Let's work in the half plane model, and consider the vertical geodesic at $x = 0$. We need only find two geodesics that cross it, and have angle sum less than π, but remain disjoint.

For one of them, take the top half of the unit circle. For the second, take a larger circle with center slightly shifted horizontally, say, the circle of radius 3 centered at $x = -1$.

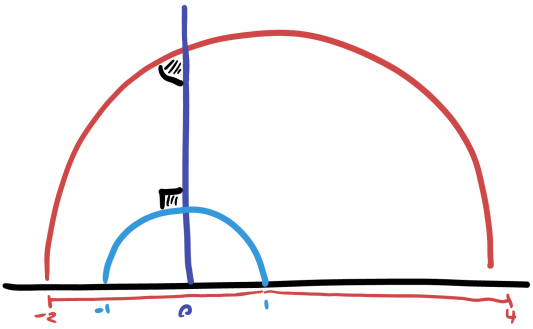


Figure 27.9.: Euclids Fifth Postulate is False in hyperbolic geometry.

Now, because the model is conformal we can figure out what these angles are using Euclidean geometry. But we don't even need their exact value: the first is $\pi/2$ and the second is less than $\pi/2$, so the sum is less than π yet the geodesics are disjoint. Contradiction. □

These two examples, rather straightforward after our classification of geodesics, finish off the Greek's largest open problem. Hyperbolic geometry satisfies Postulates 1-4 without satisfying 5, so there is no possible way to prove 5 from the first four (if so, the fact that the first four are true in $\mathbb{H}^2$ would logically imply the 5th must be true in $\mathbb{H}^2$, and its not).

27.3. Circles

A circle is the set of equidistant pts from a point. We begin our search for circles by formalizing a discussion we previously had, in the Disk Model.

Proposition 27.6. *Hyperbolic circles in the Disk Model about O are Euclidean circles.*

Proof. Let p be some point in the disk which lies at distance r from the origin. Since Euclidean rotations of the disk about O are hyperbolic isometries, we know these rotations do not change hyperbolic distances. Thus, any rotation of p about O is the same distance away from O, and so (by definition) lies on the circle of radius r about O.

But, Euclidean rotations of p about O trace out a Euclidean circle about O! Thus, the hyperbolic circle is also a Euclidean circle. □

Remark 27.1. Note however that the Euclidean radius is probably not r: the Euclidean radius is always < 1 since its inside the unit disk, whereas the hyperbolic radius could be any positive number

Like with Geodesics, we will use this information, bouncing back and forth between the two models, to learn about *all* circles. Carrying this over to the upper half plane describes the circles about $(0, 1)$. By the circle-preserving properties of our map, we know these are taken to euclidean circles in the Half Plane. So this is nice! But also confusing - these circles can't have their Euclidean centers at $(0, 1)$, because they have to *stay within the half plane.*

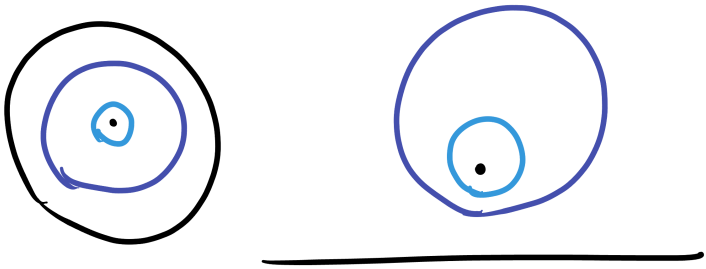


Figure 27.10.: Circles about O in the Disk and about $(0, 1)$ in the Half Plane are both Euclidean circles.

So, while hyperbolic circles appear as Euclidean circles in this model, their centers are closet to the bottom than you think: this makes sense, as distances down low are longer than they appear, and distances up high shorter than they appear, so the center - which is at the actual middle - appears to be shifted down.

What about circles based at other points? Luckily, we understand the isometries that move one point to another in the Half Plane quite well: they are Euclidean translations and similarities! Each of these types of map preserves Euclidean circles, so we see that hyperbolic circles about other points in the plane are *also* all Euclidean circles, though their centers may not be where they seem.

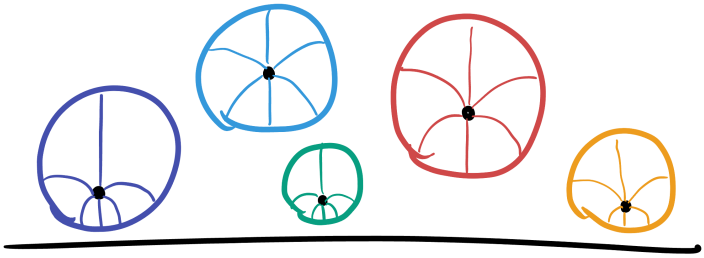


Figure 27.11.: Hyperbolic circles in the Half Plane Model.

Theorem 27.4 (Circles in the Half Plane Model)**.** *Hyperbolic circles in the Half Plane model are represented by Euclidean circles, but their hyperbolic and Euclidean centers do not coincide.*

Exercise 27.2. If a circle's (hyperbolic) center is at height h in the half plane above the x-axis and its radius is r, what are the Euclidean lengths of the radius pointed downwards, and the radius pointed upwards? What's their ratio?

Now let's transfer what we've learned back to the Disk Model. Since the transfer map preserves generalized circles, we can completely understand what happens:

Corollary 27.3 (Circles in the Disk Model)**.** *Hyperbolic circles in the disk model are Euclidean circles, though their hyperbolic center will not coincide with their Euclidean center in general.*

Let's test your hyperbolic intuition at this point: can you tell (without doing computation) if the hyperbolic center should be more towards the center of the Disk model, or more towards it's boundary?

27.4. Curvature

Now that we know a bit about circles, distances, and lines we are in a good position to be able to rigorously confirm that the curvature of our new hyperbolic world is -1. To do so, we need to use the *definition* of curvature, which requires us to know the circumference of circles, so that is where we begin.

We want to choose things to make our calculations as easy as possible: so let's consider the Disk model and look at circles centered at O. Consider the circle of *Euclidean* radius a about the center of the disk. Let's try and find its hyperbolic circumference.

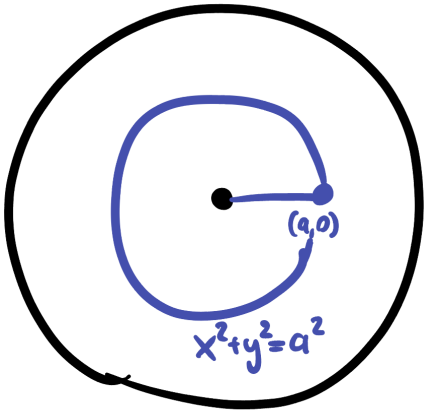


Figure 27.12.: A circle about O in the disk, quantitatively.

We know its Euclidean circumference is $2\pi a$, and we also know that at a distance of a from the origin the scaling factor of the Disk Model is $2/(1-a^2)$. Because this is the same scaling factor at every point along the circle, we can just multiply the Euclidean length by this to get its hyperbolic counterpart:

$$C = 2\pi a \frac{4}{1-a^2} = \frac{4\pi a}{1-a^2}$$

However, this isn't all that we need. Our formula is expressed in terms of the *euclidean* radius a, which is a meaningless quantity in hyperbolic geometry. To find this, we need to do some more calculus.

We do know that straight Euclidean lines through O are geodesics of the model, so to measure the length of the radius of our circle, all we need to do is find the hyperbolic length of $\gamma(t) = (t, 0)$ on $[0, a]$. Calculating infinitesimal length,

$$\gamma' = \frac{2}{1-t^2} \|\langle 1, 0\rangle_{\mathbb{E}^2} = \frac{2}{1-t^2}$$

Thus, the length we seek is

$$\text{length}(\gamma) = 2\int_0^a \frac{1}{1-t^2} dt$$

In a calculus 2 course, you may have seen this integral and immediately thought *ooh, integration by partial fractions!* and that's totally do-able here: in fact it works out rather nice! But another technique works out even nicer, now that we have put in

the effort to learn hyperbolic trigonometric functions: we can do a hyperbolic trig sub!

We know that $\operatorname{sech}^2 + \tanh^2 = 1$ so, $1 - \tanh^2 = \operatorname{sech}^2$, and so substituting $t = \tanh x$ gives

$$\int \frac{1}{1-t^2} dt = \int \frac{1}{1-\tanh^2 x} d(\tanh x) = \int \frac{1}{\operatorname{sech}^2 x} \operatorname{sech}^2 x \, dx$$
$$= \int 1 \, dx = x$$

Converting back to t, since $t = \tanh x$ we have that $x = \operatorname{arctanh} t$, and so the hyperbolic radius is

$$r = \text{length}(\gamma) = 2 \operatorname{arctanh} x \Big|_0^a = 2 \operatorname{arctanh}(a)$$

Now, we have the two pieces of information we need to figure out the relationship between circumference and radius: we just need to eliminate mention of the Euclidean a:

Exercise 27.3. A hyperbolic circle of radius r has circumference $2\pi \sinh(r)$.

*Hint: use the fact that we know $C(a)$ and $r(a)$: solve for a in terms of r, substitute into the circumference, and then use hyperbolic trigonometric identites to simplify.*s

Now we *already proved* that if a space had this relationship between circumference and radius, then its curvature was precisely -1. So, we're done! These maps really do describe the space humanity missed for two thousand years!

Corollary 27.4 ($\mathbb{H}^2$ has constant curvature -1.)**.**

27.5. Polygons

We've seen that the area of a hyperbolic triangle is determined by its angle sum. And, more surprisingly - that its bounded above by π! Can we come to an understanding of this?

Let's think in the Disk Model for a bit. Since space is infinitely large it seems absurd that a triangle can't get very big! But this all has to do with the way geodesics behave. Imagining a large triangle means (in our Disk model) imagining a triangle whose three vertices are all very far away from the center, and thus appear out by the unit circle.

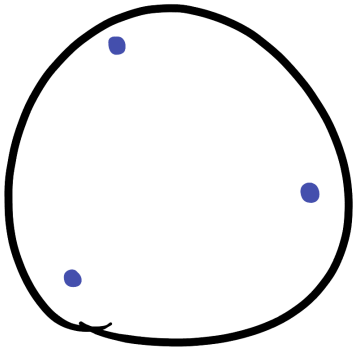


Figure 27.13.: Three far away points make for a large triangle.

To form a triangle from these points, they must be connected together with geodesics. And we know what the geodesics are - they're arcs of circles which are orthogonal to the boundary. Thus, the triangle has *very skinny angles* as the geodesics are almost tangent to one another. And being so skinny, these arms of the triangle can't contain that much area.

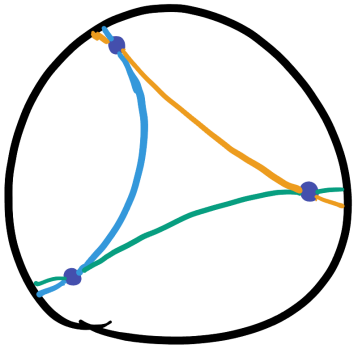


Figure 27.14.: Toward the boundary, geodesics are almost tangent. Thus large triangles have very skinny 'legs', which do not contribute much to their area.

In fact, the biggest triangle one could imagine making would have *infinitely long sides*, and would consist of three geodesics going all the way out to infinity.

Remark 27.2. Of course this isn't actually a triangle as it has no vertices! Mathematicians call it an *ideal triangle*

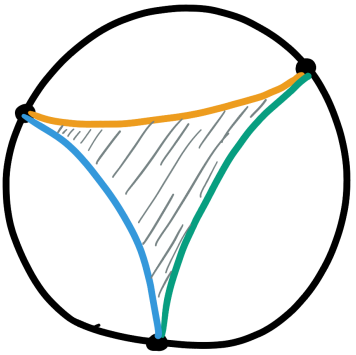


Figure 27.15.: An ideal triangle in the hyperbolic plane.

How big is an ideal triangle? To calculate, its easiest to hop over to the upper half plane. We can choose our three points so that two of them are ± 1 and the other is the projection point - off at infinity. This means our triangles geodesic sides are the unit circle, and two parallel vertical lines

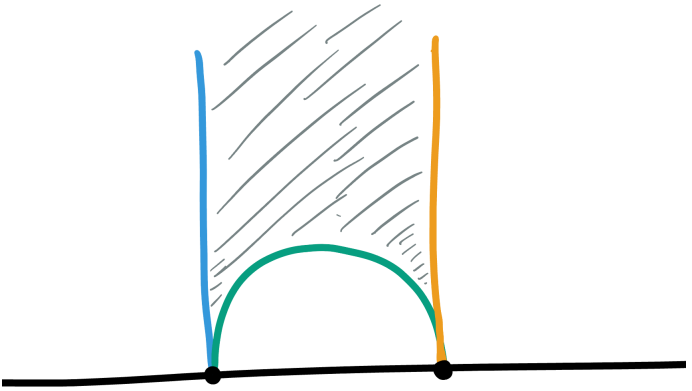


Figure 27.16.: An ideal triangle in the Half Plane model

Its area is given by a double integral of dA: using the scaling factor,

$$dA = \frac{dxdy}{y^2}$$

Setting up the bounds (bottom bound = unit circle, top goes to infinity) we see

$$\text{area} = \iint_T dA = \int_{-1}^{1} \int_{\sqrt{1-x^2}}^{\infty} \frac{dydx}{y^2}$$

Doing the inner integral first:

$$\int_{\sqrt{1-x^2}}^{\infty} \frac{dy}{y^2} = \frac{-1}{y}\Big|_{\sqrt{1-x^2}}^{\infty} = \frac{1}{\sqrt{1-x^2}}$$

Thus, the total area is

$$\text{area} = \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} dx$$

But this integral is quite familiar to us by this point in the course! Its the arc-length of the top half of the unit circle (the integral that defines arccosine).

$$\text{area} = \pi$$

Thus we have it: all triangles in $\mathbb{H}^2$ have area less than or equal to π, because that's the area of the ideal triangle!

28. Life in Curved Space

Its one thing to *do* geometry, the way we do in a geometry course - deriving relationships between triangles circle and geodesics. But its another thing to *feel* it: to try and imagine yourself in the world you are studying, and think hard about what that experience would be like. This is a challenging but rewarding exercise, often not requiring much new geometry but requiring a lot of deep thought, and a lot of actual calculation. In this final chapter, we will attempt to give a taste of what hyperbolic space is like, relying on the material we have developed in the course.

Just like spheres come in may different sizes, so do hyperbolic spaces: so, the first thing we must do in asking ourselves what its like is to decide *which* hyperbolic space we are talking about. This question is actually interesting at all different levels of curvature, as different effects become important at different curvatures. But here in this short chapter we will fix a curvature, and work out the consequences. A convenient way to fix the curvature is to fix the *units* that we measure space in, we can specify a distance R called the *radius of curvature*, defined so that if we measure everything in units of R, we will determine the curvature to be -1. (This is equivalent to instead fixing ahead of time some units, and then considering the hyperbolic space of curvature $-1/\sqrt{R}$ in those units.)

28.1. The Size of San Francisco

Here we fix the radius of curvature to be approximately the radius of San Francisco, $R = 5km$. This will allow us to compare the behavior of *small things* (like humans) to *medium* things (like cities) and *large* things (like planets), and see how in curved space, these different regimes behave quite differently!

In our exploration, our goal will be to ask simple questions about how the world would be, try to deduce what sort of geometry will be relevant to solving them, do the proper computations, and then try to interpret the result.

28.1.1. How Big is the Earth?

Theorem 28.1 (Volume and Surface Area)**.** *The surface area of a sphere of radius r is*

$$\mathrm{SA}(r) = 4\pi \sinh^2(r)$$

(compare this to $4\pi r^2$ in flat space). The volume of a sphere is the integral *of surface area:*

$$V(r) = \int_0^r \mathrm{SA}(r)dr = \pi(\sinh(2r) - 2r)$$

To figure out how big the earth would be, we need to think a bit about what we mean by this question. The earth formed from a collection of rocks in the early solar system, so its *volume* is fixed by the volume of the rocks that was used to make it up. The earths radius and surface area are geometric consequences of this: once we know the volume of rock we simply find the sphere of that size, and that's the earth!

Example 28.1 (Radius of the Earth)**.** First, we look to find the true earths volume, from its radius of 6300km, or 1250SF. The volume of the earth is

$$\frac{4}{3}\pi r^3 = \frac{4}{3}\pi(1250)^3 = 8,181,230,868.7\, SF^3$$

That is, the earth is a little over eight billion cubic San Franciscos! To find the radius of the sphere which has this same volume in hyperbolic space, we need to solve the following equation for r:

$$\pi(\sinh(2r) - 2r) = 8,181,230,868$$

This needs to be solved using numerical methods, but doing so yields a shockingly small answer:

$$r = 11.11.1987, SF$$

So, the earth is only 11 San Franciscos in radius! Remembering we took $SF = 5km$ this comes out to 55.98 kilometers, or 34.78 miles. In hyperbolic space, its closer to get from SF to the Earth's core than it is to get to the south bay!

Why is the earth so small? It all has to do with exponentials: since the volume of a sphere grows *exponentially* with radius, whereas in flat space it grows quadratically. This means there is just *so much more volume* as the radius grows in hyperbolic space, that it doesn't have to be that big to fit all the rocks that make up the earth! This exponential property is also shared by the formula for hyperbolic surface area, which has an amusing consequence:

Example 28.2 (Surface Area of the Earth)**.** Given a radius of 11.19 units, we can find the surface area of the earth by

$$SA = 4\pi \sinh^2(11.19) = 16,468,700,000\, SF^2$$

To understnad what it means that the hyperbolic earth has surface area of 16 billion San Franciscos, we should compare this to the *actual surface area* of the earth we live on. This is (using the Euclidean radius $6300km = 1249SF$)

$$4\pi(1249)^2 = 19,602,972\, SF^2$$

The real earth is only 19 million San Franciscos! So, in hyperbolic space the Earth has 838 times the surface area of our current planet!

This is a lot of extra real-estate! For me, one good way to conceptualize this number is to think about what the depth of the ocean would be. Here, the average depth of the ocean is 2300 meters; but spreading the same amount of water over 838 times more surface area yields a depth of only 4.389 meters, or just under 15 feet!

What about the moon? In Euclidean space the moon's radius is about 27 percent that of the earth (a little over a quarter as big), which means its volume is 2 percent that of the earth (since volume grows with the *cube of radius*).

Example 28.3. The volume of the moon is 0.02 that of the earth, which means in units of San Franciscos,

$$0.02 \times 8,181,230,868.7 = 3624617\, SF^3$$

Solving for the radius of the hyperbolic sphere with this volume, we find

$$r_{\text{moon}} = 9.8\, SF$$

So, the moon isn't that much smaller than the earth at all! This gives us a very good sense of just how quick the exponential growth of volume is. THe difference in radii between the earth and moon is

$$11.19 - 9.8 = 1.39SF = 4.3\text{miles}$$

Since the moon is only 2 percent the earth's volume, this means that 98 percent of the earths volume is contained in the outer shell of radius 4.3 miles, just the outer 38 percent of the radius! But things only get weirder from here, if we look at larger spheres, since everything is driven by exponential growth. What can we say about the sun?

Example 28.4 (Radius of the Sun)**.** The volume of the sun is 1.3 million times that of the earth, which means in units of San Franciscos, the Sun is

$$1,300,000 \times 8,181,230,868.7 = 1,063,560,012,934,045\, SF^3$$

Solving for the radius of a hyperbolic sphere that has this volume, we find that $r = 18.225$ - that is, the sun is only eighteen San Franciscos, or 56 miles in radius! Remember, the Earth is 11.19 San Franciscos in this world, meaning that the sun is only $\frac{18.225}{11.19} = 1.62$ times as big in radius.

28.1.2. How Much of the Earth is Visible?

We've already learned some rather interesting things about the Earth in negative curvature: its simultaneously much smaller (in radius) and much larger (in surface area) than we are accustomed to. But what does it *look like*?

Theorem 28.2 (Distance to the Horizon)**.** *Standing at height h above a sphere of radius R, the horizon in Euclidean space lies at a distance d of*

$$d = R \arccos\left(\frac{R}{R+h}\right)$$

and in hyperbolic space, the analogous formula is

$$d = \sinh(R) \arccos\left(\frac{\tanh R}{\tanh(R+h)}\right)$$

As a warm-up, we can use this formula to find the distance to the horizon here in flat space. At a height of 2 meters =0.002km above the ground the horizon is

$$6,300 \cdot \arccos\left(\frac{6,300}{6,300.002}\right) = 5.019\,km$$

Thus, standing on the beach we can see a little over 5 kilometers, or around 3 miles out to sea. As we know well from experience, moving up a little bit in height lets us see much more: from our classroom on the fourth floor of Harney we can easily see many miles out to sea. Quantitatively this is easy to confirm: if we were at the top of the Salesforce tower (326 meters tall), we could see

$$6,300 \cdot \arccos\left(\frac{6,300}{6,300.326}\right) = 64.09\,km$$

But, what about in hyperbolic space?

Example 28.5 (The Horizon at Different Heights)**.** Measuring in units of San Franciscos, a 2 meter tall human is $2/5000 = 0.0004$ San Francisco's tall. Using the hyperbolic radius $r = 11.19SF$ of earth, we find the horizon lies at a distance of

$$d = \sinh(11.19)\arccos\left(\frac{\tanh 11.19}{\tanh 11.1904}\right) = 0.0282786SF$$

In more useful units, two percent of a San Francisco is 141.39 meters. You can't see very far at all, only a couple hundred feet until the earth has curved enough out of the way to be below the horizon! Moving upwards helps a bit: from the top of the sales force tower (whose height is 0.0652 San Franciscos) the horizon lies at a distance of

$$d = \sinh(11.19)\arccos\left(\frac{\tanh 11.19}{\tanh 11.2552}\right) = 0.349651SF$$

This is around 1.7 kilometers, or just a bit over a mile. From the top of the Salesforce tower you wouldn't be able to see all the way to USF, or even very far out into the bay! But it gets weirder, as we continue to ascend. From the height of a commercial airliner (30,000ft, or 1.828 San Francsicos) passengers can see

$$d = \sinh(11.19)\arccos\left(\frac{\tanh 11.19}{\tanh 13.018}\right) = 0.9869SF$$

Even from miles into the sky, we can only just almost see all of San Francisco. And in fact, this is a fundamental limit: no matter how far above the sphere you are, you can only see up to 5km in any direction before the horizon. Even from space, when you look down at the earth, you would see the city stretch all the way across the earths' disk (though seeing the earth from such a height is another challenge entirely, that we will confront shortly).

Exercise 28.1. Prove this: that as the height limits to infinity you can only see 1 unit of distance along the sphere.

What area of the sphere is this? This question is actually a bit more complicated than it seems at first. We can't just use the formula for the hyperbolic area of a disk, because we're not looking at a disk - we're looking at a region on a sphere!

Theorem 28.3 (Area of a Spherical Cap)**.** *Given a sphere of hyperbolic radius r, the area of a disk of radius d drawn on its surface is given by*

$$\text{area} = 2\pi\sinh(r)^2\left(1 - \cos\frac{d}{\sinh r}\right)$$

Proof. A sphere in hyperbolic space is still a sphere - and we understand the intrinsic geometry of spheres quite well! So, we'll be able to put this to work here. Indeed, we know that on the unit sphere the area of a disk is $2\pi(1 - cosd)$ and if the sphere's Radius is ρ, then the area of a disk of radius d drawn on its surface is

$$2\pi\rho^2\left(1 - \cos\frac{d}{\rho}\right)$$

So, all we need to do is figure out the radius of our sphere. It's tempting to say that this is just r: that's the distance in hyperbolic space to its center after all - but this is not the notion we are looking for here. The radius showing up in the formula above is *the radius the sphere would have, if it were embedded in Euclidean space*, which is where we derived this formula. Since the sphere's area is $4\pi \sinh^2(r)$, we see that in Euclidean space the radius would be $\rho = \sinh(r)$ to have the same area giving

$$\text{area} = 2\pi \sinh(r)^2 \left(1 - \cos \frac{d}{\sinh r}\right)$$

□

Exercise 28.2 (Spheres from a large distance)**.** Explain why if you look at a sphere of large radius from far away, you only see approximately π square units of its surface area.

Hint: Use a taylor series for cos *and explain why its justified to only take the first terms (why is the angle you are taking* cos *of small?)*

The fact that a sphere's horizon is so nearby has far reaching consequences: one of them being the affordability of cell phones.

Example 28.6 (Price of Cell Phones)**.** Cell phones work by using cell towers to collect and re-broadcast signals from phones, but such a signal can't propagate over the horizon!

The tallest modern cell towers are around 100 meters tall (with most cell towers much shorter). From the top of such a tower in flat space, the horizon is 35.4km away, meaning the tower is accessible to approximately $3,936 km^2$ of land area. But in hyperbolic space? From $100m$ high the horizon is only

$$d = \sinh(11.19) \arccos\left(\frac{\tanh 11.19}{\tanh\left(11.19 + \frac{100}{5000}\right)}\right) = 0.198SF$$

or $0.99km$ away, and the area of such a disk is

$$\text{area} = \pi \sinh^2(11.19)\left(1 - \cos \frac{0.198}{\sinh 11.19}\right) = 0.1231SF^2$$

0.123 square units: equivalently $3.15km^2$ or $1.21mi^2$. This is $\frac{3936}{3.15} = 1249$ *times less coverage*. To get similar coverage, you need over a thousand times more towers, making the cell network over a thousand times more expensive.

But its even worse than this: remember the earth's surface area has grown by a factor of 838! Thus cell companies are hit with a double whammy: they need 1249 times more towers per fixed area, and they also have 838 times more area to cover! Overall

them, the cell network needs to be $1249 \times 838 = 1,047,100$ times larger to give the same coverage: expect your cell plan to go up in cost by a factor of one million to pay for this increased overhead!

28.1.3. What does the Earth Look Like?

Theorem 28.4 (Visual Size of a Sphere)**.** *From height h above a sphere of radius R, the angle α that the sphere takes up in your vision in Euclidean space is*

$$\alpha = 2 \arcsin\left(\frac{R}{R+h}\right)$$

and in Hyperbolic space is

$$\alpha = 2 \arcsin\left(\frac{\sinh R}{\sinh(R+h)}\right)$$

Proof.

□

Again, its useful to do some calculations on the Earth in flat space to get our bearings. From a standing height of 2 meters, the earth takes up

$$2 \arcsin\left(\frac{6300}{6300.002}\right) = 3.1365\text{rad} \approx 179.666^\circ$$

If the earth were a flat plane it would take up half of our field of view, or 180 degrees. So, the earth is rather indistinguishable from an infinite plane at human-height (as searching the uninformed corners of the internet show unfortunately all too well). From the top of the sales force tower the earth takes up

$$2 \arcsin\left(\frac{6300}{6300.326}\right) = 178.833^\circ$$

which is just slightly smaller. Even from the height of an airplane (30,000ft = 9.144 kilometers), earth takes up almost half our field of view.

$$2 \arcsin\left(\frac{6300}{6309.144}\right) = 173.83^\circ$$

And from the international space station at 254mi = 408km high, the Earth still looms large, taking up a wider field of view than our eyes provide us (we can see approximately 114 degrees with binocular vision)

$$2 \arcsin\left(\frac{6300}{6708}\right) = 139.8^\circ$$

Now, what happens in hyperbolic space?

Example 28.7 (Size of Earth from Different Heights)**.** At 2 meters above the ground, the earth looks slightly smaller than its Euclidean counterpart, but perhaps not noticeably so.

$$\alpha = 2\arcsin\left(\frac{\sinh 11.19}{\sinh 11.1904}\right) = 176.8^\circ$$

From the height of the the Salesforce tower,

$$\alpha = 2\arcsin\left(\frac{\sinh 11.19}{\sinh 11.2552}\right) = 139.1^\circ$$

That is - the earth looks slightly smaller from this skyscraper than it does in flat space *from the actual space station.* And it only gets wilder, due to the exponential nature of the hyperbolic sine. From the height of an airliner,

$$\alpha = 2\arcsin\left(\frac{\sinh 11.19}{\sinh 13.018}\right) = 18.5^\circ$$

The earth is only eighteen degrees across in your vision! This is about the size of your hand held at arms length away. In hyperbolic space, your airplane flight is rather dark, and you have to look almost straight down to even catch a glimpse of the tiny earth below.

The height the space station orbits above the earth is quite unrealistic in hyperbolic space (for some reasons we'll encounter shortly), but its a good benchmark to evaluate nonetheless, to really appreciate the unrelenting growth of the exponential. At a height of 252 miles = 81 San Franciscos, the earth would appear to be

$$\alpha = 2\arcsin\left(\frac{\sinh 11.19}{\sinh 92.19}\right) =$$

$$0.000000000000000000000000000000007609^\circ$$

This is so fantastically small that not only would the earth be completely invisible to the stations inhabitants, but it would not be detectable by any form of future supertelescope. A ride to the space station would be quite terrifying as the earth rapidly shrinks below you, fading forever from view into the black.

So, when we move away from a sphere of fixed radius, it shrinks rapidly in our vision. But what happens if we stand at a fixed distance from a sphere of different radii?

In Euclidean space, if you were distance h from the surface of a sphere of radius x, the larger the sphere the more and more it would appear to take up half your vision. Precisely, we can see this using our formula:

$$\alpha = 2\arcsin\left(\frac{x}{x+h}\right)$$

As x grows the argument of the arcsine approaches 1, and so the arcsin approaches $\pi/2$ and α approaches π, half your field of view. But what happens in hyperbolic space?

Example 28.8 (Visual Size of Growing Sphere)**.** At a distance h from its surface, the visual size of a sphere of radius x in hyperbolic space is

$$\alpha = 2\arcsin\left(\frac{\sinh x}{\sinh(x+h)}\right)$$

What is this behavior like when x is sufficiently large? Well, $\sinh(x)$ rapidly approaches $\frac{1}{2}e^x$ for large inputs, so we can simplify the argument of arcsine in this approximation as

$$\frac{\sinh x}{\sinh(x+h)} \approx \frac{\frac{1}{2}e^x}{\frac{1}{2}e^{x+h}} = \frac{e^x}{e^x e^h} = \frac{1}{e^h}$$

Thus, for large spheres, how big they actually are is *essentially irrelevant* to how big they appear in your field of view. Even if you know the distance to a sphere well, its impossible to gauge its size visually: even spheres that differ in size by millions of times still take up the same area in your field of view:

$$\alpha \approx 2\arcsin(e^{-h})$$

28.1.4. What's the Gravity Like?

Remark 28.1. There is a precise way to do all of this, using the formulation of gravity in terms of a *gravitational potential*: if ρ is the mass density in space, the gravitational potential U solves $\Delta U = \rho$, where Δ is the *laplacian* differential operator (in Euclidean space, this is $\partial_x^2 + \partial_y^2 + \partial_z^2$). In hyperbolic geometry, gravity follows the same equation, where we simply replace Δ with the *hyperbolic laplacian.* Solving this for a point mass gives the gravitational potential, whose (negative) gradient is the gravitational force. And finally finding the magnitude of this recovers the law stated below $GM/\sinh^2 r$.

Theorem 28.5 (Inverse Area Law of Gravity)**.** *Netwon's law of gravity is usually referred to as the* inverse square law*, but this is only true in flat space. In fact, a more careful reading of the law might be read to say that* gravity, like light, spreads out evenly in all directions*. Thus, a mass M causes a gravitational acceleration on an object at distance r away proportional to M and inversely proportional to the surface area of the sphere.*

$$\frac{M}{\mathrm{SA}(r)}$$

So for Euclidean space we would have $a = \frac{kM}{4\pi r^2}$ *where* k *is the proportionality constant giving the* strength of gravity. *Traditionally, the* 4π *in the surface area of the sphere is absorbed into the constant which is then renamed* $G = k/4\pi$*, or* Newton's constant, *giving*

$$a = \frac{GM}{r^2}$$

In hyperbolic space, the area of a sphere $4\pi \sinh^2(r)$ *also contains a* 4π *so we can continue to use Netwon's constant, getting the adjusted formula*

$$a = \frac{GM}{\sinh^2 r}$$

Conceptually it will be more helpful often to speak of the relative difference of this from Euclidean gravity, than to speak of the absolute numbers. The ratio between these two quantities is

$$\frac{a_{\mathbb{H}}}{a_{\mathbb{E}}} = \frac{\frac{GMm}{\sinh^2(r)}}{\frac{GMm}{r^2}} = \frac{r^2}{\sinh(r)^2}$$

To use this to understand the experience of gravity on the Earth's surface, we will use Newton's other insight, that the gravity of a spherically symmetric body acts just like a point mass located at its center.

Example 28.9 (Gravity on Earth's Surface)**.** On the surface of the earth in Hyperbolic space ($r_{\mathbb{H}} = 11.19$ SF), the gravitational acceleration felt by people is

$$a = \frac{GM_{\text{earth}}}{\sinh(r_{\mathbb{H}})^2}$$

But to actually get the numerical value here we need to think about $units*: we should express Newtons constant not in meters, but in San - Franciscos! Instead of doing this, its much easier to just compute the ratio with Euclidean gravity, which will automatically measure our result in g-forces:

$$\frac{r_{\mathbb{E}^2}}{\sinh(r_{\mathbb{H}})^2}$$

Because the hyperbolic component requires us to work in units of San Franciscos, we need to measure Euclidean radius in those units: $r_{\mathbb{E}} = 6300km = 1249SF$. Thus

$$\frac{r_{\mathbb{E}}^2}{\sinh(r_{\mathbb{H}})^2} = \frac{(1249)^2}{\sinh(11.19)} = 0.00119g$$

Thus, gravity on the hyperbolic earth is only 0.1% as strong as on earth!

This means that we are only held very weakly to the surface of the earth: you can multiply your weight here by 0.00119 to find how much you'd weigh there! This has some pretty scary consequences: since you are just as strong as you are here in Euclidean space (nothing about *you* changed, just the gravitational pull of you to the earth), you can imagine that it's pretty easy to jump *very high.* Too high - I'd say, it turns out even the feeblest jump will launch you into space, never to return again. To see this, we need to calculate the *escape velocity.*

Theorem 28.6 (Escape Velocity)**.** *On the surface of a gravitating object, when you jump you can either fall back to the ground, go into orbit, or escape forever into the void. The dividing line between these* bound states *(falling back, or getting stuck in orbit) and the* unbound states *(getting ejected to infinity) is the* escape velocity*, a speed where if you jump any slower you'll end up bound to the planet, but any faster will send you away forever. Our goal here is to compute the escape velocity for a hyperbolic earth.*

The first step is to figure out how much energy *is required to escape. The gravitational field tries to pull us back down all the way along our trajectory, and the total work it does on us is the integral of its force along our path. Thus, if we have mass m and were to escape all the way to infinity, this would be*

$$\begin{aligned} W &= \int_{r_{\mathbb{H}}}^{\infty} \frac{GMm}{\sinh^2 r} dr \\ &= GMm \int_{r_{\mathbb{H}}}^{\infty} \operatorname{csch}^2(r) dr \\ &= GMm \left(-\coth r \Big|_{r_{\mathbb{H}}}^{\infty} \right) \\ &= GMm(\coth(r_{\mathbb{H}}) - \lim_{r\to\infty} \coth(r)) \\ &= GMm(\coth(r_{\mathbb{H}}) - 1) \end{aligned}$$

This number (which we still have to compute in the appropriate units of San Franciscos) tells us how much energy is needed to escape. But how fast to we need to go? The kinetic energy *of an object with mass m is $\frac{1}{2}mv^2$. To escape to infinity, we need to give ourselves enough kinetic energy to be able to cancel out the pull of the gravitational field: thus, we need*

$$W = \frac{1}{2}mv^2 \implies v = \sqrt{\frac{2W}{m}}$$

Putting these together, we find the escape velocity

$$v_{\text{escape}} = \sqrt{2GM(\coth(r_{\mathbb{H}}) - 1)}$$

To actually compute this quantity we would need to be careful about units: convert everything to San Franciscos, do the calculation, and then convert to something more reasonable to interpret it in the end. However, like before, its easier to instead compute the *ratio* of the hyperbolic to the Euclidean escape velocities, as all these annoying constants cancel out.

Exercise 28.3 (Euclidean Escape Velocity)**.** In Euclidean space with $F = \frac{GMm}{r^2}$, show that the escape velocity is $\sqrt{\frac{GM}{r}}$

Example 28.10 (Ratio of Escape Velocities)**.** Let $v_{\mathbb{H}}$ be the escape velocity from hyperbolic earth, and $v_{\mathbb{E}}$ be the escape velocity from Euclidean earth. Then

$$\frac{v_{\mathbb{H}}}{v_{\mathbb{E}}} = \frac{\sqrt{2GM(\coth r_{\mathbb{H}} - 1)}}{\sqrt{\frac{2GM}{r_{\mathbb{E}}}}} = \sqrt{r_{\mathbb{E}}(\coth r_{\mathbb{H}} - 1)}$$

Now we can plug in some actual numbers and get a value.

Example 28.11 (Escaping from Earth)**.** Measuring all radii in San Franciscos, $r_{\mathbb{E}} = 1249$ and $r_{\mathbb{H}} = 11.19$, we get the ratio

$$\frac{v_{\mathbb{H}}}{v_{\mathbb{E}}} = \sqrt{1249(\coth(11.19) - 1)} = 0.000000476$$

Thus, it takes less than one ten-millionth the speed to escape the hyperbolic earth as it does its euclidean counterpart! Since the escape velocity here is 11.2km/s (or 6.95 miles per second), in hyperbolic space this becomes

$$0.000000476 * 11.2 = 0.0000053 km/s = 0.533 cm/s$$

Thus if you make any movement faster than half a centimeter per second, you'll be immediately ejected from the earth, never to return!

Life on such a world would be very perilous indeed: it's impossible to walk as the mere act of taking a step will launch you beyond orbit! Perhaps we would all live below ground so there was a solid roof above our heads at all times, or put plungers on our feet to hold ourselves fast to the ground.

28.1.5. Can we see the Sun or Moon?

As we live our strange lives on the hyperbolic earth, what do we see in the sky around us? Is it blue? Is there a familiar sun and moon marking the times of day and month? Or are we forever shrouded in blackness?

To answer this, we first need to think about *how far the earth should be from the sun.* Just as the size of the earth and sun have shrank in negative curvature, the size of orbits and solar systems would also shrink, as the suns gravity drops off much quicker. A reasonable model to investigate then is what would life be like if we set the earth-sun separation so the suns gravitational pull is equal to its true value in Euclidean space? Since Euclidean gravity is inverse square and hyperbolic gravity is inverse sinh-square, this amounts to finding the radius r such that $\sinh(r)$ equals the Euclidean distance of 93 million miles, or

$$\sinh(r) = 29933798.4SF$$

Taking arcsinh, we find $r = 17.907$ - that is, the earth is less than 18 San Franciscos from the sun! This is more wild when we realize that the Sun is also 18 San Franciscos across: we are only one sun-diameter away from the sun!

So, at this distance, how big does the sun look in the sky? Calculating much as we did for the earth-size previously, we find

$$\alpha = 2\arcsin\left(\frac{\sinh 18.225}{\sinh(18.225 + 17.904)}\right) = 0.0000192^\circ$$

This is absurdly small - the sun would appear star-like in the sky. This is too small to be interesting, so let's ask another question: how far away would the earth have to be from the sun for it to be as big as we see it here in flat space?

Example 28.12 (Distance to the Sun)**.** The angular diameter of the sun in the sky is about half a degree, so we are looking to solve for at which distance h a sphere of radius 18.225 appears to be half a degree, or 0.00872665radians. This requires solving

$$2\arcsin\left(\frac{\sinh 18.225}{\sinh(18.225 + h)}\right) = 0.0087$$

We can solve this with some algebra:

$$h = \operatorname{arcsinh}\left(\frac{\sinh 18.225}{\sin\frac{0.0087}{2}}\right) - 18.225 = 5.43758SF$$

This is pretty wild - the earth is 11 SFs in radius, and the sun is 18SFs, but for us to be able to see the sun at normal size in our sky, we need to be just over 5 SFs away from it! But the story gets weirder: because the size of the earth is actually *larger* than the distance between the earth and the sun, the size of the sun in the sky varies throughout the day! At high noon, we are at the location on the earth closest to the sun, with only 18 San Franciscos separating us from its firey surface.

But, at sunrise or sunset we have rotated away, and are actually much farther from the sun: of course, the same is technically true on the earth, where we are approximately 6,000km closer to the sun at noon than at sunset. But this is absolutely negligible in comparison to the 93 *million* miles that separate us. In hyperbolic space these numbers are 11 SFs and 18SFs however, which are of the same order of magnitude!

Example 28.13 (The Size of the Sun)**.** At noon we are at a distance of 4.74 SFs from the sun, which itself has radius 18.225, and appears in our sky to be half a degree across. How far are we away from the sun at sunrise or sunset? Drawing ourselves a picture, we see that this distance is the hypotenuse of a right triangle, and so we can use the hyperbolic pythagorean theorem:

$$d = \operatorname{arccosh}\left(\cosh(11.19)\cosh(5.43)\right) = 15.9345SF$$

This is much farther from the sun (though, still less than a single sun-radius away from its surface!). How big does the sun appear from here?

$$\alpha = 2\arcsin\left(\frac{\sinh 18.225}{\sinh(18.225 + 15.9345)}\right) = 0.0000069^\circ$$

Over the course of the day, the sun has changed in size by a factor of $\frac{0.5}{0.0000069} = 72,306$ *times*! It rises in the sky as an almost invisibly small star, and then midday quickly grows by tens of thousands of times to briefly bathe the world in light, before fading into the abyss once more.

Where's the moon in this story? We calculated the size of the moon above to be $r = 9.8SF$, and we can play the same game and ask what distance it must be from the earth so that it appears the same visual size in the sky (at least at some point in time). By sheer coincidence the the moon and sun are both half a degree in our skies, so we are looking to solve the same equation we did for the sun, just with a different input radius.

Example 28.14 (Distance to the Moon)**.** The angular diameter of the moon in the sky is about half a degree, so we are looking to solve for at which distance h a sphere of radius 9.8 appears to be half a degree, or 0.00872665radians. This requires solving

$$2\arcsin\left(\frac{\sinh 9.8}{\sinh(9.8 + h)}\right) = 0.0087$$

We can solve this with some algebra:

$$h = \operatorname{arcsinh}\left(\frac{\sinh 9.8}{\sin\frac{0.0087}{2}}\right) - 9.8 = 5.43758SF$$

This is *the same distance, to five figures after the decimal point*!

This seems very, very strange at first: the sun and moon are very different sizes, why should they orbit the earth at essentially *the exact same distance* to look the same in the sky? But this goes back to something we already calculated: the size a sphere appears in the sky is pretty much independent of its actual size (so long as it is large enough that $\sinh r \approx \frac{1}{2}e^r$), it only depends on the distance to it. So, for the moon and sun to look the same size they must be at the same distance!

This of course has disastrous consequences, as if we orbit the sun, and the moon orbits us, the moons orbit will *pass directly through the center of the sun.* Goodbye moon! But, before the moon is burned - its reasonable to ask what we would see if we looked at its surface from earth at "lunar noon" - when it is highest in the sky. The amount of the moon thats visible would only be a disk of radius

$$d = \sinh(9.8)\arccos\left(\frac{\tanh 98}{\tanh(9.8+5.43)}\right) = 0.999990SF$$

As we expect, we can see only $d \approx 1SF = 5km$ in radius across the moon - meaning when we look up in the sky we will see the lunar disk with just a single crater or two across its surface at a time!

Part VII.

Lorentzian Geometry

29. A Strange Inner Product

In the rest of this book, we bring our tools of geometry to study an space that shares many similarities with Euclidean geometry, but one major difference right at the beginning: the dot product that we use to measure infinitesimal distances and angles has a *minus sign* in one place it used to have a plus.

Definition 29.1 (Minkowski Inner Product)**.** We write $\mathbb{R}^{2,1}$ to mean a 3-dimensional real vector space together with the inner product

$$\langle a, b, c\rangle \star \langle x, y, z\rangle = ax + by - cz$$

Its often useful to write this in matrix form, where J is the diagonal matrix $J = \text{diag}(1, 1, -1)$:

$$v \star w = v^T J w$$

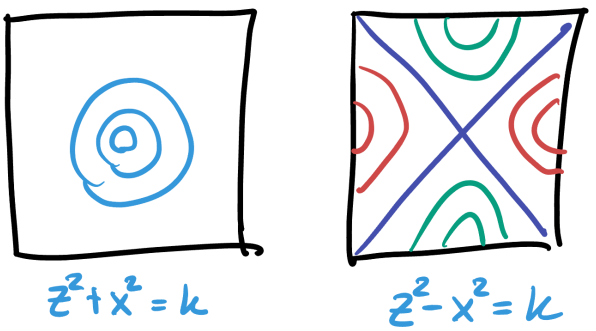


Figure 29.1.: The level sets of the Euclidean inner product are circles. The level sets of the Minkowski inner product are hyperbolas, together with a union of two lines as the degenerate level set $x^2 - z^2 = 0$.

More generally, we write $\mathbb{R}^{n,m}$ for the $n+m$ dimensional real vector space whose inner product is of the above form, with n pluses and m minuses, and as shorthand write $\mathbb{R}^n$ for $\mathbb{R}^{n,0}$ (the standard inner product, where all signs are +)

Definition 29.2. Minkowski space is the 3-dimensional geometry where every tangent space is a copy of $\mathbb{R}^{2,1}$ (just like Euclidean space has the standard inner product on the tangent space at every point).

Our primary goal is to practice using the techniques we have developed in geometry in a new an unfamiliar setting, but this particular space was chosen for a reason: there will be some exciting payoffs along the way.

This is the geometry of special relativity, but our first introduction to it actually comes through its intimate relationship to hyperbolic geometry. We will see that just like the sphere ^{2} naturally lives inside of Euclidean 3-space (as the level set of the dot product $v \cdot v = 1$), hyperbolic geometry $\mathbb{H}^2$ naturally lives inside of Minkowski space (as part of the level set of the dot product $v \star v = -1$).

$$\{(x, y, z) \in \mathbb{R}^{2,1} \mid x^2 + y^2 - z^2 = -1\}$$

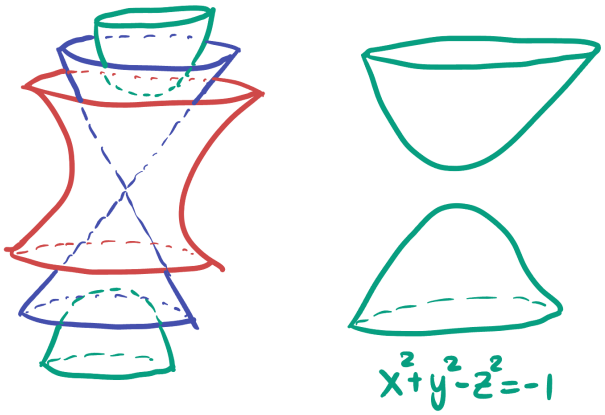


Figure 29.2.: The level sets of the Minkowski inner product in $\mathbb{R}^{2,1}$, and the -1 level set: a hyperoloid of two sheets.

29.1. Isometries of Minkowski Space

As a first step to understanding this, we aim to take as much of our understanding of ^{2} inside of $\mathbb{R}^3$ as we can and build analogies to the hyperboloid in Minkowski space. In Euclidean geometry, we found the isometries of the sphere inside of the isometries of Euclidean space, as precisely the isometries which fix the origin (all others were translations). So here we will attempt to classify the isometries of Minkowski space, and inside of this, find those that preserve the hyperboloid.

Definition 29.3. Recall that if X is any geometry (where we write $\langle -, \rangle$ for its inner product on each tangent space), an isometry is a map that preserves this inner product, a $\phi : X \to X$ where at each point p, and for every $v, w \in T_p X$ we have

$$\langle v, w \rangle = \langle D\phi_p(v), D\phi_p(w) \rangle$$

The same is true for Minkowski space, where a Minkowski isometry is any map $\phi : \mathbb{R}^{2,1} \to \mathbb{R}^{2,1}$ such that

$$v \star w = D\phi_p(v) \star D\phi_p(w)$$

Theorem 29.1. *Translations $T(x, y, z) = (x+a, y+b, c+z)$ are isometries of Minkowski space.*

Exercise 29.1. Prove this via a computation

Since a translation can take any point to any other point, we see already that Minkowski space is *homogeneous*. Thus, the only remaining isometries to consider are those that fix the origin (every isometry can be built as a composition of one that fixes the origin, and a translation). Studying these generally sounds difficult, so we start by studying linear isometries: recall an isometry ϕ is linear if it is given by matrix multiplication: $\phi(p) = Ap$ for some 3×3 matrix A, for every point $p = (x, y, z)$.

Theorem 29.2. *Show that a linear map $\phi(p) = Ap$ is a Minkowski isometry if and only if $A^T JA = J$ for $J = \mathrm{diag}(1, 1, -1)$.*

This reduces the classification of linear isometries to understanding solutions of a matrix equation. We can understand such solutions in analogy with the more familiar equation $A^T A = I$ specifying linear isometries of $\mathbb{R}^3$

Theorem 29.3. *Show that solutions of the equation $A^T JA = J$ are matrices*

$$A = \begin{pmatrix} | & | & | \\ a_1 & a_2 & a_3 \\ | & | & | \end{pmatrix}$$

Where a_i and a_j are Minkowski-orthogonal if $i \neq j$, and each column is of Minkowski norm ± 1 (precisely, $a_1 \star a_1 = a_2 \star a_2 = 1$ but $a_3 \star a_3 = -1$)

Its easy to find some such matrices, using things we already know about Euclidean geometry.

Theorem 29.4. *Show that any Euclidean rotation of the x, y plane that fixes the z axis is a Minkowski isometry. Show that reflecting in the xy plane, so $R(x, y, z) = (x, y, -z)$ is also a Minkowski isometry.*

Thus, we can understand isometries that rotate the hyperboloid or exchange the two sheets easily, as they are just the Euclidean isometries we are used to! But there are more isometries of Minkowski space than just this: thinking back to the sphere in $\mathbb{R}^3$, we have essentially just found the "rotations about the z axis" so far, and a reflection. What are the analogs of rotations about the other axes?

Since we understand rotations in the xy plane just fine, we can restrict ourselves to isometries which do no such rotation: for specificity, we can focus on isometries for the moment which fix the x axis. This let's us drop the dimension of our problem by one, and think about just the slice $x = 0$, or the yz plane.

29.2. Isometries of $\mathbb{R}^{1,1}$

On this slice of the hyperboloid, our inner product has went from $\langle a, b, c\rangle \star \langle x, y, z\rangle = ax + by - cz$ to just

$$\langle b, c\rangle \star \langle y, z\rangle = by - cz$$

This is a 2-dimensional plane with one $+$ and one $-$, so is a copy of $\mathbb{R}^{1,1}$. This is an incredibly useful space because it incorporates all the strange behavior of Minkowski space in a nice 2 dimensional picture we can see. Here, we imagine this space as what we get from looking at Minkowski 3-space with the x axis pointed directly at us: isometries of $\mathbb{R}^{1,1}$ correspond to isometries fixing the x-axis, which in turn we can think of as the Minkowski analog of 'rotations about the x axis' in $\mathbb{R}^3$.

Theorem 29.5. *Show that the Linear isometries of $\mathbb{R}^{1,1}$ are given by matrices of the form*

$$\begin{pmatrix} a & b \\ b & a \end{pmatrix}, \qquad a^2 - b^2 = 1$$

As well as these followed by a reflection in the z axis, $(y, z) \mapsto (y, -z)$ (which then switches the upper hyperboloid and the lower one) Hint: Find the Minkowski-Orthonormal Bases of $\mathbb{R}^{1,1}$!

The matrices given above are exactly the linear isometries of $\mathbb{R}^{1,1}$ which preserve the upper hyperbola, and so correspond to isometries of Minkowski space that preserve the upper hyperboloid and fix the x axis! Like in Euclidean space, we can parameterize these isometries using trigonometric functions: there we used that $a^2 + b^2 = 1$ is satisfied by $a = \cos t, b = \sin t$, and here we use $a^2 - b^2 = 1$ is the defining trigonometric relation for $a = \cosh t$ and $b = \sinh t$

Corollary 29.1. *The linear isometries preserving the upper hyperboloid of* $\mathbb{R}^{1,1}$ *are all of the form*

$$\phi(y,z) = \begin{pmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix}$$

Its really instructive (and confusing) to use these matrices to try and understand the geometry of $\mathbb{R}^{1,1}$. Starting with the vectors $e_1 = \langle 1, 0\rangle$ and $e_2 = \langle 0, 1\rangle$, we can easily check these are Minkowski orthogonal ($e_1 \star e_2 = 0$): thus if ϕ is any linear isometry we know $D\phi(e_1)$ and $D\phi(e_2)$ will also be Minkowski-orthogonal.

Exercise 29.2. Choose some isometries (just pick some value of t and plug it into the matrices above and get decimal approximations, to make it easy to work with) and use them to get a collection of different Minkowski-orthonormal bases in the plane. Draw them! What do they look like? When t gets large, what is happening to the vectors?

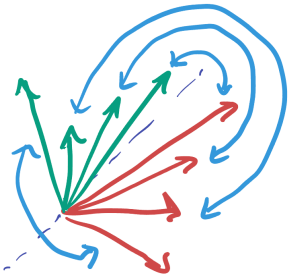


Figure 29.3.: Various pairs of Minkowski-orthonormal vectors (each pair is grouped by the blue arrows). Thus our coordinate description of Minkowski space can be very misleading: each vector in this diagram is length 1, and within each pair the green and red vector are orthogonal!

29.3. Distances in Minkowski Space

Like in any of the geometries we've previously visited, computing the lengths of various curves in Minkowski space will teach us a lot about it, and eventually will help us track down the geodesics. There is one small difficulty in generalizing what we already know however - it's that our inner product can now spit out negative numbers, so our definition of *infinitesimal length* is not well defined, as it had a square root in it. This is an easy fix: just introduce an absolute value! If $v \in \mathbb{R}^{n,1}$ is a tangent vector, we define its Minkowski infinitesimal length as

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

Given a definition of infinitesimal length, we can define the length of curves exactly as in Geometry class: if $\gamma : I \to \mathbb{R}^{n,1}$ is a curve its derivative $\gamma'(t)$ gives a tangent vector based at $\gamma(t)$ for each t, and the infinitesimal length $\|\gamma'\|$ represents an infinitesimal segment of arc. Thus, the lenght of γ is recovered via integration:

$$\text{Length}(\gamma) = \int_I \|\gamma'(t)\|\, dt$$

Where $I = [a, b]$ is the domain of γ.

Happily, since the inner product is so close to the Euclidean inner product, many simple lengths are easy to compute. The first three exercises below show this by having you find the lengths of several straight lines in the Minkowski plane. But beware - easy to compute does not mean easy to understand!

Example 29.1. In $\mathbb{R}^{1,1}$ that the length of a segment of the y axis the same as its Euclidean length: that is, if $\gamma(t)$ parameterizes the y-axis from $y = a$ to $y = b$, then $\text{Length}(\gamma) = b - a$.

Example 29.2. In $\mathbb{R}^{1,1}$ that the length of a segment of the z axis the same as its Euclidean length: that is, if $\gamma(t)$ parameterizes the z-axis from $z = a$ to $z = b$, then $\text{Length}(\gamma) = b - a$.

Example 29.3. In $\mathbb{R}^{1,1}$ show the length of any segment of one of the diagonals $y = \pm z$ is *ZERO*! Thus, Minkowski space has the extremely strange property that curves connecting two distinct points are allowed to have zero length!

In the final exercise here of 'simple-to-compute-lengths' we will look at measuring the length of the hyperboloid $y^2 - z^2 = -1$ in $\mathbb{R}^{1,1}$: amazingly, due to the strange dot product we are using, this length *also* turns out to be easy to compute!

Example 29.4. In $\mathbb{R}^{1,1}$, we can parameterize the upper sheet of the hyperbola $y^2 - z^2 = -1$ by $(y, z) = (\sinh t, \cosh t)$. Show that the Minkowski length of this hyperbola between $(0, 1)$ and $(\sinh T, \cosh T)$ is exactly T.

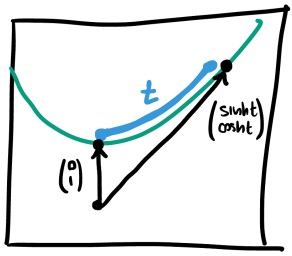


Figure 29.4.: The arclength along a hyperbola in Minkowski space.

The length of segments of hyperbolas is a useful thing to compute in hyperbolic geometry - as these are the geodesics! The fact that this computation comes out so cleanly in the hyperboloid model is one reason this model is good for computation: in fact in almost all computer programs I write using hyperbolic geometry, I do all the computations in the hyperboloid model because of this.

29.4. Proving the Hyperboloid is Hyperbolic Space

We have learned enough about Minkowski space through its isometries and distances to embark on our main goal: proving that the hyperboloid really is a copy of hyperbolic space! Below is an approach to showing they are the same space, without constructing an isomorphism directly, which we will do so in a series of steps.

Proposition 29.1. ***STEP 0: The Hyperboloid is Preserved by Linear Minkowski Isometries***

Proof. All linear isometries of Minkowski space preserve the inner product $x^2+y^2-z^2$, and so they preserve the Minkowski norm of vectors. That is, if ϕ is any isometry and p is a point of Minkowski space with $\|p\| = c$ then $\|\phi(p)\| = c$ as well. Since the hyperboloid is the level set $\|p\| = -1$, any Minkowski isometry takes the hyperboloid to itself. □

Proposition 29.2. ***STEP I: The Hyperboloid is Homogeneous*** *For any two points p, q on the hyperboloid $x^2 + y^2 - z^2 = -1$, there is an isometry of the hyperboloid taking p to q.*

Proof. Like in previous cases, its enough to show your favorite special point can be taken to any other point, if $o \to p$ and $o \to q$ we can compose one with the inverse of the other to send $p \to q$.

Here, a natural special base point for the hyperboloid is $(0, 0, 1)$. If $p = (x, y, z)$ is any other point on the hyperboloid we may use a Euclidean rotation in the x, y direction to rotate p into the xz plane (such Euclidean rotations are Minkowski isometries, as they preserve the Minkowski inner product - it agrees with the Euclidean inner product on the xy directions!) Thus, without loss of generality we may take $p = (x, 0, z)$.

But now, ignoring the y coordinate, we have a point of $\mathbb{R}^{1,1}$, which lies on the hyperboloid $x^2 - z^2 = -1$. We can parameterize this hyperboloid by $(\sinh t, \cosh t)$, so for some particular $t_0 \in \mathbb{R}$ we may write $(x, z) = (\sinh(t_0), \cosh(t_0))$. But, by our classification of isometries, we know that the matrix

$$\begin{pmatrix} \cosh(t_0) & \sinh(t_0) \\ \sinh(t_0) & \cosh(t_0) \end{pmatrix}$$

Is an isometry, and this takes $(0,1)$ to $(\sinh t_0, \cosh t_0)$. Thus, combining this wtih our rotation, we can take $(0,0,1)$ to p, and thus p to q. □

Proposition 29.3. ***STEP II: The Hyperboloid is Constant Curvature*** *Isometries preserve curvature, so whatever geometry this is, the curvature is the same at every point. Since its constant curvature, its either the sphere the Euclidean plane or hyperbolic space.*

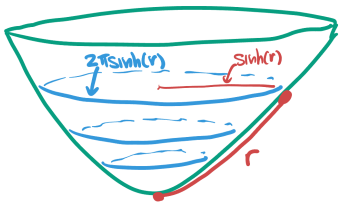


Figure 29.5.: Circles on the hyperboloid of Minkowski radius r have circumference $2\pi \sinh(r)$.

Proposition 29.4 (STEP III: The Curvature is Constant -1).

Proof. To compute curvature, we need to find the formula of circumference of a circle in terms of radius $C(R)$, and compute the limit

$$\kappa = \frac{-1}{2\pi} C'''(0)$$

We can find circles of the hyperboloid using isometries: we know that rotations in the xy plane are Minkowski isometries that fix $(0,0,1)$, so if $p = (x,y,z)$ is any point of the hyperboloid and R is any rotation about the z-axis,

$$\text{dist}(O,p) = \text{dist}(R(0), R(p)) = \text{dist}(O, R(p))$$

Thus, p and $R(p)$ lie on the same circle about O! Thus, taking an arbitrary point $(x,0,z)$ on the hyperboloid, we find that the circle containing this point is

$$\begin{pmatrix} \cos t & -\sin t & 0 \\ \sin t & \cos t & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ 0 \\ z \end{pmatrix} = \begin{pmatrix} x\cos t \\ x \sin t \\ z \end{pmatrix}$$

So we know the circles about O of the hyperboloid, but we don't know what the radii are: for that, we need to find a geodesic. Recall that geodesics can be found as the lines of symmetry - curves that are fixed by isometries. And, one isometry of Minkowski space is $(x,y,z) \mapsto (x,-y,z)$ This fixes the set $y = 0$, so the hyperbola containing $(x,0,z)$ is in fact a geodesic!

For any such $(x, 0, z)$ we know that since $x^2 - z^-1$, we can find some R such that $x = \sinh R$ and $z = \cosh R$. And, from our previous calculation, we know that the length between O and $(\sinh R, \cosh R)$ is just R! Thus, the circle parameterized by

$$\gamma(t) = \begin{pmatrix} \sinh R \cos t \\ \sinh R \sin t \\ \cosh R \end{pmatrix}$$

has a radius of R, and all we need to compute is its circumference. This is done via an integral of $\|\gamma'\|$, which is just $\sinh R$: thus

$$\text{Length} = \int_0^{2\pi} \sinh R dt = 2\pi \sinh R$$

So $C(r) = 2\pi \sinh r$ and $\kappa = -1$: this is hyperbolic space! □

How to tell? Compute some quantity we know in these geometries: say the circumference to radius of circles. Start around (0,0,1): we found the segment of length R ends at (0,sinh R, cosh R). Thus, a circle is a horizontal circle of (Euclidean) radius sinh R. Easy to find arclength in Minkowski space is $2\pi \sinh R$. Thus its hyperbolic space!

29.4.1. Calculating Distance in the Hyperboloid

Distance between p and q is the length of a geodesic. If p, q are arbitrary points then use an isometry to move one to the basepoint $(0, 0, 1)$. Now need to measure just distance from this to some p.

Geodesics in hyperbolic space are equivalently (1) distance minimizing (2) straightest curves (3) lines of symmetry. Easy to find lines of symmetry of the hyperboloid: a reflection in a vertical plane is an isometry, so the geodesic from $(0, 0, 1)$ to p is just such a segment of a hyperbola. Without loss of generality we can actually rotate things so p lies in the (x, z) plane, and is $(\sinh a, \cosh a)$ for some a. But we know from before the length of this segment is just a! Thus this is the distance.

How can we write this in a geometrically meaningful way? Note that

$$(0, 0, 1) \star (\sinh a, 0, \cosh a) = -\cosh a$$

And so we can express the distance between points $X = (0, 0, 1)$ and $Y = (\sinh a, 0, \cosh a)$ as

$$d = \operatorname{acosh} |X \star Y|$$

Now that we have expressed things in terms of the Minkowski dot product, we can learn alot! The transformations we used to take our general pair of points p, q to X and Y were all Minkowski isometries, and so they preserved distances on the hyperboloid as well as the inner product! That means $X \star Y = p \star q$, and

$$\operatorname{dist}(p, q) = \operatorname{acosh} |p \star q|$$

We can compute distances in the hyperboloid model by just using dot products! This makes something as complicated as computations in curved space reducible to linear algebra! This makes it very helpful to employ the hyperboloid model in computations done on a computer.

30. Geometry of Minkowski Space

Our introduction to Minkowski space in the last chapter essentially provided a playground for us geometers,

- We got to try out our definitions (of isometries, distances, etc) in an unfamiliar context, which was still quite close to things we've seen before (all we switched was one negative sign in the dot product!)
- We were able to use our geometric skills to discover a new model of hyperbolic geometry living inside this world!

This provides a nice end to our story of hyperbolic space, and our approach to geometry as a whole. But it also serves as the *beginning* to a new and wonderful story of geometry and its interaction with physics. In these final chapters we endeavour to tell a small bit of this story, where Minkowski space itself (and not merely the hyperboloid lying within it) takes center stage.

To begin, we need to take a deeper dive into the geometric properties of Minkowski space.

30.1. Positive and Negative

As a geometry, Minkowski space is a $n+1$ dimensional real vector space M where each tangent space space $T_pM \cong \mathbb{R}^{n,1}$ comes equipped with the inner product $\star$.

At each point $p \in M$ there is a null cone of vectors of length zero. Any isometry of M preserves null cones: it sends the null cone at p to the null cone at $\phi(p)$.

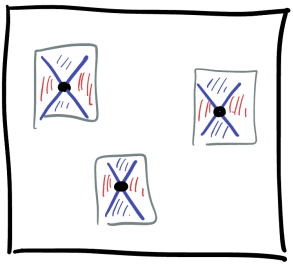


Figure 30.1.: The X shapes (cones) in the tangent space at each point of Minkowski space are the level sets of the dot product, much like circles in the tangent space of Euclidean/Hyperbolic geometry were level sets of their dot products.

One of the most effective ways to study a geometric space is with curves: for example, in Riemannian geometry we used curves to help us define geodesics, and then learned a ton of geometry from trying to classify geodesics. Let's attempt the same here.

First, we recall the definition of a regular curve. In Riemannian geometry, we said this is a curve $\gamma : I \to M$ where the derivative is never zero (or, equivalently, its norm is never zero). We copy this definition in Minkowski space

Definition 30.1 (Regular Curve:)**.** A curve $\gamma : I \to M$ is regular if $\|\gamma'(t)\| \neq 0$ for all $t \in I$.

However note here that $\gamma' \neq 0$ is not equivalent to $\|\gamma'\| \neq 0$ as there are null vectors. Nonetheless the definition we chose for regular is the "correct" one, as the reason one imposes regularity on curves is that one often wishes to do computations that require dividing by $\|\gamma'|$.

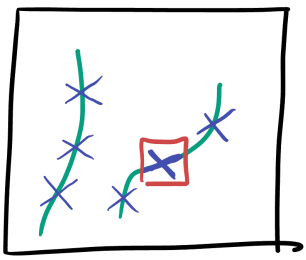


Figure 30.2.: A regular curve (left) and a non-regular curve (right). The curve on the right has a single point where its derivative is tangent to the null cone, meaning the magnitude is zero.

This condition makes the collection of regular curves in Minkowski space quite different than in Riemannian geometry: we can sort all regular curves into two classes (which for now, we will call positive and negative).

Definition 30.2 (Positive and Negative Curves.)**.** We say a regular curve is positive if $\gamma'(t)$ has positive norm for all t, and is negative if the norm is negative for all t.

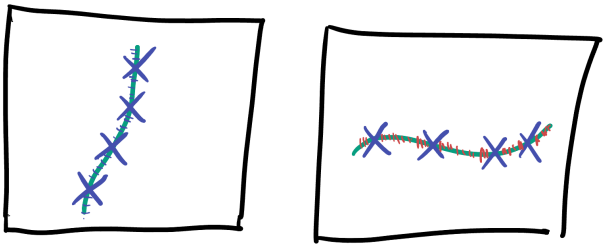


Figure 30.3.: Positive (right) and Negative (left) curves in $\mathbb{R}^{1,1}$.

Theorem 30.1. *Regular Curves are Either Positive or Negative. Precisely, let γ be a regular curve: then it is either a positive or negative curve.*

Proof. Apply the intermediate value theorem to the function $\|\gamma'\|$. □

Of the non-regular curves, most have points of positive and negative norm - but a few special ones do not: they have zero norm *everywhere*. These are worthy of a special name

Definition 30.3. A curve γ is null if $\|\gamma'(t)\| = 0$ for all t.

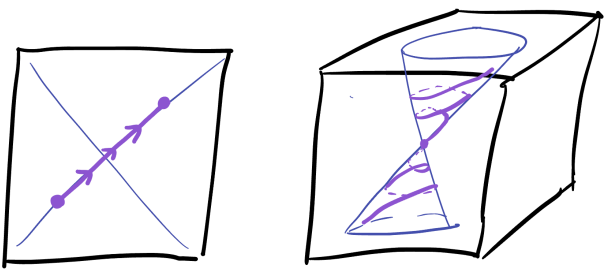


Figure 30.4.: Null curves in $\mathbb{R}^{1,1}$ and $\mathbb{R}^{2,1}$. In $1+1$ dimensions, any null curve is just a subset of one of the lines forming the null cone. In higher dimensions affine lines lying on the cone are also null curves, but there are more interesting examples as well (right).

This sorting of all regular curves curves into two classes has profound implications: it let's us actually sort the *points* of M into classes

Definition 30.4 (Positive and Negative Pairs of Points.)**.** Let $p, q \in M$ be two points. We say that p and q are positively separated if there is a positive curve starting at p and ending at q. They are negatively separated if there is a negative curve starting at p and ending at q. And, they are null separated if there is a null curve connecting them.

This definition requires some checking to be sure its well-defined:

Theorem 30.2. *If two points are connected by a negative curve, then any regular curve connecting them must be negative. Same for positive pairs.* Hint: A sneaky use of Rolle's theorem to the z coordinate, for negative curves!

Using affine lines (which are easy to tell when they are positive or negative curves, just using the norm of their derivative) we can sort points into positives and negatives with this definition.

Exercise 30.1. For the origin, show that the points that are positively separated from O are those outside (horizontally) of the X made by the lines $z = \pm x$, and the points that are negatively separated are inside (vertically above an d below) the X.

This same thing holds true at every point: given a point $p \in M$ we can sort all other points q so that the positive pairs (p,q) are all the points lying outside of an X centered at p and the negative paris are all the points inside that same X.

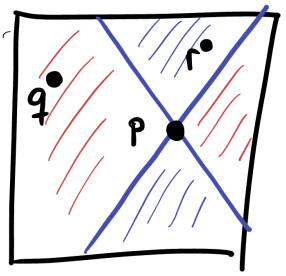


Figure 30.5.: Points with positive separation (q) and negative separation (r) from a point p.

30.1.1. Geodesics

Now that we understand a bit of isometries, regular curves, and positively/negatively separated points, we can start to talk about geodesics in Minkowski space. This discussion is more subtle than in Riemannian geometry, for several reasons.

We've already dealt with one major difference: the fact that not all pairs of points in Minkowski space are created equal - for some pairs, *every* regular curve connecting them is a positive curve, and for others every regular curve is a negative curve (and finally, for the remaining pairs of *null-separated* points, there are *no regular curves at all joining them*).

Thus, we already expect there to be two notions of geodesic, a *positive geodesic* as some optimization problem over the space of positive curves, and correspondingly a *negative geodesic* for pairs of negatively separated points.

But there's one more subtlety to confront: we already know that Minkowski space has curves of *total length zero* between distinct points. This radically different behavior that Euclidean space suggests a potential problem with our usual notion of geodesic as *minimizing* - if we can make a regular curve nearby to a curve of zero length, we might expect that regular curve to have a *very short length* - and perhaps taking the infimum over all regular curve lengths actually gives zero, which is not realized by any regular curve.In fact, exactly this worry happens.

Theorem 30.3 (A truly mindbending example)**.** *Consider the following two curves in* $\mathbb{R}^{2,1}$*: the first curve* γ *is just the vertical segment from* $(0,0)$ *to* $(0,4)$*. THe second curve* c *is piecewise: it begins with the affine line segment connecting* $(0,0)$ *to* $(1,2)$ *and then continues as the affine line segment connecting* $(1,2)$ *to* $(0,4)$*.*

Which is longer? Now do the same for the piecewise curve that bends at the point $(a,2)$ *instead of* $(1,2)$*, for* $a \in [0,2]$*. Show that there is a curve connecting* $(0,0)$ *to* $(0,4)$ *with arbitrarily short nonzero length:*

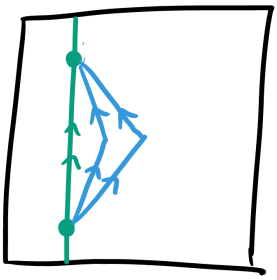


Figure 30.6.: The blue curve is *shorter* in Minkowski length.

Exercise 30.2. Can you construct a similar example, for a pair of positively separated points?

One concern here might be that these curves described are not regular - they have a corner (they are piecewise regular, however). This is not actually a technical concern as it is fine (and often more convenient) to just work with the class of piecewise regular curves from the start. But even if you choose not to do so, these examples still point the way:

Example 30.1. Given two negatively separated points (without loss of generality we can take them to be $(0,0)$ and $(0,a)$ after applying isometries), there are regular curves of arbitrarily short length connecting them. The idea is to approximate the piecewise curve above by something smooth, in this case, a segment of a hyperbola

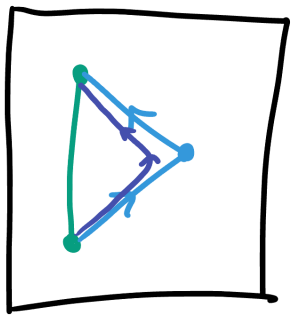


Figure 30.7.: Smoothing a piecewise regular curve to a regular curve, while maintaining the property of arbitrarily short length.

The situation is even worse that these examples make it appear - here we found curves that were very short, but were also *very far away* from our original curve. Perhaps one might hope there aren't any *actually nearby* to our original curve - so maybe its still "locally" nice. But this intuition is rather shaky; by modifying the idea above to introduce lots of small crinkles instead of one big deviation....

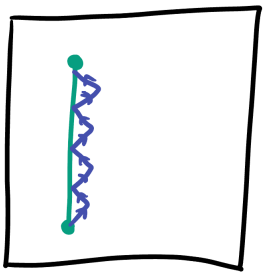


Figure 30.8.: Finding a very short piecewise regular curve arbitrarily close to a long curve.

Exercise 30.3. If you constrain a curve to never get more than ϵ away from a vertical line in its x coordinate, what can you say about its length? Can it be arbitrarily close to zero, or is there some lower bound?

What about if its a regular curve?

Corollary 30.1. *There are no length minimizing regular curves in Minkowski space: given any pair of positively or negatively separated points, the infimum of the lengths of all regular curves joining them is zero, but there is no regular curve of length zero joining them.*

Thus, the inifmum is not *a minimum, so the minimum does not exist!*

This example teaches us two important things: first the formal definition of geodesic can't be directly borrowed from Riemannian geometry, but secondly we can see its clearly not even the right notion! In Riemannian geometry, its easy to make a curve *longer*, by wiggling it, curves of minimal length are the right sort of *optimal objects* to seek. But in Minkowski space, its easy to make a curve *shorter* by wiggling it; and in fact, its difficult to make a curve longer! Almost everything you try shortens it…so perhaps the right thing to do is turn our intuition on its head and define our optimal objects as the *length maximizing curves*. Amazingly, for negative curves this works!

Definition 30.5 (Negative Geodesics)**.** Let p, q be two negatively separated points in Minkowski space. Then in the set of all regular curves joining p to q, there is a unique curve of *maximal length*.

We call this curve the *geodesic from p to q*.

This is a definition that justifies itself with a claim: (that there is a maximum, and as a bonus its unique!) So, we should check this!

Theorem 30.4. *If γ is an affine line connecting two negatively separated points, then γ is globally length maximizing.*

Proof. Without loss of generality we can take our points to be $(0,0)$ and $(0,a)$ after using some isometries, and so γ is the curve $\gamma(t) = (0,t)$. Now let $\alpha(t) = (x(t), z(t))$ be any other curve joining $\alpha(0) = (0,0)$ and $\alpha(1) = (0,a)$. Writing out its length, we see

$$\begin{aligned}\text{Length}(\alpha) &= \int_I \sqrt{|\alpha' \star \alpha'|}\, dt \\ &= \int_I \sqrt{|(x')^2 - (z')^2|}\, dt \\ &= \int_I \sqrt{(z')^2 - (x')^2}\, dt\end{aligned}$$

Where the last equality follows as since α is regular and joins negatively separated points, we know that $\alpha'(t)$ is negative for all t, and so taking the absolute value is the same as multiplying by a negative.

But, no matter what $x(t)$ is we know $x'(t)^2 \geq 0$ and so for all $t<$

$$(z')^2 - (x')^2 \leq (z')^2$$

Both sides of this are positive and the square root is an increasing function, so this implies

$$\sqrt{(z')^2 - (x')^2} \leq \sqrt{(z')^2} = |z'|$$

and finally, $|z'| \geq z'$, so stringing these inequalities together and integrating yields

$$\int_I \sqrt{(z')^2 - (x')^2}\, dt \leq \int_I z'\, dt$$

The first integral here is none other than the length of α, and the second integral here is easily evaluated via the fundamental theorem of calculus:

$$\int_{[0,1]} z'\, dt = z(1) = z(0) = a - 0 = a$$

Thus for any such curve α we have $\text{Length}(\alpha) \leq a$. As this is precisely the length of the affine line γ we have

$$\text{Length}(\alpha) \leq \text{Length}(\gamma)$$

□

As a corollary of this, looking closer at the argument above we can see that in fact no other curve can be as long as γ, so this is the *unique* maximum

Exercise 30.4. Geodesics connecting negatively separated points are unique.

Hint: if α is distinct from γ then its x coordinate must be nonzero at some point. Because it is continuous, this means there must be some small interval where the x coordinate is nonzero, and on this interval you can show the length of α is strictly less* than the length of γ. On the rest of the curve we can get $\leq$ as above, and putting it together we get the inequality is *strict*: α must actually be shorter than γ!

In fact, there's a way to make this craziness sound not so strange after all. Remember that we *defined* infinitesimal arclength by using an absolute value for negative curves, since the dot product yields a negative number. So, finding the *maximal length* is really finding the *maximum absolute value* which is the same as finding the *most negative* (since we know the original numbers are all negative). But the most negative is the minimum! So, we could simply modify our definition of the length of a negative curve to *remember that the dot product is negative*

$$L(\gamma) = -\int_I \sqrt{|\gamma' \star \gamma'|}\, dt$$

And with this new definition all curves have negative length, the *maximum* is not achieved (as curves can have lengths arbitrarily close to zero) but the *minimum is*: curves of minimal length are again geodesics! This is a totally fine approach to take, and perhaps a convenient one if you are very good at not missing minus signs. However when it comes to our use case for Minkowski geometry (the physics of relativity) we will see that the length of negative curves really corresponds to time intervals, and if we put a negative here, we'll have to negate it once more to think of intervals of time as *positive* like we do in daily life. So, we will opt not to do this, and instead just deal with the fact that geodesics are *maximizing*.

This turns out to be alright actually - as its strangeness actually forces us to think carefully about what is going on, and this careful though reveals things are even stranger for positively separated points!

In $\mathbb{R}^{1,1}$ where the inner product has one positive and one negative direction, things are symmetric, and so nothing stranger happens (positive geodesics are also length maximizing). But, as soon as there is more than one positive direction, things can get rather strange indeed.

Exercise 30.5. Let $p = (0,0,0)$ and $q = (2,0,0)$ be two positively separated points in $\mathbb{R}^{2,1}$, and let $\gamma(t) = (t,0,0)$ be the affine line connecting them.

- Show that there are nearby curves to γ which are *shorter*, by varying the curve slightly into the negative z direction (for example, look at the curve connecting $(0,0,0)$ to $(1,0,z)$ and then continuing to $(2,0,0)$ for $z \in (0,1)$.)

- Show that there are nearby curves to γ which are *longer*

To study this a bit further, it will be useful to have a little more terminology available to us, so we introduce the idea of a *variation*:

Definition 30.6. If γ is an affine line between two positively separated points, a *variation of* γ is a nearby curve $\gamma + \eta$, where η is some curve with $\eta(a) = \eta(b) = 0$ (so that γ and its variation start and end at the same point).

We call $\gamma+\eta$ a *negative variation* if $\eta(t) = (0, 0, z(t))$ is nonconstant only in the direction where the inner product returns negative values, and analogously we call $\gamma + \eta$ a *positive variation* if $\eta(t) = (x(t), y(t), 0)$ is non-constant only in the directions where $\star$ is positive.

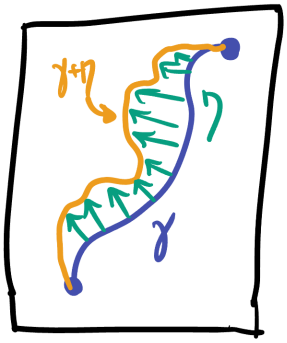


Figure 30.9.: A variation of γ.

Exercise 30.6. Let p, q be two positively separated points in Minkowski space, and γ the affine line connecting them.

- Show that γ is a local *minimum* of the length functional over the set of all positive variations of γ.
- Show that γ is a local *maximum* of the length functional over the set of all negative variations of γ.

This shows that γ is a local minimum of length if you vary the curve in the positive directions of the inner product, but is a local maximum if you instead vary the curve a bit in the negative direction. From multivariable calculus we might recognize points that are either a maximum or minimum depending on the direction you slice in as *saddle points*, and indeed this is the case here.

Theorem 30.5. *Let p, q be a pair of positively separated points in Minkowski space, and γ the affine line connecting them. Then γ is a* saddle point *of the length functional over all regular curves.*

We do not prove this theorem, as understanding the precise definition of a saddle point in infinite dimensions, and ensuring to do the calculation correctly would take us rather far afield. And, seeing that we will never use this result (in our upcoming physics application, it will turn out that *only* the negative curves are relevant) its just not worth it here.

However, it does tell us that if we want a general definition of geodesic in Minkowski space, we need to work a little harder: we can't just replace local minimum with local maximum and call it a day; instead we should seek a definition which captures both max/mins and saddles.

Definition 30.7 (Minkowski Geodesics)**.** A minkowski geodesic is a regular curve $\gamma : \mathbb{R} \to \mathbb{R}^{3,1}$ which is a *critical point* of the Length functional

$$\text{Length}(\gamma) = \int_I \|\gamma'\| \, dt = \int_I \sqrt{|\gamma' \star \gamma'|} \, dt$$

In fact, one could take this more general definition and apply it back in Riemannian geometry as well. Even though it appears to allow for more possible behaviors, one can prove that with a positive inner product at each tangent space, the only critical points of length are *actually minima* - that is, they are the same geodesics we have already found! So, there is no harm in replacing the definition with this, and using it universally across both Riemannian and Minkowski spaces.

This small change turns out to be the a hint to much wider generalizations, bringing geometry deep into the study of quite a lot of fields of math and physics. We do not have the time nor background to develop such here, and should remain focused on our goal. But I cannot help but mention one: Lagrange managed to rewrite the laws of classical mechanics as an *optimization problem*, where the solutions to Newton's laws appeared to correspond to the minima of a certain function (called the *Action Functional*). But on closer inspection - this didn't always work! Instead we discovered physics is not seeking *minima* but rather *critical points* and so this generalized notion of geodesic is the correct notion here as well. (Look up *The Principle of Stationary Action* to learn more)

31. Geometry of Spacetime

We begin by defining spacetime to be the set $\mathbb{R}^4 = \{(x,t) \mid x \in \mathbb{R}^3, t \in \mathbb{R}\}$ of all possible points in space, at all possible moments of time. We call a point (x,t) an *event*, and we call a point $x \in \mathbb{R}^3$ a *location*, and a point $t \in \mathbb{R}$ a *moment*. Our goal is to try to understand the mathematical structures that determine the behavior of spacetime - a lofty goal with very little to go off of initially!

31.1. Axiomatization

Like many things in mathematics, one way to study spacetime is to study its group of symmetries. This is analogous to how we study Euclidean space by discovering its group of isometries, and then use our newfound knowledge of translations rotations and reflections to simplify all further calculations.

But how do we go about finding the symmetries to start with? In Euclidean geometry, perhaps we start by realizing that the Euclidean plane has a particular mathematical structure on it - it has a *distance function* (induced by a dot product on all tangent spaces), and we can rigorously define the isometry group as the set of all transformations that preserve this dot product. But for spacetime, the situation seems much more difficult - we don't know what sort of math structure best describes spacetime, that's one of the things we are looking to discover! So we can't just explicitly define spacetime symmetries to be the things that preserve this (unknown!) structure.

However, if we dig back deep enough into history, we can find a close analogy between our current predicament and geometry. The greeks after all did not know about infinitesimal tangent spaces and all that: instead, they described geometry *axiomatically*, by specifying properties that they observed to be true of space, and then using these as the foundations of their mathematical theory. Us moderns then could take the axioms and try to rigorously find which mathematical spaces satisfy them: we saw in Geometry class that if you take Axioms 1-4 there are two possible ways the world could be (Euclidean or Hyperbolic), but once you have all the Axioms 1-5, there is a unique structure (Euclidean dot product on every tangent space) which instantiates them.

Could we attempt an axiomatic description of spacetime? That is, could we list some rules we claim to be true (based on our observations of the world around us) and use these as *constraints* on our mathematical theory? Perhaps, if we choose a good set of

axioms, we will be able to find a small number of possible solutions (like Euclidean / Hyperbolic space, for Euclid's Axioms 1-4). Then, if we were left without a unique solution, we could try to find *more axioms* to add (based on other observations of the world around us) which further narrow down to a unique mathematical structure, which would allow us to begin a rigorous study of spacetime.

Here's a proposal for three first axioms: in the first we take the ideas of Euclid, in the second the ideas of Galileo, and while the third observation doesn't have a name, it certainly predates the others, going back to the first humans to imagine themselves as living *in* space while time *passes*.

Definition 31.1 (Axioms of Spacetime Symmetries)**.**

- **Symmetries of space and time are symmetries of spaceitme:** If $A : \mathbb{R}^3 \to \mathbb{R}^3$ is a Euclidean isometry, then $\mathscr{A}(p,t) = (A(p),t)$ is a symmetry of spacetime. Similarly, if B is any isometry of $\mathbb{R}$ (a translation, reflection or combination) then $\mathscr{B}(p,t) := (p, B(t))$ is a symmetry of spacetime.
- **Galileo's Principle:** All symmetries of spacetime must preserve the class of constant speed trajectories. And, if ℓ_1 and ℓ_2 are the worldlines of any two constant speed observers, there is a symmetry of spacetime taking ℓ_1 to ℓ_2.
- **Space is different than time:** There is no symmetry of spacetime which takes the vertical axis $(0,t)$ to a line in space $(\ell, 0)$.

The symmetries of spacetime form some group G of transformations $\mathbb{R}^4 \to \mathbb{R}^4$. But which group? These axioms put strong constraints on this group of transformations, and so our immediate goal is to try and discover which groups G satisfy these axioms. Perhaps surprisingly, the list is small! There are only two different groups, up to isomorphism, whose transformations satisfy Axioms (1)-(3). Said colloquially, these axioms alone imply there is only two possible ways that spacetime could logically behave.

31.1.1. Step 1: Interesting Symmetries Change Velocity

By Axiom (1) we know that spacetime is homogeneous, as Euclidean space is homogeneous, and the real line is homogeneous. We can use this to reduce the class of isometries we are interested in. Let $\phi : \mathbb{R}^4 \to \mathbb{R}^4$ be an arbitrary symmetry of spacetime, and let $\phi(0,0) = (x_0, t_0)$. Then we can find a Euclidean isometry $A : \mathbb{R}^3 \to \mathbb{R}^3$ with $A(x) = 0$, and the time isometry $B(t) = t - t_0$ for which $B(t_0) = 0$, and build (via axiom (1)) a spacetime symmetry $\psi(x,t) = (A(x), B(t))$. Then notice that

$$\psi\phi(0,0) = \psi(x,t) = (0,0)$$

So $\eta = \psi\phi$ is a symmetry of spacetime (both ϕ and ψ were in G, so $\eta = \psi\phi \in G$) where $\eta(0,0) = (0,0)$. But now, taking an inverse, we see

$$\phi = \psi^{-1}\eta$$

So an arbitrary symmetry of spacetime is the composition of a symmetry which fixes the origin, and a symmetry which is just a translation in space and time. Since we understand the symmetries of space and time separately (from Euclidean geometry!) this implies that if we can understand the spacetime symmetries that fix the origin, we can understand the entire group G.

So, what are the symmetries fixing the origin? Some of these we also already understand: if we take a Euclidean rotation R that fixes $0 \in \mathbb{R}^3$ for instance, the symmetry $\mathscr{R}(x,t) = (R(x),t)$ fixes the origin. Thus, what we are *really* interested in are spacetime symmetries that fix the origin of *spacetime* but do not fix the origin of space for all time (as this is a Euclidean isometry, just applied at each time!).

This means we are looking for isometries that fix the origin but do not fix the t axis! These are the only sort that we do not already understand. But by Galileo's principle (axiom 2) the t axis is a constant speed trajectory (moving at speed zero) and so any symmetry of spacetime must send it to another constant speed trajectory, which is another affine line. But, since without loss of generality our isometry fixes the origin, this is just some other line through $(0,0)$! All such lines are of the form (vt,t) for $v \in \mathbb{R}^3$, or of the form $(vt,0)$ if they lie directly in the Euclidean space. However Axiom (3) rules out this second class (this would be a symmetry that took the time direction to a space direction) so, we know that $(0,t)$ must be sent to (vt,t). Such a trajectory represents something that moves v units of space for every t units of time, and thus represents someone moving at velocity v.

Thus, the only isometries we are interested in are the ones that take $(0,t)$ to (vt,t): the rest we already understand! Let's call such a symmetry $S(v)$.

One thing to ask ourselves here; $S(v)$ unique (if we are trying to give it a name, after all)? What if we had two symmetries A and B which both took $(0,t)$ to (vt,t)? Then B^{-1} would take (vt,t) to $(0,t)$, and so the combination $B^{-1}A$ fixes the t axis pointwise - it restricts to a Euclidean isometry only in space! That is, we can write $B^{-1}A = \mathscr{R}$ where $\mathscr{R}(x,t) = (R(x),t)$ is just a Euclidean transformation. Composing with B yields $A = B\mathscr{R}$, so A and B differ by a Euclidean isometry. BUT we already understand Euclidean isometries! So A and B are 'essentially' the same. (Precise note: really what we've seen here is that if A and B both take $(0,t)$ to (vt,t) then they lie in the *SAME COSET* of the Euclidean group inside the group of spacetime symmetries!)

%Note to future self: prove that things fixing the t axis pointwise must be the SAME euclidean isometry on each slice

31.1.2. Step 2: The Group is Linear

The first step is to show that the only groups that satisfy these axioms are groups of linear transformations, that is, $G < \mathrm{GL}(4, \mathbb{R})$. We did not do a full and rigorous job of this in the independent study, (let me know if we should return, and dot all our i's and cross our t's): instead we pointed to two ways one can prove this with more work

- Using Galileo's axiom, we see that if $\phi : \mathbb{R}^4 \to \mathbb{R}^4$ is a symmetry it must take constant speed trajectories to constant speed trajectories. Since constant speed trajectories are affine lines, this means it must preserve (at least some subset of) affine lines. ϕ also fixes the origin, so it sends affine lines through the origin to affine lines through the origin: like a linear map! (It's more work, but in fact this is the only option, it is a linear map)
- We could instead proceed and just look for the *subgroup* of G that is linear - and accept the fact that there could also be nonlinear symmetries out there! We will end up finding a bunch of linear symmetries, and in the end, can go back and argue that we have actually found *all the symmetries*: there were no nonlinear ones out there to be worried about!

31.1.3. Step 3: We can Work with 1 Space Dimension

Let $S(v)$ be a symmetry of spacetime taking $(0, t)$ to $(\vec{v}t, t)$, and let $v = \|\vec{v}\|$ be the speed. Then, just in $\mathbb{R}^3$ there is a Euclidean isometry rotating about 0 which sends $\vec{v}$ to $(0, 0, v)$. For a proof recall that Euclidean space is isotropic, so we can take any direction on the unit sphere to any other direction on the unit sphere. Thus there's an isometry taking $\vec{v}/\|\vec{v}\|$ to $(0, 0, 1)$. But, isometries preserve length, so this same isometry must send the vector $\vec{v}$ to a vector of the same length in the direction of $(0, 0, 1)$: that is, by definition $(0, 0, v)$. If we call this Euclidean rotation R, we can build an associated spacetime symmetry $\mathscr{R}(x, t) = (R(x), t)$.

Using this, consider the $X = \mathscr{R}^{-1}S(v)$: this takes the $(0, t)$ axis to $((0, 0, s)t, t)$. Then $\mathscr{R}X = S(v)$, so we can understand our arbitrary isometry $S(v)$ as a composition of X, which takes $(0, t)$ to a velocity along the z axis, and a Euclidean rotation. Since we already understand Euclidean isometries, this lets us *further simplify* what we are interested in: it is enough if we understand the isometries which take a stationary observer to one moving along the z axis!

Using the fact that we *also* know that such resulting transformations are linear, we can write this as a matrix:

$$S = \begin{pmatrix} \star & \star & \star & \star \\ \star & \star & \star & \star \\ \star & \star & \star & \star \\ \star & \star & \star & \star \end{pmatrix}$$

Our goal is to fill in the missing entries! (right now, all of them!). One first thing to notice, is that with our motion in the z direction, the x and y directions don't get mixed up with z and t after the transformation (an argument from axiom 2 goes as follows: if distances in the x and y directions were affected by me moving in the z direction, I could tell whether or not I was moving - violating Galileo's principle! - by seeing how the length of something I was carrying with me changed! I would like a better argument here though)

This implies the matrix is block diagonal: the xy do not mix with the zt, so we have

$$S = \begin{pmatrix} \star & \star & 0 & 0 \\ \star & \star & 0 & 0 \\ 0 & 0 & \star & \star \\ 0 & 0 & \star & \star \end{pmatrix}$$

Now, we have a transformation which translates along the z-axis, and does something in the x, y direction. Say that it does the Euclidean rotation R in the xy plane, which acts on spacetime as $R(x, y, z, t) = (R(x, y), z, t)$. Because we are free to use Euclidean isometries to simplify our situation, we can compose our map above with R^{-1} without changing the property that we care about (that it translates along the z axis) so we can without loss of generality assume that the 2×2 euclidean block is the identity! Thus, our matrix is

$$S = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \star & \star \\ 0 & 0 & \star & \star \end{pmatrix}$$

This leaves a very manageable sized problem - everything about the symmetries of spacetime can be totally understood so long as we know how they work with one space and one time dimension!

31.1.4. Step 4: Some Matrix Calculations

This is where all the real work is!!

Now we have gotten ourselves into a sufficiently restricted situation that we can do some actual calculations. We are only interested in symmetries fixing the origin, and taking $(0, t)$ to (vt, t) for some $0 \neq v \in \mathbb{R}^3$. We know these symmetries are linear, and after a Euclidean isometry can actually assume without loss of generality that v is parallel to the z axis, so the matrix representing our symmetry is block diagonal with only one 2×2 undetermined block:

$$S(v) = \begin{pmatrix} a(v) & b(v) \\ c(v) & d(v) \end{pmatrix}$$

We then did a bunch of matrix calculations that I do not feel like typing tonight (sorry!) but I emailed out a handwritten copy of. Together, this implies that there are two possibilities for the group of symmetries of $\mathbb{R}^4$ satisfying axioms (1) to (3).

Possibility 1: The group of symmetries of spacetime consists of the following matrices

$$G = \left\{ \begin{pmatrix} 1 & -v \\ 0 & 1 \end{pmatrix} \middle| v \in \mathbb{R} \right\}$$

Possibility 2: The group of symmetries of spacetime consists of the following matrices, for c some positive constant.

$$L = \left\{ \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \begin{pmatrix} 1 & -v \\ -v/c^2 & 1 \end{pmatrix} \middle| v \in (-c, c) \right\}$$

31.2. Implications

Our next goal, as mathematicians is to try and study these two possible worlds, and derive some properties they have. We will refer to Possibility I as the Galilean world, and Possibility II as the Lorentzian world.

Proposition 31.1 (All Lorentzian Worlds are Isomorphic)**.** *At first, it appears that there are really* uncountably many *different possibilities for spacetime: one possible Galilean world, but a continuum of Lorentzian worlds, one for each value of $c \in (0, \infty)$. But, it turns out that this entire continuum of Lorentzian worlds are qualitatively the same: we can make this formal by saying that the group of symmetries for any two lorentzian worlds are* isomorphic

Exercise 31.1 (Prove This:)**.** When $c = 1$ we have the Lorentz $\mathscr{L}$ group with matrices

$$\frac{1}{\sqrt{1 - v^2}} \begin{pmatrix} 1 & -v \\ -v & 1 \end{pmatrix}$$

For the Lorentz group $\mathscr{L}_c$ with an arbitrary c, show that the map

$$\mathscr{L}_c \to \mathscr{L}_1$$

defined by sending v to v/c is

- Injective
- Surjective
- A group homomorphism

31.2.1. Velocity Addition

These two worlds have very different rules for velocity addition: in the Galilean world, velocities add:

Proposition 31.2. *For any $v, w \in \mathbb{R}$,*

$$G(v)G(w) = G(v + w)$$

But in the Lorentzian world, velocities satisfy a rather different formula, which makes sure that the overall velocity always remains within $(-c, c)$ as the formula seems to require.

Proposition 31.3. *For any $v, w \in (-c, c)$,*

$$L(v)L(w) = L\left(\frac{v + w}{1 + \frac{vw}{c^2}}\right)$$

Exercise 31.2. Do these calculations.

Exercise 31.3. Then, using the velocity addition for $\mathscr{L}$, show that no matter what speed you start out moving at, and no matter how much you speed up by, you will never go faster than c. That is, our mystery constant is actually a *universal speed limit*.

31.2.2. Existence of a Constant Speed

Show that in the Galilean world, if there is an object moving at any speed v you can catch up to it: there's a symmetry of spacetime that boosts your velocity to v, and now the object is standing still. In the Lorentzian world, we already know that it is impossible to boost an *initially stationary* observer to any speed beyond $(-c, c)$. But what if we had an object moving at the universal speed c to begin with? This would follow a trajectory $(x, t) = (ct, t)$ through spacetime.

Exercise 31.4. Say you observe an object to be moving at speed c.

What would happen if you speed up? Would you see its speed change at all? Show that no matter what speed you go (so, no matter which Lorentz transformation $L(v)$ you apply to (ct, t)) you will always see this object move at the same speed.

This is an incredibly weird prediction: first, we saw that it is impossible to start out stationary and move at any speed outside of $(-c, c)$. So, you may have thought that if there were an object moving at speed c, even though you could never catch up to it,

31.2.3. Metric

The type of mathematical structure spacetime has in these two possibilities is also rather different. Perhaps surprisingly, it turns out that the Lorentzian spacetime (with the more complicated looking matrices!) actually has a nicer mathematical structure.

What should we be looking for here? Well, if Geometry taught us anything, the geometric properties of a space are usually stored in the form of a dot product on infinitesimal tangent vectors (we saw this even for strange geometries, like Minkowski space). So, now that we have two potential symmetry groups of spacetime, we should ask if these correspond to any sort of geometric structure.

Exercise 31.5 (The Galilean World)**.** Show that if $(x, y) \star (u, v) = \alpha xu + \beta yv$ is any inner product on the space $\mathbb{R}^2 = (z, t)$ of $1 + 1$ dimensional Galilean spacetime which is preserved by all transformations $G(v)$ then either $\alpha = 0$ of $\beta = 0$. That is, its not really an inner product on spacetime at all but rather only notices information about space, or information about time. This tells us that in the Galilean world, there is no geometry of spacetime, but rather space and time are just separate entities.

Exercise 31.6 (The Lorentzian World)**.** Show that the inner product $(x, y) \star (u, v) = xu - c^2yv$ is preserved by all lorentzian transformations $L(v)$. This is a Minkowski dot product (just scaled by the quantity c^2): and says that this spacetime cannot be thought of as space and time separately any more than the x and y axes in the plane can be thought of separately - they are just parts of one larger geometric object (this time with the geometry of Minkwoski space!)

This was Hermann Minkowski's big realization upon reading Einstein's work: in 1908 he said in a lecture, announcing this that

> "Gentlemen! The views of space and time which I wish to lay before you ... They are radical. Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality."

Remark 31.1 (Reminder: Showing a dot product is preserved). If L is a transformation and v, w are tangent vectors based at p, to show that L preserves the dot product $\star$ one must show that $DL_p(v) \star DL_p(w) = v \star w$. For us, since L is linear we know $DL_p = L$ and so we just need to compute L applied to vectors v and w and then take the dot product. Alternatively, remember we can just show infinitesimal length is preserved, so $\|Lw\|^2 = Lw \star Lw = w \star w$ for a single arbitrary vector w.

This proves our main theorem: we've discoverd the mathematical model for spacetime in this case, and identified it with something we already understand!

Theorem 31.1. *The geometry of spacetime equipped with the Lorentz symmetries is isomorphic to Minkowski space.*

32. Relativity

32.1. Symmetries of Spacetime

We've shown there are two possible ways the symmetries of spacetime could behave: it can either follow the laws of the Galilean Group G:

$$G = \left\{ \begin{pmatrix} 1 & -v \\ 0 & 1 \end{pmatrix} \middle| v \in \mathbb{R} \right\}$$

Or, it can follow the symmetries of the *Lorentz Group* L

$$L = \left\{ \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \begin{pmatrix} 1 & -v \\ -v/c^2 & 1 \end{pmatrix} \middle| v \in (-c, c) \right\}$$

Where these are applied as linear transformations to the spacetime of events $e = (x, t)$ for $x \in \mathbb{R}, t \in \mathbb{R}$. While the Galilean group has much more intuitive consequences, it turns out the real world exhibits Lorentz symmetry, and we focus on the strange consequences of that here.

32.1.1. Observers in Spacetime

Objects in the world are modeled by *curves in spacetime* $\gamma(s) = (x(s), t(s))$. An object is said to be *stopped* at some instant s_0 if its x coordinate is not changing: $x'(s_0) = 0$. An object is stopped for a interval if this holds for all points in that interval: equivalently, if x is constant for $s \in [a, b]$. An object described by γ is *stoppable* at s_0 if there is some Lorentz transformation $L(v)$ where the curve $L(v)\gamma(s)$ is stopped at s_0: equivalently if $L(v)\gamma'(s_0)$ is parallel to $\langle 0, 1 \rangle$.

In the coordinates (x, t) of an observer O who is stationary at $x = 0$, we define the *speed of an object* γ *relative to* O to be the change in x over the change in time. That is, the average speed over an interval $s, s + h$ is

$$\frac{x(s+h) - x(s)}{t(s+h) - t(s)}$$

Exercise 32.1. Show the *instantaneous speed* is expressible as a derivative

$$v = \frac{x'(s)}{t'(s)}$$

An observer is called *constant speed* if this v is constant as a function of s. Such constant speed trajectories describe affine lines in spacetime.

Galileo's principle says that anyone's viewpoint on the world is equivalent - that is, that anyone can imagine themselves as *not moving*. This gives a constraint on the set of curves that describe trajectories we can move on, which we will call *observer trajectories*. Every observer trajectory must be locally stoppable at each point, in order to obey Galileo's principle.

In a world with Lorentz symmetry, this already poses a strong constraint: because $L(v)$ is defined only for $v \in (-c, c)$ and $L(v)$ takes the vector $\langle -v, 1\rangle$ to $\langle 0, 1\rangle$, we see that ONLY trajectories with speed less than c are stoppable, and so observers can only travel at speeds less than the constant c!

While trajectories moving at speed c are ruled out for observers, they are not immediately disallowed overall. However such trajectories have some very strange properties:

- An object moving at speed c can never be stopped: it must always move!
- You cannot run away from an object moving at speed c: no matter your speed, you will always see it approaching you at speed c!

Thus not only must such objects never stop moving, but they can never even slow down! They must *always* travel at speed c.

32.2. Measuring Time along a Trajectory

If $\gamma(s)$ is the trajectory of some observer through spacetime, one may ask between the points $\gamma(a)$ and $\gamma(b)$, how much time elapses for that observer. We need to find a way to calculate this quantity. Here's an alternative argument to the one given in class (this is perhaps more like the 'physics way' starting with a simple case and then building up)

If the observer is stationary, this is easy: the vertical axis is *defined* as the time axis of a stationary observer, so we just measure the difference in t coordinates. But for any other observer this is not possible. However, this suggests a strategy to calculate the time duration of a constant speed observer: if an observer moves at speed v can use the Lorentz boost $L(v)$ to make them stationary, at which point we can read off from the time axis exactly how long of a time they experienced. For example, using

homogenity of spacetime to have our constant speed observer pass through the origin, we may write

$$\gamma(s) = (vs, s)$$

and then

$$L(v)\gamma(s) = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\begin{pmatrix} 1 & -v \\ -v/c^2 & 1 \end{pmatrix}\begin{pmatrix} sv \\ s \end{pmatrix}$$

$$\frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\begin{pmatrix} sv - sv \\ -sv^2/c^2 + s \end{pmatrix} = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\begin{pmatrix} 0 \\ s\left(1-\frac{v^2}{c^2}\right) \end{pmatrix}$$

So, between $s = a$ and $s = b$ this observer *themselves* really experiences the time

$$T = (b-a)\sqrt{1-\frac{v^2}{c^2}}$$

On a general curve γ we need to instead take an infinitesimal approach, as at each point the observer may be moving at a different speed $v = x'/t'$. For an infinitesimal interval around this point, the time dT which passes is proportional to $ds = b - a$ by the above formula, and

$$\Delta T = \Delta s\sqrt{1-\frac{v^2}{c^2}}$$

Taking the limit as the interval shrinks to zero and integrating gives the following result for curves of the form $\gamma(s) = (x(s), s)$ (this constraint is only because we used the curve (sv, s) as our initial starting point!)

$$T = \int \sqrt{1-\frac{v^2}{c^2}}ds = \int \sqrt{1-\frac{x'(s)^2/t'(s)^2}{c^2}}ds$$

32.2.1. Another Way

We could also carry out this calculation more abstractly (this is the 'math way' that we did in the independent study): we know that spacetime with Lorentz symmetries preserves the dot $(x')^2 - c^2(t')^2$ on tangent vectors, from our previous assignments. Looking at this dot product, the units we have written it in are *space-units* (if x is in meters and c is in meters per second, then ct is also in meters...). But, if we wanted to measure everything in time-units (perhaps more intuitive for us, who are trying to understand durations) we could do that just as easily. Dividing this entire dot product through by the constant c^2 gives

$$\frac{(x')^2}{c^2} - (t')^2$$

Which is also preserved (as its just a constant multiple of the original) but now has the same units that t does (so, if we are thinking of time in seconds, the output of the norm will be seconds). If γ is the trajectory of some observer, then $\gamma' = (x', t')$ is moving at speed $v = x'/t'$ less than c, and so there is a Lorentz transformation $L(v)$ bringing it to a stop. This transformation must take its tangent vector to a vector along the vertical t axis, but also cannot change the length with respect to this preserved inner product! Thus, the duration we would measure for time must simply be the norm

$$\|\gamma'\| = \sqrt{\left|\frac{(x')^2}{c^2} - (t')^2\right|} = \sqrt{(t')^2 - \frac{(x')^2}{c^2}} =$$

$$= t'\sqrt{1 - \frac{(x')^2/(t')^2}{c^2}} = t'\sqrt{1 - \frac{v^2}{c^2}}$$

Integrating this over a curve gives

$$\text{Time}(\gamma) = \int_a^b t'(s)\sqrt{1 - \frac{v(s)}{c^2}}ds$$

And, for curves of the form $\gamma(s) = (x(s), s)$, we see $t' = 1$ and this agrees with our previous formula from the physics derivation. A side note: we can *always* parameterize a curve like this if we want, so there is no loss of generality here! Because our curve is a negative curve, we know that t' is always positive, and so t is an increasing function. Thus, t has an inverse function and we can use that inverse to reparameterize: the equations $(x(s), t(s))$ and $(x(t^{-1}(s)), t(t^{-1}(s)))$ trace out the same curve in spacetime, but $t(t^{-1}(s)) = s$ so the second is just of the form $(f(s), s)$ for some f.

Now we can proceed from this general statement applicable to all curves, to specialize for constant speed trajectories. Say that γ is a constant speed curve connecting the spacetime events $p = (x_p, t_p)$ and $q = (x_q, t_q)$. Then one way to parameterize this affine line is

$$\gamma(s) = p + s(q - p)$$

where $\gamma(0) = p$ and $\gamma(1) = q$. The derivative of this is just $(x', t') = q - p = (x_q - x_p, t_q - t_p)$, and so

$$\text{Time}(\gamma) = \int_0^1 \|\gamma'\|ds = \int_0^1 \sqrt{(t')^2 - \frac{(x')^2}{c^2}}ds = \int_0^1 \sqrt{(t_q - t_p)^2 - \frac{(x_q - x_p)^2}{c^2}}ds$$

$$= \sqrt{(t_q - t_p)^2 - \frac{(x_q - x_p)^2}{c^2}} = \sqrt{\Delta t^2 - \frac{\Delta x^2}{c^2}}$$

This is a remarkably understandable formula! (Even if the consequences are wild) In a diagram, we are measuring the duration experienced along a trajectory simply by the *Minkowski Pythagorean Theorem!* Take the difference in time coordinates Δt, the difference in space coordinates $\Delta x/c$ (dividing by c puts it in units of time) and then apply $a^2 - b^2$!

Side Note: we can rewrite this in terms of the speed $v = \Delta x/\Delta t$ so it matches up with the physics derivation, factoring out Δt from the root:

$$\text{Time}(\gamma) = \Delta t\sqrt{1 - \frac{(\Delta x/\Delta t)^2}{c^2}} = \Delta t\sqrt{1 - \frac{v^2}{c^2}}$$

32.3. The Flash vs a Flashlight

Imagine the following setup: you and The Flash are standing still right next to one another, at a distance of d from your friend who is also standing still relative to you both, and is holding a flashlight. At some point in time, your friend turns on the flashlight, and immediately The Flash takes off, running away from the light beam at speed v.

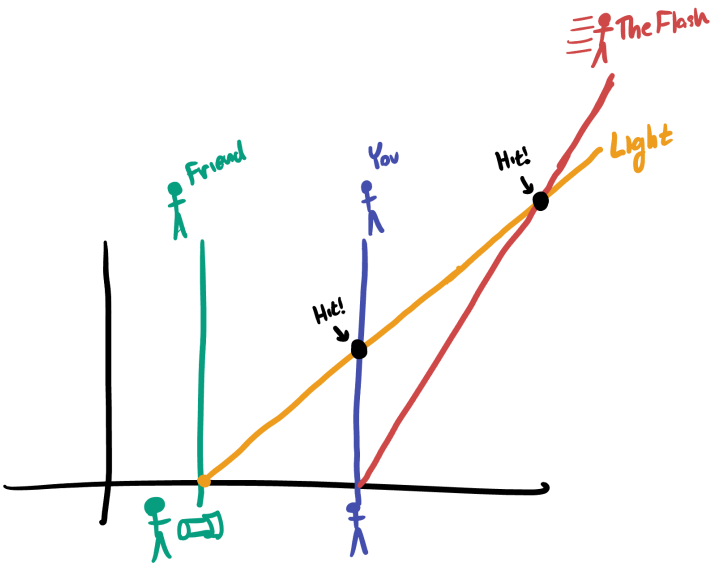


Figure 32.1.: Flash running away from a flashlight.

Exercise 32.2.

- How much time passed on your clock, by the time the light hits you?
- How much time passes on The Flash's clock, by the time the light hits him?
- What time will your clock read at the moment that The Flash is hit?
- What time will your clock read when you see The Flash get hit?

32.4. The Twin Paradox

Two twins start out right next to one another. One of decides to stay where they are, and just hang out. The other gets into an fast moving ship and goes on a joyride, following some trajectory $\gamma(s)$ through spacetime, but eventually coming back to where they started.

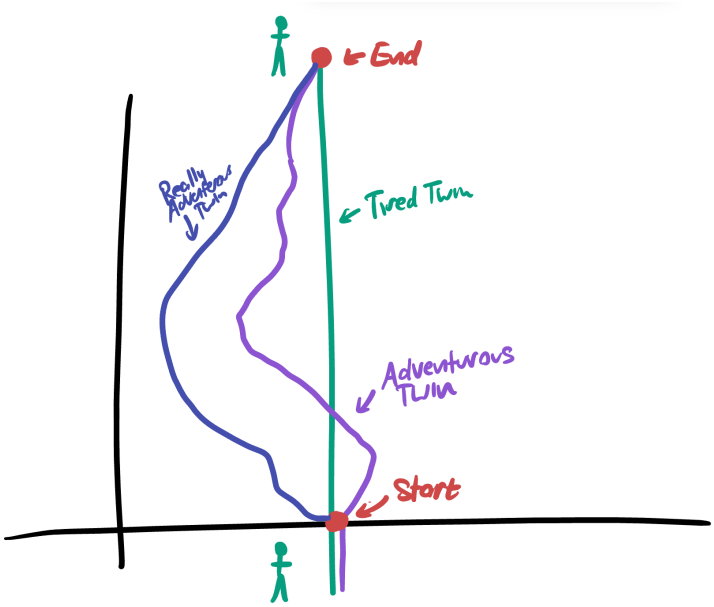


Figure 32.2.: A sedentary twin, and two possible adventures.

Exercise 32.3. Prove that no matter what trajectory γ the adventurous twin travels on, so long as they leave home they return younger than the twin who stayed at home.

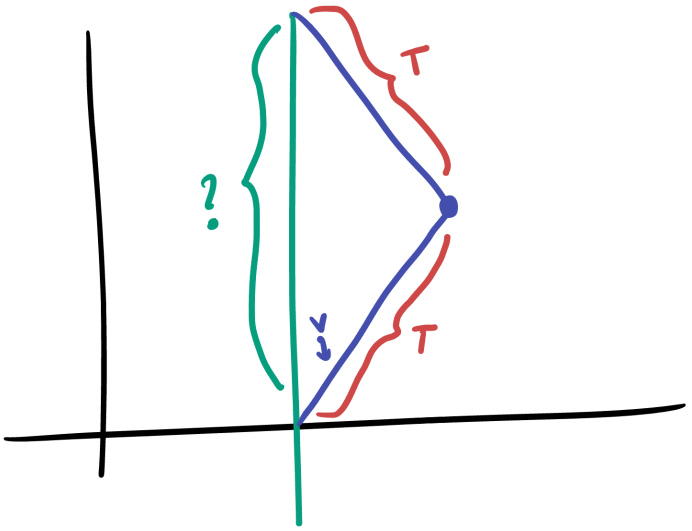


Figure 32.3.: How long is this out-and-back journey for the twin that stayed home?

Exercise 32.4. Say that the adventurous twin leaves home at velocity v, and travels for T units of time (on his clock), and then turns around and heads home by the same manner (traveling at velocity v for time T). How old is his twin when he gets back?

32.5. Neither Before nor After

In the Galilean world, everyone agrees which slices of spacetime counts as "space". These slices define a *universal notion of time*: given any two events $p = (x_p, t_p)$ and $q = (x_q, t_q)$, the time difference $t_p - t_q$ is a number that any Galilean observer could compute from p and q and would agree on. We've seen already that things are much stranger in the Lorentzian world, where the time difference between events (like two twins departing, and then reuniting again) differs wildly depending on who you ask.

But, surely the *order* events occur in is invariant across observers, even if the precise length of time elapsed between them is not - right? Right??

Exercise 32.5 (Wrong)**.** Consider an observer O who is not moving with respect to (x, t), and imagine two events in spacetime p and q that he agrees occur at the same time. Show that there is also an observer who claims that p happens *before* q, and an observer who claims that p happens *after* q.

Hint: If p and q occur at the same time for O then they have the same t coordinate. Apply a Lorentz boost $L(v)$ and compare the resulting t coordinates. What about $L(-v)$?

But not all is lost: there are still pairs of events where everyone agrees the order they occurred in: for instance, it is a fact of the matter that I lived in Minnesota before I lived in California, and not the other way around. Everyone, even fast traveling aliens watching us through super telescopes agrees with this.

:::{#rel-prob-5} Show that if p and q are two events on the trajectory of some observer, and p happened before q for that observer, then everyone agrees that p happened before q. :::

Hint: Show if γ is the trajectory of some observer with $\gamma(a) = p$ and $\gamma(b) = q$ then $\|q-p\|^2$ is negative. And, because spacetime is homogeneous, you might as well consider $p = (0,0)$ is the origin. Then show that if $q = (x,t)$ is a point with $t > 0$ and $\|q - p\| = \|q - 0\|^2 = \|q\|^2$ negative, then every Lorentz transformation applied to q leaves its time coordinate positive, so it always occurs after p, whose time coordinate is zero.

Thus spacetime has events where it is *undefined* which came first, and others where it is *unambiguous* which was first. And, there are no events which unambiguously occur at the same time! (This is what physicists call the "Relativity of Simultaneity"). In fact these two types of pairs of events have a nice math interpretation: the ones which can be ordered in time are the pairs of points with negative separation (from our linear algebra chapter) and those which cannot are those with positive separation.

If you feel like getting really confused, this is a good time to go back and think about The Flash and the Flashlight: remember we said that from our friend with the flashlight's perspective, the following events are simultaneous:

- Him turning on the flashlight
- You and the Flash being right next to one another.

Since you are not moving relative to your friend, you both agree that these events are simultaneous. But The Flash does not! In fact, from the Flash's perspective, the flashlight is not turned on until *much later than he started running*: this means from his perspective, there is much more distance for the light to cover before overtaking him (as he had a head start).

Exercise 32.6. Draw in space slices for The Flash into the diagram, and see that the space slice through the point where the flash runs away from you intersects your friend far before he turned on the flashlight. Use geometry and Lorentz transformations to figure out how long after the flash starts running he thinks the flashlight is turned on.

32.6. You Can't Go Back

The *grandfather paradox* is a classic time travel story that highlights one of the absurd consequences of entertaining time travel into the past. In it, starting from wherever you are right now, you travel along some path in spacetime and eventually you end up back inside the past light cone of your starting point: somewhere where actions you undertake could *affect you* where you started!

The classical example is rather violent: you arrive at a time before your parents were born, and kill one of your grandfathers. Thus, no parents, and no you. But how could this be possible, since *you exist* - after all its *you* who went back and did this! The one can be much less violent and arrive at similar absurdities: what if along your journey you bring your favorite book to read; and upon your arrival in your past, you meet the author before he has written it. You give him a copy, remarking how much you loved it. If he then reads it, agrees, and submits it for publication, who wrote the book?

In fact, this book example provides a much better picture of what must be going on mathematically than the grandfather case. Indeed, track the book: starting in your hand, it follows you (on an observer's trajectory) until it meets the author, at which point it changes hands and follows the author (also an observer's trajectory) until it ends up on a bookstore's shelf, and then gets into your hands! At each point in time the book was following an observer's trajectory (either literally with you or the author, or sitting motionless on a shelf), but at the end of its journey it ended up *at the same spacetime event where it began.*

That means the trajectory of our book is a closed curve: a loop in spacetime!

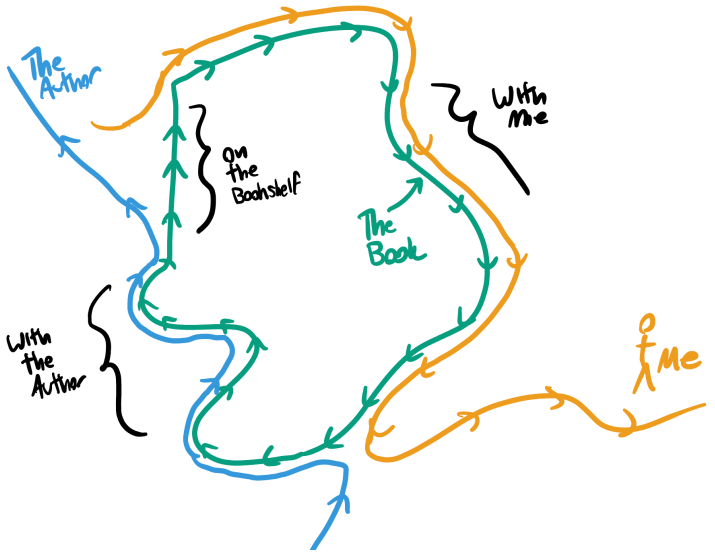


Figure 32.4.: A time travel paradox

Exercise 32.7. Show that there are no closed observer trajectories in spacetime with Lorentzian symmetries. Thus, you can never visit the past!

Hint: An observer trajectory must always have its tangent vector lying inside the negative cone. But if $\gamma(s) = (x(s), t(s))$ *were a closed curve, then use some real analysis to show that the* t *component must have a point where* $t'(s_0) = 0$*. But this is a problem!! Why?*

Assignments

Problem Set 1

Exercise 32.8 (Constructing an Isoceles Triangle)**.** Start with a line segment of length a. Prove that you can construct a triangle with one side of length a, and two sides of length $2a$.

Exercise 32.9 (Inscribing an Equilateral Triangle)**.** Prove that inside of an equilateral triangle, you can inscribe an upside down equilateral triangle of exactly half the side length, shown

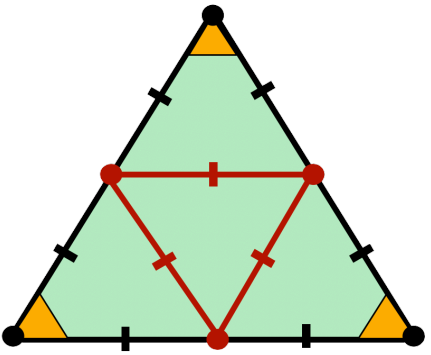


Figure 32.5.: An equilateral triangle inscribed within a larger one.

Exercise 32.10 (Angle Sums of Polygons)**.** A polygon is *convex* if all of its angles are less than 180°, so that it has no "indents". Equivalently, a *convex* polygon is one where any line segment with endpoints on the boundary of the polygon lies *inside* the polygon.

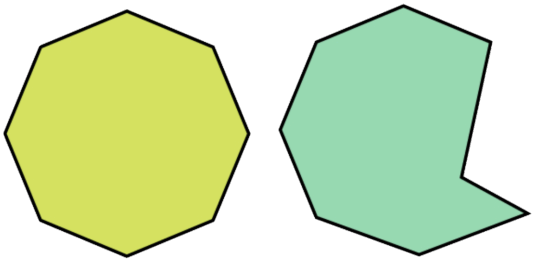


Figure 32.6.: A convex and non-convex octagon.

Prove that the angle sum of convex quadrilaterals is a constant, for all quadrilaterals. Prove the angle sum of convex pentagons is also a constant. What are these constants?

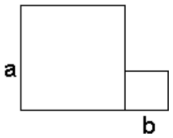
What do you think the formula is for the sum of angles in a convex n-gon? (Optional: If you have seen mathematical induction, prove your guess!)

Exercise 32.11 (Rectangles Exist)**.** Prove that Rectangles exist using Euclid's Postulates (and also Playfair's Axiom, if you like it), and the propositions proven in the sections *Euclid* and *Parallels*.

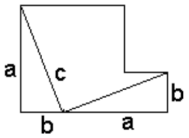
Hint - we know how to make right angles now, and parallel lines through points. Start making some!

Exercise 32.12 (Diagonal Bisectors)**.** If the diagonals of a quadrilateral are bisect one another, then that quadrilateral is a parallelogram.

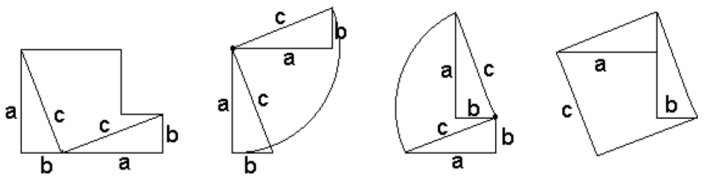
Exercise 32.13 (Proving the Pythagorean Theorem)**.** The following is an ingenious *rearrangement proof* of the Pythagorean theorem.



We start with two squares with sides **a** and **b**, respectively, placed side by side. The total area of the two squares is a^2+b^2.



The construction did not start with a triangle but now we draw two of them, both with sides **a** and **b** and hypotenuse **c**. Note that the segment common to the two squares has been removed. At this point we therefore have two triangles and a strange looking shape.



As a last step, we rotate the triangles 90°, each around its top vertex. The right one is rotated clockwise whereas the left triangle is rotated counterclockwise. Obviously the resulting shape is a square with the side c and area c^2.

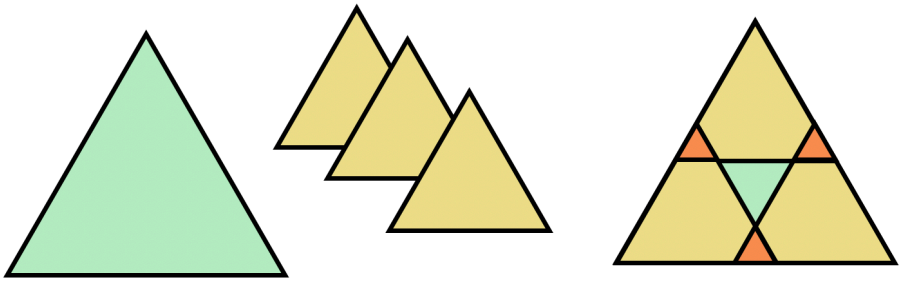
Prove that the final shape shown here is a square, using what we have learned (the Postulates and Propositions).

Problem Set II

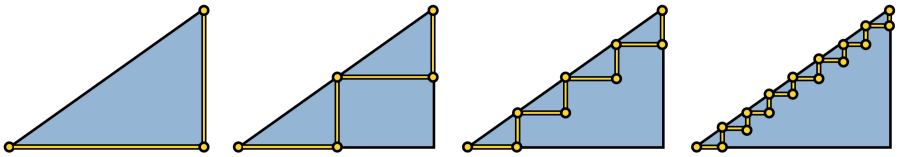
Exercise 32.14 (The Square Root of 3)**.** Read carefully the geometric proof of Theorem 3.2, which proves $\sqrt{2}$ is irrational by showing its impossible to make two integer side-length squares where one has twice the area of the other.

Construct a similar argument showing that it is impossible to find two integer side-length equilateral triangles where one has three times the area of the other.

Hint: try to mimic the argument in the book, but now use the diagram below for inspiration



Exercise 32.15 (Convergence to the Diagonal)**.** Consider a simpler analog of Archimedes' situation, where instead of trying to measure a curve using straight lines, we are trying to measure a straight diagonal line using only horizontal and vertical segments. The following sequence of paths converges pointwise to the diagonal of the square, but what happens to the lengths?

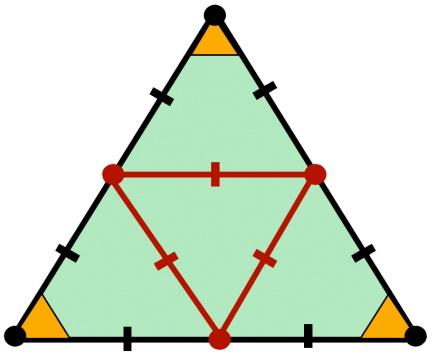


If you believed that because this sequence of curves limits to the diagonal, its sequence of lengths must limit to the length of the diagonal, what would you have conjectured the pythagorean theorem to be?

Exercise 32.16. Use the result of last week's problem Exercise 32.9 (that you can inscribe an equilateral triangle with half the side lengths) to produce an alternative proof of Archimedes sum

$$\sum_{n=0}^{\infty} \left(\frac{1}{4}\right)^n = \frac{4}{3}$$

By dividing up a triangle instead of a square. Draw some nice pictures (its pretty!)



Exercise 32.17. Construct an argument in the same spirit as Archimedes' geometric series to show the following equality:

$$\sum_{n=1}^{\infty} \left(\frac{1}{3}\right)^n = \frac{1}{2}$$

Can you cut something iteratively into thirds? It may not be as pretty as Archimedes', but thats ok!

Fractals

The final two problems involve the Koch Snowflake fractal. In these problems you should still *explain* why things you are doing are valid geometrically, but you **do not need to prove every thing you do from the axioms**. We are getting ourselves ready for a calculus mindset!

This shape is the limit of an infinite process, starting at level 0 with a single equilateral triangle. To go from one level to the next, every line segment of the previous level is divided into thirds, and the middle third replaced with the other two sides of an equilateral triangle built on that side.

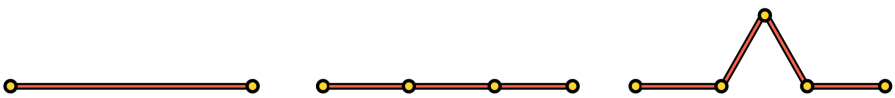


Figure 32.7.: The Koch subdivision rule: replace the middle third of every line segment with the other two sides of an equilateral triangle.

Doing this to *every line segment* quickly turns the triangle into a spiky snowflake like shape, hence the name. Denote by K_n the result of the n^{th} level of this procedure.

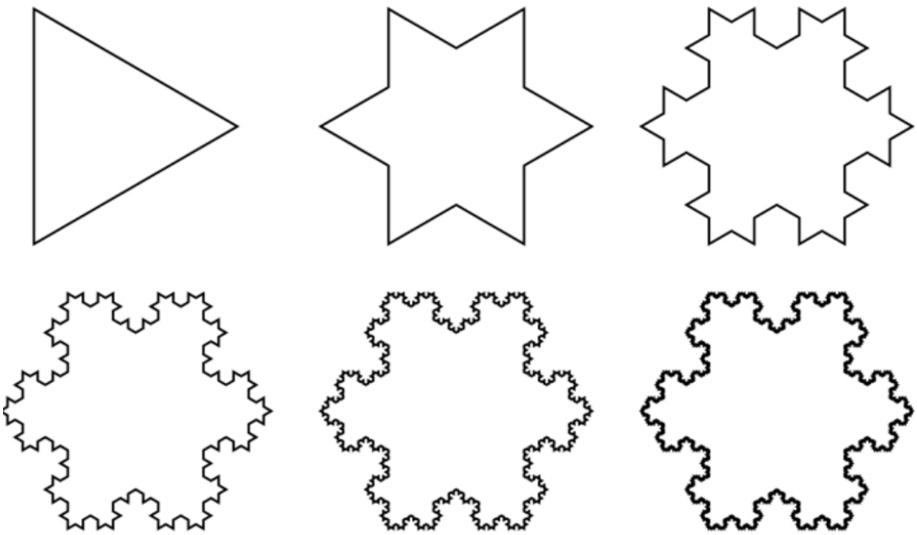


Figure 32.8.: The first six stages K_0, K_1, K_2, K_3, K_4 and K_5 of the Koch snowflake procedure. K_∞ is the fractal itself.

Say the initial triangle at level 0 has perimeter P, and area A. Then we can define the numbers P_n to be the perimeter of the n^{th} level, and A_n to be the area of the n^{th} level..

Exercise 32.18 (The Koch Snowflake Length)**.** What are the perimeters P_1, P_2 and P_3? Conjecture (and prove by induction, if you've had an intro-to-proofs class) a formula for the perimeter P_n.

Explain why as $n \to \infty$ this diverges (using the type of reasoning you would give in a calculus course): thus, the Koch snowflake fractal cannot be assigned a length!

Before doing the next problem: ask yourself what happens to the area of an equilateral triangle when you shrink its sides by a factor of 3? Can you draw a diagram (similar to that from last week's Exercise 32.9 but larger) to see what the ratio of areas must be?

Exercise 32.19 (The Koch Snowflake Area)**.** What are the areas A_1, A_2 and A_3 in terms of the original area A?

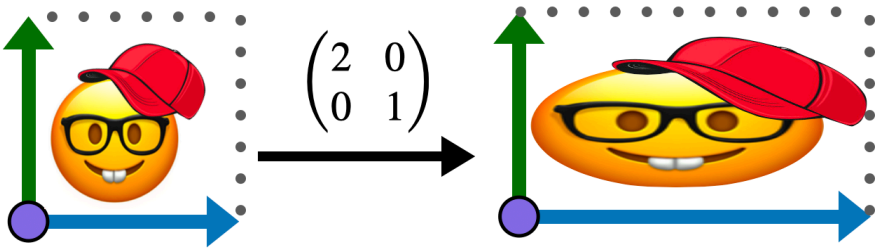
Find an infinite series that represents the area of the n^{th} stage A_n (if you've taken an intro to proofs class or beyond - prove it by induction!). Use calculus reasoning to

sum this series and show that while the Koch snowflake does not have a perimeter, it *does* have a finite area!

Problem Set III

Linear Transformations

This exercise goes with similar examples in the text, on visualizing linear transformations action on the plane by what they do to the points of the unit square. For instance, we saw that the transformation $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ scales the x axis by a factor of 2 and leaves the y axis invariant, so it performs the following *stretch* to our little smiley face



Exercise 32.20. Choose your own image on the plane (hand-drawn is great!), and draw a reference image of it undistorted, inside the unit square. Then draw its image under each of the following linear transformations:

$$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \qquad \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \qquad \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \qquad \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

Determinants & Area

Recall the following definition: the **determinant** of a linear transformation $M = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ is

$$\det M = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

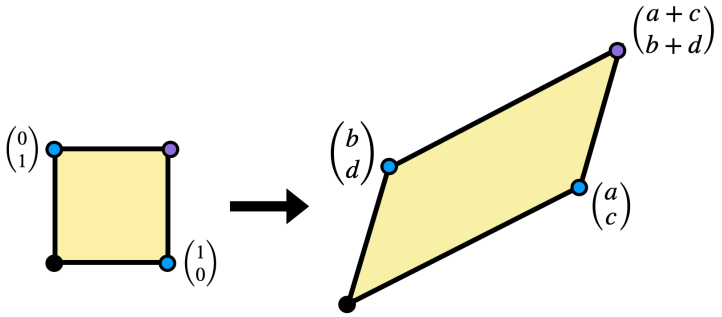


Figure 32.9.: The determinant measures the change in area under a linear map.

In class, said this measured the area change of the unit square under the linear transformation M, but now we will confirm it. We can actually find this area in a pretty satisfying way using just what we've proven about Euclidean geometry so far. We know the areas of squares, rectangles, and right triangles, so let's try to write the area we are after as a difference of things we know:

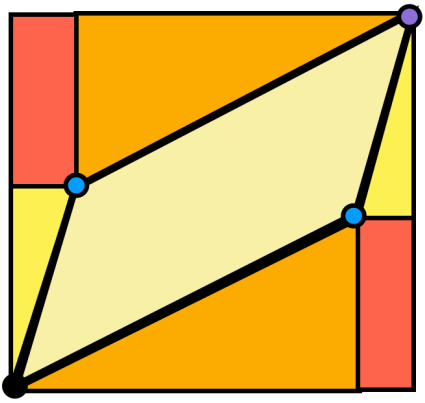


Figure 32.10.: A formula for the determinant can be found knowing only the area of squares, rectangles, and right triangles. (I learned this awesome diagram from Prof Daniel O'Connor!)

Exercise 32.21. Show the area of the parallelogram spanned by $\langle a, c\rangle$ and $\langle b, d\rangle$ is $ad - bc$, using the Euclidean geometry we have done, and the diagram above.

Calculating Derivatives

Practice calculating the derivative of multivariate functions as matrices, and applying them to vectors. No proofs, here, just some computations!

Exercise 32.22. Find the derivatives of the following functions, at the specified points.

- The function $f(x, y) = (xy, x + y)$ at the point $p = (1, 2)$.
- The function $\phi(x, y) = \left(xy^2 - 3x, \frac{x}{y^2+1}\right)$ at the point $q = (3, 0)$.

Now use these to compute the following quantities:

- $Df_{(1,2)}\langle 3, 4\rangle$
- $D\phi_{(3,0)}\langle a, b\rangle$

Differentiating Compositions

This is another problem which focuses on the new *computational skills*, using linear algebra. No proofs here either!

Exercise 32.23. If F, G, H are the following multivariate functions

$$F(x, y) = (x - y, xy)$$

$$G(x, y) = (-y, x)$$

$$H(x, y) = (x^3, y^3)$$

Differentiate the following compositions:

- $F \circ G$ at $(1, 1)$
- $G \circ G$ at $(0, 2)$
- $F \circ G \circ H$ at $(-1, 3)$.

When the Derivative is Constant

In class, we proved that *if* a function is linear, *then* its derivative is constant. But is this the only time a function's derivative is constant? Certainly no - the derivative of $(x, y) \mapsto (x + 1, y)$ is constant (equal to the identity matrix!), even though this function is not linear.

We call a function **affine** if it is the composition of a linear function and addition of a constant. For instance, $2x + 3$ or $5x + 2y - 7$ are affine functions. We call a multivariable function affine if each of its component functions is affine.

Exercise 32.24 (When the derivative is constant)**.** Prove that a function $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ has a constant derivative if and only if the function is *affine*: that is, a linear map plus constants.

Hint: if the derivative is a constant matrix, can you integrate each entry (with respect to the right variable) to figure out what the original functions were?

Problem Set IV

Parameterization Invariance

Below are four different curves which all trace out the same set of points in the plane: the segment of the x axis between 0 and 4.

$$\begin{aligned} \alpha(t) &= (t, 0) & t &\in [0, 4] \\ \beta(t) &= (2t, 0) & t &\in [0, 2] \\ \gamma(t) &= (t^2, 0) & t &\in [0, 2] \end{aligned}$$

Because these all describe the same set of points, we of course want them to have the same length! But our *definition* of the length function involves integrating infinitesimal arclengths (derivatives), and these curves don't all have the same derivative! Thus, to *really* make sure our definition makes sense, we need to check that it doesn't matter which parameterization we use, we will always get the same length.

Exercise 32.25. Check these three parameterizations of the segment of the x-axis from 0 to 4 all have the same length.

After doing this exercise, read the proof of Theorem 10.1 (which follows this exercises' original location in the text): you don't have to write anything here, but it's good to see how do to this in general with the chain rule!

Non-Isometries

Exercise 32.26. Write down a linear map that sends both $\langle 1, 0\rangle$ and $\langle 0, 1\rangle$ to unit vectors, but is not an isometry.

This shows there's not a shortcut to checking something is an isometry by just seeing what happens to the basis vectors!

Composition and Inversion of Isometries

Exercise 32.27. If ϕ and ψ are two isometries of $\mathbb{E}^2$, prove that both the composition $\phi \circ \psi$ is an isometry, and the inverse ϕ^{-1} is an isometry.

Remember, you will need to explain why at every point $p \in \mathbb{E}^2$, these maps do not change the lengths of tangent vectors. This will probably involve the multivariable chain rule, whether you do it in words, or in equations!

Homogenity and Isotropy

In class we built a couple different sorts of isometries from the basic ones we constructed by hand (translations and rotations about 0). In this exercise, you are to prove the existence of another very useful isometry. We will use this homework problem all the time

Exercise 32.28 (Moving from p to q.)**.** Given any two pairs p, v_p and q, w_q of points p, q in Euclidean space and unit tangent vectors $v_p \in T_p\mathbb{E}^2$, $w_q \in T_q\mathbb{E}^2$ based at them, prove that there exists an isometry taking v_p to w_q.

Hint: try to combine pieces we know about, and prove the result does what you need by applying it both to the point p, and applying its derivative to the vector v.

Lines of Symmetry

In this exercise we will investigate the third potential definition of line, which involves isometries.

Definition 32.1 (Line of Symmetry)**.** A *fixed point* of an isometry $\phi : \mathbb{E}^2 \to \mathbb{E}^2$ is a point p with $\phi(p) = p$.

A curve γ is called a *line of symmetry* of $\mathbb{E}^2$ if there exists an isometry which fixes $\gamma(t)$ for all t.

In this exercise, you show that the curves which are lines of symmetry are exactly the same as the curves which are lines under Archimedes' definition!

Exercise 32.29 (Reflections in Any Line)**.**

- Show that map $\phi(x, y) = (x, -y)$ is an isometry of $\mathbb{E}^2$. Explain why this shows that the x-axis is a line of symmetry of the plane.

- Show that every curve which is distance-minimizing in the plane is also a line of symmetry. *Hint: given an isometry that reflects in the x axis, can you build an isometry that reflects in any other line? Consider moving the line to the x axis, reflecting, and then moving back.*

Equilateral Triangles Revisited

In this question we will revisit two problems from Greek geometry. That is we will be *re-proving* things we knew before, so we know they are still true in our new foundations!

This problem requires the distance function on the Euclidean plane, which we did not get to in class on Thursday, but will cover on Tuesday. However - you all areadly know the distance function so you can absolutely complete the homework now if you like!

Definition 32.2. If $p = (x, y)$ and $q = (a, b)$ are two points in the Euclidean plane, the **distance** from p to q is the length of the shortest curve connecting them. Working this out, we find the familiar pythagorean theorem:

$$\text{dist}(p, q) = \sqrt{(x-a)^2 + (y-b)^2}$$

First, we re-prove the very first proposition of Euclid, the existence of an equilateral triangle. Then we redo your earlier homework problem on finding a smaller equilateral triangle inside of it, of half the side lengths (but this is *much easier* with our new tools!).

Exercise 32.30.

- Beginning with the segment $[0, \ell]$ along the x-axis, construct an equilateral triangle by finding the coordinates of a point $p = (x, y) \in \mathbb{E}^2$ which is equidistant from both endpoints of the segment.
- Re-prove that inside of this equilateral triangle, you can inscribe a smaller one with exactly half the side length. *Hint: just find where the vertices should be, and then measure the distances between them!*

Length of a Parabola

Arclength integrals give a good opportunity to practice a lot of Calculus II integration techniques. Even for relatively simple curves like the parabola, the answers can be quite nontrivial!

Exercise 32.31 (The Length of a Parabola)**.** Find the length of the parabola $y = x^2$ between from $x = 0$ to $x = a$, following the steps below.

- Paramterize the curve as $c(t) = (t, t^2)$, show the arclength integral is $L(a) = \int_{[0,a]} \sqrt{1 + 4t^2}$
- Perform the trigonometric substitution $x = \frac{1}{2}\tan\theta$ to convert this to some multiple of the integral of $\sec^3(\theta)$.
- Let $I = \int \sec^3(\theta)d\theta$ and do integration by parts with $u = \sec\theta$ and $dv = \sec^2\theta$.
- After parts, use the trigonometric identity $\tan^2\theta = \sec^2\theta - 1$ in the resulting integral to get another copy of $I = \int \sec^3\theta d\theta$ to appear.
- Get both copies of I to the same side of the equation and solve for it! To check your work at this stage, you should have found that

$$\int \sec^3\theta d\theta = \frac{1}{2}\sec\theta\tan\theta + \frac{1}{2}\ln|\sec\theta + \tan\theta|$$

- Relate this back to your original integral, and undo the substitution $x = \frac{1}{2}\tan\theta$: can you use somet trigonometry to figure out what $\sec\theta$ is?
- Finally, you have the antiderivative in terms of x! Now evaluate from 0 to a.

Minimizing a Function by Minimizing its Square

Here's a problem that's straight up single variable calculus, but turns out to be a quite useful "trick" in geometry! Oftentimes we want to minimize a function in geometry (like arclength, or distance) but this turns out to be technically hard because of the square root. One might wonder - what happens if I square the function, and try to minimize that instead? That will have an easier formula (no roots), but will I get the right answer?

This exercise shows, yes you will!

Exercise 32.32 (Minimizing the Square: A Very Useful Trick!)**.** Let $f(x)$ be a differentiable positive function of one variable, and let $s(x) = f(x)^2$ be its square. Show that the minima of $s(x)$ and $f(x)$ occur at the same points, by following the steps below:

- First, assume $x = a$ is the location of a minimum of f. What does the first and second derivative test tell you about the values $f'(a)$ and $f''(a)$? Use this, together with the fact that $f(a) > 0$ to show that $x = a$ is also the location of a minimum of s (using the second derivative test).
- Conversely, assume $x = a$ is the location of a minimum of $s(x)$. Now, you know information about the derivatives $s'(a)$ and $s''(a)$. Use this to conclude information about $f'(a)$ and $f''(a)$ to show that a is a minimum for f as well.

Problem Set V

The Parallel Postulate

Recall that *Playfair's Axioms* (already suggested by Proclus in the 400s) was a simpler re-phrasing of Euclid's original fifth postulate on parallel (that is, nonintersecting) lines.

Proposition 32.1. *Given any line L in $\mathbb{E}^2$, and any point $p \in \mathbb{E}^2$ not lying on L, there exists a unique line Λ through p which does not intersect p.*

Exercise 32.33. Prove Proposition 32.1.

Hint: use isometries to help you out!
First, use an isometry to move L to the x-axis. Then, use another isometry to keep L on the x axis, but to move p to some point along the y axis. Then, prove that through any point on the y axis there is a unique line that does not intersect the x-axis.

Similarities and Lines

We saw in Theorem 12.2 that any isometry will carry a line to another line. The same is true more generally of similarities:

Exercise 32.34 (Similarities Send Lines to Lines)**.** Let $\gamma : \mathbb{R} \to \mathbb{E}^2$ be a line, and $\sigma : \mathbb{E}^2 \to \mathbb{E}^2$ be a similarity. Prove that $\sigma \circ \gamma$ is also a line.

*Hint-replicate the proof of Theorem 12.2 as closely as possible, replacing the isometry ϕ with the similarity σ, and keeping track of the scaling factors of σ versus σ^{-1} (Proposition 11.4).

Distance to a Line.

In this problem it's probably helpful to use the 'calculus trick' offered as an optional problem last week: that is, if you are looking to minimize a positive function $f(x)$, you can instead try to minimize the function $f(x)^2$, and you'll find the same x-value achieves the minimum.

The reason this is useful to us is that the distance function in geometry has a square root in it, and differentiating roots can be a lot of work. So this says instead we can minimize the *square of distance* to find the right point.

*his if you like
an you use an
ine where the
e isometry to
final answer.*

Exercise 32.35 (Closest Point on Line)**.** Let L be the line traced by the affine curve $\gamma(t) = \begin{pmatrix} at + c \\ bt + d \end{pmatrix}$, and O be the origin.

Alternatively, do this for the line $(2t + 1, 3t - 4)$**, to save yourself a lot of** *abcd's***.**

- Find the point $p \in L$ which is closest to O.
- Calculate dist(O, L).
- What angle does the segment connecting p to O make with the line L? *(We haven't reviewed the definition of angle yet, so just use your knowledge of angles from precalculus: pick a line an compute an example!)*

This problem shows how we can use calculus as a *tool* to discover a geometric fact: here we learned something about where the closest point on a line is located (by finding it with calculus, then calculating the angle formed).

Often calculus can provide the tools needed for discovery of a new fact, but once its known, one can often go back and find a *geometric* (more in the style of Euclid) proof that all of a sudden makes the conclusion feel inevitable.

Exercise 32.36 (Closest Point: Geometric Reasoning)**.** Now that we know the answer, formulate in your own words a geometric proposition that describes which point p on a line is closer to a given point q not on that line.

Prove this propostion without just taking a derivative, try to reason more like the Greeks, using other facts and theorems that we've proven.

Hint: if you pick some other point r along L, can you draw a triangle using p, q, r? How can we use the geometry of this triangle to show that r is farther from q than p was? Does the Pythagorean theorem say anything useful?

Intersecting Circles

Recall that in all of Euclid's axioms, conditions for intersections with circles were never specified! Indeed - Euclid intersected two circles in his construction of the equilateral triangle. Now that we have a precise description of circles in our new foundations, we can fix this gap:

Exercise 32.37. Prove that two circles intersect each other if the distance between their centers is less than or equal to the sum of their radii.

Alternatively: do this for the specific case where the circles radii are 3 and 4, and the distance between thier centers is 5.

Hint: start by applying an isometry to move one of the circles to have center $(0, 0)$*, and then another isometry to rotate everything so the second circle has center* $(x, 0)$ *along the x-axis. This will make computations easier!*

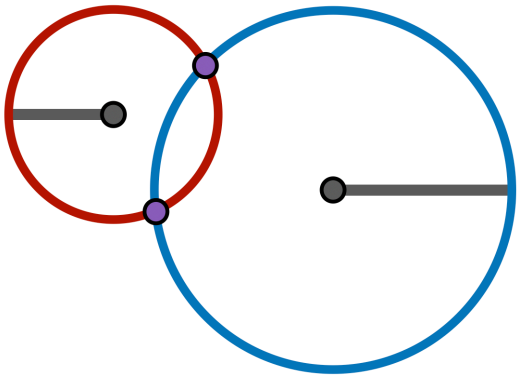


Figure 32.11.: Two circles intersect if the distance between their centers is less than or equal to the sum of their radii.

Parabolas

A parabola was one of few curves that the Greeks understood. However, while we often think of a parabola *algebraically* that was not their original definition!

Definition 32.3 (Parabola)**.** Given a point f (called the focus) and a line L not containing f (called the directrix), the parabola with focus f and directrix L is the curve C of points where for each $p \in C$ the distance to the focus equals the distance to the line:

$$\text{dist}(p, f) = \text{dist}(p, L)$$

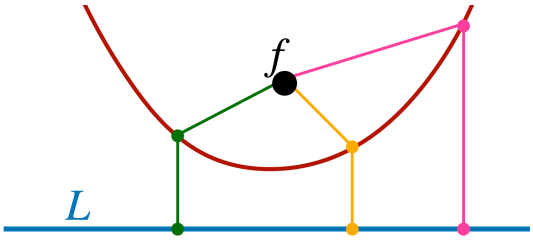


Figure 32.12.: A parabola is the set of points which are the same distance from a point (the focus) and a line (the directrix). In this figure, line segments of the same color are supposed to be the same length.

Exercise 32.38. In this problem we confirm that $y = x^2$ is indeed a parabola! Let L be a horizontal line intersecting the $y-$axis at some point $(0, -\ell)$, and $f = (0, h)$ be a point along the y-axis for $\ell, h > 0$.

- Write down an algebraic equation for the coordinates of a point (x, y) determining when it is on the parabola with focus f and directrix L.

- Find which point f and line L make this parabola have the algebraic equation $y = x^2$.

Problem Set VI

Right Angles

In this set of two problems, we make use of the fact that we can *finally* measure angles rigorously in our new geometry, to reprove an important fact we already know, and to prove the one remaining postulate of Euclid: the 4th.

Exercise 32.39. Prove that rectangles exist, using all of our new tools! (Ie write down what you know to be a rectangle, explain why each side is a line segment, parameterize it to find the tangent vectors at the vertices, and use the dot product to confirm that they are all right angles).

To prove Euclid's 4th postulate, we need to first phrase it more precisely than Euclid's original *all right angles are equal.*

Proposition 32.2 (Euclids' Postulate 4)**.** *Given the following two configurations:*

- *A point p, and two orthogonal unit vectors u_p, v_p based at p*
- *A point q, and two orthogonal unit vectors a_q, b_q based at q*

There is an isometry ϕ of $\mathbb{E}^2$ which takes p to q, takes u_p to a_q, and v_p to b_q.

Exercise 32.40. Prove Euclids' forth postulate holds in the geometry we have built founded on calculus.

Hint: there's a couple natural approaches here.

- *You could directly use Exercise 32.28 from a previous homework to move one point to the other and line up one of the tangent vectors. Then deal with the second one: can you explain why its either already lined up, or will be after one reflection?*
- *Alternatively, you could show that every right angle can be moved to the "standard right angle" formed by $\langle 1, 0\rangle, \langle 0, 1\rangle$ at O. Then use this to move every angle to every other, transiting through O*

Measurement of the Circle

The half angle identities played a crucial role in Archimedes' ability to compute the preimeter of n gons in his paper *The Measurement of the Circle*. Indeed, to calculate the circumference of an inscribed n-gon, its enough to be able to find $\sin \tau/(2n)$:

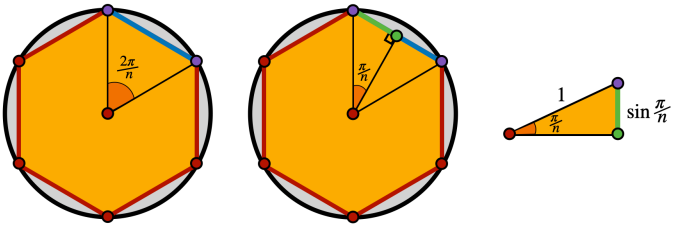


Figure 32.13.: The side-length of an inscribed n-gon is $2 \sin \frac{\tau}{2n}$, found via bisecting the side to form a right triangle. The perimeter of the n-gon is just n times this.

By repeatedly bisecting the sides, we can start with something we can directly compute - like a triangle, and repeatedly bisect to compute larger and larger n-gons.

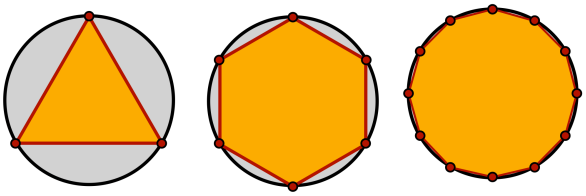


Figure 32.14.: Archimedes' method: repeatedly doubling the number of sides of the n-gon to get polygons approaching the circle.

In the book, I use the half-angle identities to compute the exact value of $\sin \tau/12$. Follow that example further, to retrace the steps of Archimedes.

Exercise 32.41. Continue to bisections until you can compute $\sin(\tau/(2 \cdot 96))$. What is the perimeter of the regular 96-gon (use a computer to get a decimal approximation, after your exact answer).

Explain how we know that this is *provably an underestimate* of the true length, using the definition of line segments.

Optional: *Be brave - and go beyond Archimedes! Compute the circumference of the 192-gon.*

Quadrature of the Parabola

This is also a two-problem series, where we complete Archimedes famous *Quadrature of the Parabola* using modern tools from calculus. Archimedes problem was about a *parabolic segment*: that is, the region enclosed by a parabola and a line segment connecting two points on the parabola. Instead of working in complete generality like Archimedes, we will be content to just study a special case in this problem. We will look at the parabola $y = x^2$, and the parabolic segment cut out by this and the line connecting $(-1, 1)$ to $(2, 4)$.

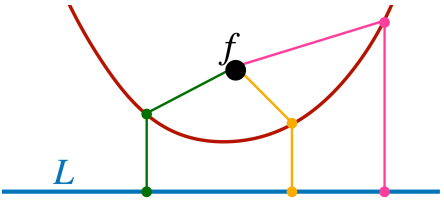


Figure 32.15.: The Parabolic Segment in this Problem

Recall Archimedes main result: the area of this parabolic segment is 4/3rds the area of the largest inscribed triangle, which is the triangle whose base is the line segment, and third vertex lies on the parabola at the point where the tangent line is parallel to the base.

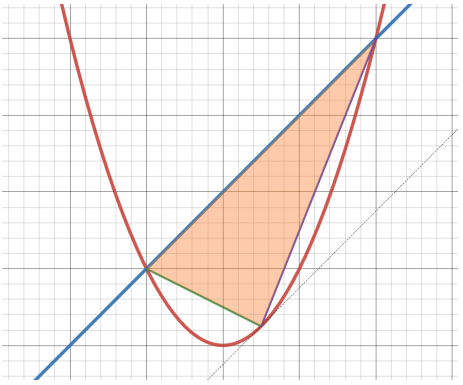


Figure 32.16.: The overall segment's are is 4/3rds that of this triangle.

Exercise 32.42.

- Write down a formula for the area of the triangle whose third vertex lies at (x, x^2)

Hint: instead of finding the height of the triangle to use $\frac{1}{2}bh$, can you use the fact that the determinant of a matrix calculates the area of a parallelogram whose sides are the column vectors, and that the area of a the triangle you want is half a parallelogram?

- Use Calculus I to find the point x where the inscribed triangle has maximal area. Then show that Archimedes was right: the slope of the tangent line to the parabola at this point is exactly the same as the slope of the line segment forming the triangle's base!

This gives us the starting point: the area for which archimedes wishes to compare the parabolic segment. Next - we need to find the parabolic segment's area. We could of course follow Archimedes' original method (and if you choose to, this can be your class project!) But here, we will use our modern tools and confirm the answer with calculus:

Exercise 32.43. Compute the area of the parabolic segment (via integration, as the area between two curves). Show that its exactly 4/3rds the area of the triangle!

The Area and Circumference Constants

A circle has a *circumference constant*: the ratio of its radius to its circumference, which we've named τ. But it also has an *area constant: the ratio of its area to the square of its radius, which we've named π.

It was Archimedes who first showed that these two constants were intimately related, by finding that $\tau = 2\pi$. Here we will again use our modern tools to reprove Archimede's result.

Tau is the circumference of the unit circle $x^2 + y^2 = 1$. We can parameterize the top half of this circle via

$$\gamma(t) = (t, \sqrt{1-t^2})$$

And then compute its arclength via the integral

$$\frac{\tau}{2} = \int_{-1}^{1} \|\gamma'(t)\| dt = \int_{-1}^{1} \frac{1}{\sqrt{1-t^2}} dt$$

But we can also write down the *area* of the circle as an integral: the top half of the circle is $y = \sqrt{1-x^2}$ and the bottom half is $y = -\sqrt{1-x^2}$ so the area is

$$\pi = \int_{-1}^{1} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} dy dx = \int_{-1}^{1} 2\sqrt{1-x^2} dx$$

Your goal in this problem is to show these two integrals are equal to one another!

Exercise 32.44. Prove that

$$\int_{-1}^{1} \frac{1}{\sqrt{1-t^2}} dt = \int_{-1}^{1} 2\sqrt{1-x^2} dx$$

Thus showing that $\frac{\tau}{2} = \pi$.

Hint: Do u-substitutions to the integrals to make them into the same integral. The goal isn't to evaluate them and get a number! This is just a Calc II problem - but a tricky one, so here's one outline you could follow:

(1) Rewrite the area integrand $\sqrt{1-x^2}$ as $\frac{1-x^2}{\sqrt{1-x^2}}$. Use properties of integrals to break this into two integrals, and see

$$\pi = \tau - \int_{-1}^{1} \frac{2x^2}{\sqrt{1-x^2}} dx$$

(2) Now we just have to evaluate this new integral: Do the u-substitution $u = \sqrt{1-x^2}$ to this, to show that

$$\int_{-1}^{1} \frac{2x^2}{\sqrt{1-x^2}} dx = \int_{-1}^{1} 2\sqrt{1-u^2} du = \pi$$

(This u-sub requires some work: you'll need at some point to solve for x in terms of u!)

(3) Now just assemble the pieces! You never completed a single integral, but you still managed to prove that $\tau = 2\pi$.

Problem Set VII

Trigonometric Identities:

The following exercise has you compute with some trigonometric identities, which we needed to find the volume of spheres:

Exercise 32.45 (Integrating $\cos^4(\theta)$)**.** Start with the angle sum-identity we derived in class some lectures back

$$\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b)$$

Use this to derive an identity relating $\cos(2\theta)$ to $\cos^2(\theta)$ (we did most of this in class - but repeat it for yourselves). Now use this identity twice to show

$$\cos^4(\theta) = \frac{3}{8} + \frac{\cos 2\theta}{2} + \frac{\cos 4\theta}{8}$$

by writing $\cos^4\theta = (\cos^2\theta)^2$. Then use this to confirm that

$$\int_0^{\frac{\tau}{4}} \cos^4\theta = \frac{3}{8}\frac{\tau}{4}$$

This result was crucial to our calculation of the volume of the four dimensional sphere in class, where we showed (via a trigonometric substitution)

$$\text{vol} = \frac{8\pi}{3}\int_0^{\tau/4} \cos^4(\theta)d\theta$$

Using this result gave us the final answer:

$$\begin{aligned}\text{vol} &= \frac{8\pi}{3}\frac{3}{8}\frac{\tau}{4} \\ &= \frac{\pi\tau}{4} \\ &= \frac{\pi^2}{2}\end{aligned}$$

The 5-Dimensional Sphere

The unit sphere in five dimensions is given by the set of points (x, y, z, w, u) satisfying

$$x^2 + y^2 + z^2 + w^2 + u^2 = 1$$

Exercise 32.46. Calculate the volume of this space by slicing along the u direction. Show that the slices are four dimensional spheres: what's the radius? Use the volume formula in four dimensions we derived in class

$$\text{vol}(r) = \frac{\pi^2}{2}r^4$$

to write down the volume of each slice, and then perform the integral to confirm that the 5-dimensional volume is

$$\frac{8\pi^2}{15}$$

(This will not need any trigonometric substitution) Once you have this, find the "*surface area*" of this 5-dimensional sphere by differentiation.

High Dimensional Spheres and Cylinders

Exercise 32.47. The discovery Archimedes was most proud of was that the surface area of the sphere in 3-dimensions was the same as the area of the smallest cylinder surrounding it.

Is the same true in four dimensions? (A four-dimensional cylinder has a sphere's surface as its "base", just like a three dimensional cylinder has a circle's length as its base!)

Problem Set VIII

The Dot Product

One thing we used in our arguments building up the basics of spherical geometry was the fact that the dot product has a nice derivative rule.

Exercise 32.48 (Product Rule for Dot Product)**.** Let $f(t) = \langle f_1(t), f_2(t), f_3(t)\rangle$ and $g(t) = \langle g_1(t), g_2(t), g_3(t)\rangle$ be two vector functions. Prove that the dot product satisifes the product rule:

$$\frac{d}{dt}\left(f(t)\cdot g(t)\right) = f'(t)\cdot g(t) + f(t)\cdot g'(t)$$

Our use of the dot product overall is as a tool to give the sphere *geometry* it defines what we mean by infinitesimal length and by angle. Often we will use this just as a theoretical tool - but its good to get some hands-on practice at the beginning, measuring some actual angles.

Exercise 32.49. Consider the curves $\alpha(t) = (\cos t, \sin t, 0)$ (the equator of the sphere), and $\beta(t) = (0, \sin(t), \cos(t))$ (a line of longitude). Prove that they

- Intersect each other at the $t = \pi/2$
- Form a right angle at their point of intersection.

Draw a picture of this situation in 3D on a sphere.

Isometries

Recall the definition of an isometry of ^{2} is a function $\phi : {}^2 \to {}^2$ which preserves the dot product (or equivalently, preserves infinitesimal lengths).

Exercise 32.50. A *permutation matrix* is a square matrix where every row and column has exactly one "1", and the other entries are zero. Prove the following permutation matrix

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

can be used to define an isometry of ^{2} by the formula

$$\phi(x, y, z) = A \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

directly from the definition of isometry.

In class we are working our way to prove some facts about isometries, mimicking what we did in the Euclidean plane. In particular, on Tuesday we will prove the following two important facts.

Theorem 32.1 (The Sphere is Homogeneous)**.** *Given any two points p and q on the sphere, there is an isometry taking p to q:*

Proposition 32.3. *Let N be the north pole, and v be any unit vector in $T_N{}^2$. Then there exists an isometry ϕ of the sphere which fixes N and takes $\langle 1, 0, 0\rangle \in T_N{}^2$ to v.*

The first is an analog of translations of the Euclidean plane: we can always find an isometry of the sphere that takes any point to any other. And the second is similar to when we proved that you could build rotations of $\mathbb{E}^2$ about the origin (we'll actually prove it, using that exact theorem!)

Your goal in these next problems is to *use these theorems* to prove even more: first, prove that you can actually rotate the sphere fixing any point you wish (not just the north pole!)

Exercise 32.51. Use Proposition 32.3 and Theorem 32.1 to show the sphere is isotropic: that given any point $p \in {}^2$ and any two unit vectors $v, w \in T_p{}^2$, there exists an isometry of ^{2} fixing p and taking v to w.

(Hint: first show you can do this when p is the north pole! Then use homogenity and a *conjugation*. Be inspired by the Euclidean proof!)

Next, just like in Euclidean space we often find it useful to combine homogenity and isotropy into one condition: we can move any point and any tangent vector to any other point and tangent vector that we like!

Exercise 32.52. Let p, q be any two points on the sphere, and v a unit tangent vector at p and w a unit tangent vector at q. Then there is an isometry of ^{2} taking (p, v) to (q, w).

Hint: Look back at Ex 4 Homework 4!

Problem Set IX

Acceleration

We saw in class that the *acceleration* of a curve on the *unit sphere* is the projection of its $\mathbb{E}^3$ acceleration vector onto the tangent plane. In terms of the curve $\gamma(t)$, that worked out to

$$\text{acc}_\gamma(t) = \gamma''(t) - (\gamma''(t) \cdot \gamma(t))\,\gamma(t)$$

Exercise 32.53 (Geodesic Curvature)**.** The magnitude of the acceleration of a unit speed curve is called its *geodesic curvature*: its a way to measure how much that curve differs from a geodesic. This exercise, you will calculate the geodesic curvature of the circle of radius r on ^{2}:

- Let C be a circle on ^{2} of radius r (to make calculations easy, let C be centered at the north pole if you like).
- Write down a parameterization of C *(hint: you know the plane that C lies in, and its Euclidean radius in that plane!)*
- Find a parameterization of C that has speed 1 (*hint if you wrote down a parameterization above, what speed does it travel at? Can you adjust it so the new curve has speed* 1*?*)
- What is the acceleration felt on ^{2} if you go around a circle of radius r at unit speed? What is it's magnitude?

The idea of measuring acceleration along a surface in $\mathbb{E}^3$ as the projection of the second derivative onto the tangent space is foundational to the study of surfaces beyond just the sphere (it is one of the fundamental concepts in *differential geometry*). When the acceleration of a curve γ is equal to zero on a surface, then we say that curve is a *geodesic* on the surface. So, the equations we get by setting the acceleration equal to zero give us a set of *differential equations* that tell us what the geodesics are! These are called the *geodesic equations*

In this next problem, we will get a small taste of what happens in differential geometry, when our space is not nice and symmetric like the plane or the sphere, and we have to resort to finding the geodesic equations and solving them (you won't have to solve them! Just find them....)

Exercise 32.54 (Geodesics on Surfaces)**.** Let S be the surface $z = x^2 + y^2$, which is the output of the function $F : \mathbb{E}^2 \to \mathbb{E}^3$ given by

$$F(x, y) = (x, y, x^2 + y^2) \in \mathbb{E}^3$$

.
Let $p = F(x, y)$ be a point on S.

- Use calculus to find two tangent vectors to the graph at a point p: *Hint:* $DF_{(x,y)}\langle 1, 0\rangle$ *and...?*
- Find a normal vector n to the graph at p using these and the cross product.
- Write down the projection of a vector $v = \langle a, b, c\rangle$ onto the normal vector n.
- Write down the projection of $v = \langle a, b, c\rangle$ onto the tangent plane T_pS.

This is all the data we need to be able to compute acceleration *along the surface S*! Let $\gamma(t)$ be a curve that lies on the surface, so

$$\gamma(t) = F(x(t), y(t)) = \left(x(t), y(t), x(t)^2 + y(t)^2\right)$$

for some function $x(t)$ and $y(t)$.

- Find γ''
- Find the acceleration of γ on the surface S, in terms of $x(t)$ and $y(t)$.
- What are the geodesic equations for S?

Curvature

We saw in class that the circumference of a circle of radius r on ^{2} is given by

$$C(r) = 2\pi \sin(r)$$

Furthermore, we saw that the area is given by

$$A(r) = \int_0^r C(r)dr = 2\pi(1 - \cos(r))$$

The idea of *curvature* is a

Exercise 32.55. Check this, that as $r \to 0^+$ the following limits both exist, and are both equal to zero:

$$\lim_{r\to 0} \frac{C_{\mathbb{E}^2}(r) - C_2(r)}{r} = 0$$

$$\lim_{r\to 0} \frac{C_{\mathbb{E}^2}(r) - C_2(r)}{r^2} = 0$$

But

$$\lim_{r\to 0} \frac{C_{\mathbb{E}^2}(r) - C_2(r)}{r^3} = \frac{\pi}{3}$$

Hint: L'Hospital's rule.

Because the first two limits here are zero, they tell us that the difference between the circumference of a Euclidean circle and a Spherical is *very small* indeed - they agree to first and second order, and their difference is only revealed at the next (cubic) level. We use this to define the *curvature* of any surface, where we normalize things so that the curvature of the unit sphere comes out to be 1:

Definition 32.4. Let S be any surface, and C the circumference function for circles drawn on that surace at a point p. Then the curvature of S at p is

$$\kappa = \frac{3}{\pi} \lim_{r\to 0} \frac{C_{\mathbb{E}^2}(r) - C(r)}{r^3}$$

We dont need to work with circumference however; its possible to measure the curvature of space using area as well!

Exercise 32.56. Which power n is the smallest such that

$$\lim_{r\to 0} \frac{A_{\mathbb{E}^2}(r) - A_2(r)}{r^n}$$

has a nonzero value? For this n, what is the value of the limit? Use this to write down a definition of curvature in terms of the area of circles, normalized so that the curvature of the unit sphere is 1.

Spheres of Different Sizes

Unlike the Euclidean plane, spheres have no nontrivial similarities: in fact, if you apply a similarity of $\mathbb{E}^3$ to the sphere, it sends it to a larger or smaller sphere - not to itself! Because of this there is not just *one* spherical geometry like there was for the plane, but *many*. For each positive real number R we can define *spherical geometry of radius R*, denoted $\overset{2}{R}$, as follows.

Definition 32.5 (Spherical geometry of Radius R.)**.** Let $\overset{2}{R}$ denote the set of points which are distance R from the origin in $\mathbb{E}^3$. For each point $p \in \overset{2}{R}$, the tangent space $T_p\overset{2}{R}$ consists of all points in $\mathbb{E}^3$ which are orthogonal to p (definition unchanged from the unit sphere), and the dot product for measuring infinitesimal lengths and angles is the standard dot product on $\mathbb{E}^3$ (also unchanged from the unit sphere).

The development of each of these spherical geometries is qualitatively very similar to that for ^{2}: we can see without any change that $(x, y, z) \mapsto (x, y, -z)$ is an isometry so the equator is a geodesic, and orthogonal transformations are still isometries so all great circles are geodesics.

What changes is the quantitative details: the formulas for length area and curvature. In the next two problems, your job is to redo the calculations that I did for ^{2}, for the geometry ^2_R:

Exercise 32.57 (Circumference and area.)**.** What is the formula for the circumference and radius of a circle of radius r on ^2_R?

Hint: base your circles at $N = (0, 0, R)$ and look back at our arguments from class to see what must change, and what stays the same.

Exercise 32.58. Using the definition of curvature as a limiting ratio of circumference (Definition 32.4), compute the curvature of ^2_R.

Problem Set X

Platonic Solids

In these problems we will investigate regular polygons on the sphere. Recall we call a polygon *regular* if it has rotational symmetries about its center: in particular this implies that all its sides are the same length, and all its angles have the same measure (since isometries preserve both lengths and angles).

In the Euclidean plane, we know that regular polygons of all side numbers ≥ 3 exist (these are how Archimedes approximated the circle, after all!), but their angles are strictly determined by their number of sides. We proved in a previous homework that the angle sum of an n-gon is $(n-2)\pi$, and if all the angles of a regular n gon are equal, each angle must measure $\theta_n = \frac{n-2}{n}\pi$.

This puts a strong restriction on which regular polygons can be used to tile the plane. To tile the plane, a necessary (but not sufficient) condition is that we need to be able to fit k copies of each polygon around a vertex, without any gaps or overlaps. This tells us that the angles of a polygon that can tile must be $\theta = \frac{2\pi}{k}$.

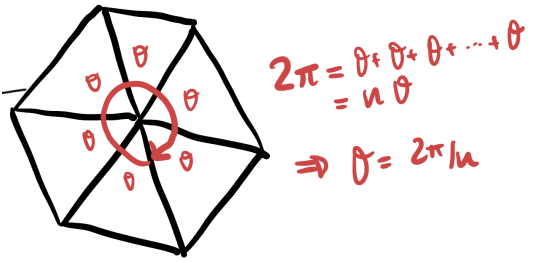


Figure 32.17.: Angles need to be an integer divisor of 2π to fit evenly around a point without gaps or overlap.

Thus, to figure out which polygons even have a chance of tiling the Euclidean plane, we want to know for which n (the number of sides) there the angle θ_n is actually 2π over an integer. We can start listing:

$$\theta_3 = \frac{3-2}{3}\pi = \frac{\pi}{3} = \frac{2\pi}{6}$$

$$\theta_4 = \frac{4-2}{4}\pi = \frac{\pi}{2} = \frac{2\pi}{4}$$

$$\theta_5 = \frac{5-2}{5}\pi = \frac{3\pi}{5}$$

$$\theta_6 = \frac{6-2}{6}\pi = \frac{2\pi}{3}$$

$$\theta_7 = \frac{7-2}{7}\pi = \frac{5\pi}{7}$$

Thus, we see that its possible to fit six triangles around a vertex, four squares around a vertex and three hexagons around a vertex, but as the angles θ_5 and θ_7 aren't even divisions of 2π, there's no nice way to fit pentagons or 7-gons around a vertex, and thus no hope of using them to tile the plane.

This is the start to the classification of regular tilings of the plane, where by what we see from the angle measures, its possible for triangles, squares and hexagons, but impossible for all other shapes!

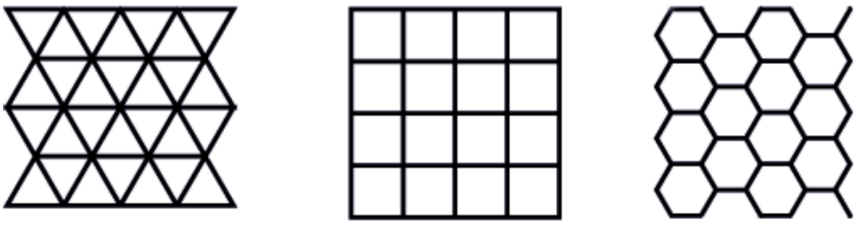


Figure 32.18.: The three regular polygons that tile the Euclidean plane.

Our goal here is to investigate what changes on the sphere.

Exercise 32.59 (Spherical Pentagons)**.**

- Find a relationship between the area A of a spherical regular pentagon and its angle measure α. *Hint: divide the spherical pentagon into five triangles*
- Show that there exists a spherical pentagon whose angle evenly divides 2π: how many of these spherical pentagons fit around a single vertex?
- What is the area of such a spherical pentagon? How many of these pentagons does it take to cover the entire sphere?

For the pentagon, there was only one possibility, as the restriction that the angle at a vertex be $2\pi/k$ is so restrictive. However, for triangles, there are *three possibilities*!

Exercise 32.60 (Spherical Triangles)**.** There are three different equilateral triangles that can be used to tile the sphere. Find them! For each triangle:

- How many fit around each vertex?
- How many are needed to cover the sphere?
- What platonic solid does this correspond to?

The Pythagorean Theorem

The fundamental formula in Euclidean trigonometry is the *Pythagorean theorem* which allows us to measure the length of the hypotenuse of a right triangle in terms of the other side lengths.

The goal of this exercise is to derive the spherical counterpart to this:

Theorem 32.2 (Spherical Pythagorean Theorem)**.** *Given a right triangle on* ^{2} *with side lengths* a, b *and hypotenuse* c*, these three lengths satisfy the equation*

$$\cos(c) = \cos(a)\cos(b)$$

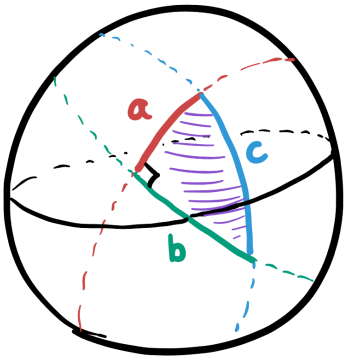


Figure 32.19.: A spherical right triangle.

Exercise 32.61 (Deriving The Pythagorean Theorem)**.** Prove that the formula given above really does hold for the legs and hypotenuse of a right triangle on ^{2}, using the distance formula that we've already calculated:

$$\cos \operatorname{dist}(p, q) = p \cdot q$$

Hint: move your triangle so the right angle is at the north pole, and the legs are along the great circles on the xz and yz plane. Now you can write down exactly what the other two vertices are since you know they are distance a and b along these geodesics from N, and these geodesics are unit circles in the xz and yz planes!*

On a sphere of radius R, a similar formula exists: here to be able to use arguments involving angles we need to divide all the distances by the sphere's radius, but afterwards an argument analogous to the above exercise yields

$$\cos\left(\frac{c}{R}\right) = \cos\left(\frac{a}{R}\right)\cos\left(\frac{b}{R}\right)$$

Its often more useful to rewrite this result in terms of the curvature $\kappa = 1/R^2$

Theorem 32.3 (Pythagorean Theroem of Curvature κ)**.** *On the sphere of curvature κ, the two legs a, b and the hypotenuse c of a right triangle satisfy*

$$\cos\left(c\sqrt{\kappa}\right) = \cos\left(a\sqrt{\kappa}\right)\cos\left(b\sqrt{\kappa}\right)$$

As a sphere gets larger and larger in radius, it better approximates the Euclidean plane. We might even want to say something like in the limit $R \to \infty$ (so, $\kappa \to 0$) the spherical geometry *becomes euclidean.* But how could we make such a statement precise? One way is to study what happens to the theorems of spherical geometry as

$\kappa \to 0$; and show that they become their Euclidean counterparts. The exercise below is our first encounter with this big idea:

Exercise 32.62 (Euclidean Geometry as the Limit of Shrinking Curvature)**.** Consider a triangle with side lengths a, b, c in spherical geometry of curvature κ. As $\kappa \to 0$, the arguments of the cosines in the Pythagorean theorem become very small numbers, so it makes sense to approximate approximate these with the first terms of their Taylor series.

Compute the Taylor series of both sides of

$$\cos\left(c\sqrt{\kappa}\right) = \cos\left(a\sqrt{\kappa}\right)\cos\left(b\sqrt{\kappa}\right)$$

in the limit $\kappa \to 0$, we can ignore all but the first nontrivial terms. Show here that only keeping up to the quadratic terms on each side recovers the Euclidean Pythagorean theorem, $c^2 = a^2 + b^2$.

Trigonometry

Following the derivation of the spherical pythagorean theorem, we might next hope to discover relationships between the sides of a spherical right triangle and its angle measures. And, indeed we can!

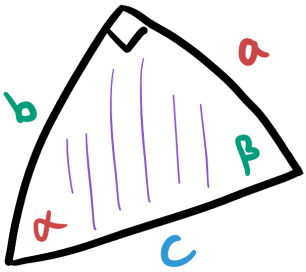


Figure 32.20.: A right triangle with angles α, β and opposite sides a, b.

The corresponding laws of spherical trigonometry are as follows:

Theorem 32.4 (Spherical Trigonometric Relations)**.** *For a right triangle with angles α, β, corresponding opposite sides a, b and hypotenuse c the following relations hold:*

$$\sin\alpha = \frac{\sin a}{\sin c} \qquad \sin\beta = \frac{\sin b}{\sin c}$$

$$\cos\alpha = \frac{\tan b}{\tan c} \qquad \cos\beta = \frac{\tan a}{\tan c}$$

You will not be responsible for the derivation of these formulas, nor for remembering them: if you ever need them they will be given to you!

One of the most biggest differences between spherical trigonometry and its Euclidean counterpart is that its possible to derive formulas for the length of a triangles' sides *in terms of only the angle information*! This is impossible in Euclidean space because of the existence of similarities: there are plenty of pairs of triangles that have all the same angles but wildly different side lengths! No so in the geometry of the sphere.

Exercise 32.63. Using the trigonometric identities in Theorem 32.4 together with the spherical pythagorean theorem Theorem 32.2, show that the side length a of a right triangle can be computed knowing only the opposite angle α and the adjacent angle β as

$$\cos a = \frac{\cos\alpha}{\sin\beta}$$

Hint: start with the formula for $\cos\alpha$*. Write out the tangents in terms of sines and cosines, then apply the pythagorean theorem to expand a term. Finally, use the definition of* $\sin\beta$ *to regroup some terms.*

Formulas such as this are incredibly useful for calculating the side lengths of polygons, by dividing them into triangles and using facts that are known about their angles.

Exercise 32.64 (Spherical Trigonometry)**.** Use spherical trigonometry to figure out the side lengths of the pentagon you discovered in the first exercise.

Hint: can you further divide the five triangles you used before, into ten right triangles inside the pentagon?

Problem Set XI

Orthographic Map

Exercise 32.65. Can you find the coordinates (x, y) of a point on the map where the vectors $\langle 1, 0\rangle$ and $\langle 0, 1\rangle$ only make a 45-degree angle with one another?

Hint: can you make the problem easier for yourself by restricting x *and* y *to lie on some line, so the problem ends up having one variable instead of two?*

You can see how this would make such a map difficult to use for navigation: it would *look* like the map is telling you to turn 90 degrees but in reality you should only turn half that!

Because having to do computations like these constantly when working with a map is a huge technical headache, mathematicians much prefer *conformal maps*, where the angle you see in the Euclidean plane accurately represents the map-angle, and all of this is unnecessary. (This is why the main map employed by mathematicians, *Stereographic Projection*, is conformal).

Archimedes Map

Read the section of the textbook on Archimedes Map (in the Examples chapter). Give a proof that this mao is *infinitesimal area preserving* following the outline below:

Exercise 32.66.

- Show that at each point $p \in M$ the vectors $\langle 1, 0\rangle$ and $\langle 0, 1\rangle$ are sent by ψ to orthogonal vectors on the sphere.

- Find their lengths on the sphere (ie the map-lengths), and use this data to find the area of the infinitesimal rectangle they form.
- Observe that everything beautifully cancels and the area is still one, even though the square was stretched into a rectangle!

Since the infinitesimal area is unchanged by the map at each point, we can finish Archimedes proof via integration (which I do below, using your exercise)

Theorem 32.5. *The surface area of the unit sphere is the same as that of its Archimedes map: that is, the same as the area of its bounding cylinder.*

Proof. Archimedes map captures every point of the sphere except the north and south poles. Since points have zero area, this omission has no effect on our actual question so we can proceed to calculate with the map M.

$$\text{area}(^2) = \iint_2 dA_2 = \iint_{\psi(M)} dA_2 = \iint_M dA_{\text{map}}$$

But now we know that $dA_{\text{map}} = dA_{\mathbb{E}^2}$, that's what you've proved in the exercise above! So we can sub this out, and then realize the resulting integral is just the *definition* of the Euclidean area of M in the plane:

$$= \iint_M dA_{\mathbb{E}^2} = \text{area}(M)$$

□

Stereographic Projection:

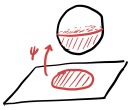
Read the stereographic projection section first (we will cover the necessary bits in class as well)

The stereographic projection map has many uses in mathematics beyond just representing points of the sphere on the plane. Because it is *conformal* (angle preserving), its often used as a tool to help build more interesting conformal maps between regions of the plane, following this general recipe:

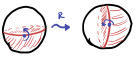
- Start with a region on the plane.
- Use ψ to map it to the sphere.
- Do something to the sphere, moving the region around
- Use ϕ to put it back on the plane.

The overall composition is a map between two regions on the plane, that was created by going to the sphere and back! In these exercises, we will deal with a fundamental example of this, and construct a map from the unit disk onto half of the entire plane!

The strategy above is summarized for this case in the following three figures:



(a) Mapping the unit disk to the lower hemisphere of ^{2} via the parameterization ψ.



(b) Rotating the sphere about the x axis by a quarter turn takes the lower hemisphere to the hemisphere of positive y.

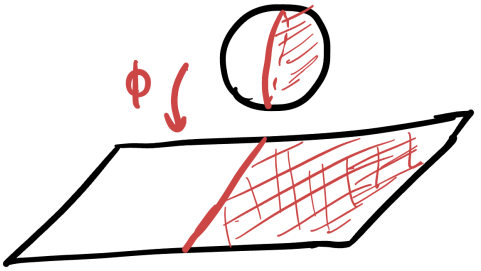


Figure 32.22.: Projecting the hemisphere of positive y to the plane with ϕ gives the half plane with positive y.

Exercise 32.67 (Disk and Half Plane: Construction)**.** Let $\mathbb{D}$ be the unit disk $\mathbb{D} = \{(x, y) \mid x^2 + y^2 < 1\}$ and let $\mathbb{U}$ be the upper half plane $\mathbb{U} = \{(x, y) \mid y > 0\}$. Let $T : \mathbb{D} \to \mathbb{U}$ be the map described above. Prove that $T can be expressed as

$$T(x,y) = \left(\frac{2x}{1+x^2+y^2-2y}, \frac{1-x^2-y^2}{1+x^2+y^2-2y} \right)$$

By building it step by step:

- Start with (x, y) in the unit disk.
- Apply ψ to get the disk onto the sphere.
- Rotate the sphere about the x axis in the appropriate way so that the south pole goes to $(0, 1, 0)$.
- Apply ϕ to return to the plane.

This map is *conformal* - meaning that it preserves all angles! And even more than that, it takes *generalized circles to generalized circles.*

Exercise 32.68 (Disk and Half Plane: Understanding)**.** Prove that these claims are in fact true: that our new function is conformal, and sends generalized circles to generalized circles. *Hint: what kinds of maps is it built out of? What do each of these maps to do angles, or to generalized circles (on the plane) / circles (on the sphere)?*

Use this to "transfer" this picture of polar coordinates in the unit disk onto the plane, via our new map.

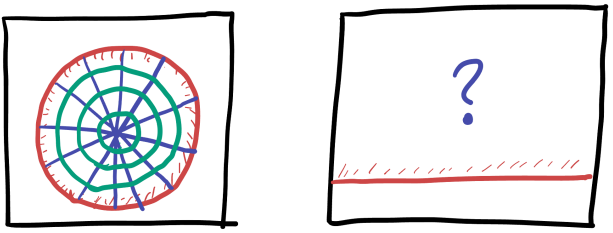


Figure 32.23.: What do these generalized circles look like when mapped to the half plane?

Spheres of Radius R:

The chapter on stereographic projection deals with the *unit sphere.* It is not too hard to generalize what we have done to spheres of other radii, and **while this may not sound super exciting at first, it actually turns out to be absolutely fundamental to how we are going to discover hyperbolic space!** So, it is a rather important exercise to work this all out for yourself.

The good news is you have this entire chapter as a guide, where I've worked out many of the details for the case of the unit sphere. The formulas will be quite similar, but

there'll be R's inserted in various places: so the second piece of good news is that I'll give you the formulas that you need to derive! That way, you can check your work.

Definition 32.6. Let $\overset{2}{R}$ be the sphere of radius R in $\mathbb{E}^3$. Then the chart ϕ for stereographic projection of this sphere is defined geometrically exactly as in the original version: given a point $p \in \overset{2}{R}$, $\phi(p)$ is where the line connecting p to the north pole $N = (0, 0, R)$ intersects the xy plane.

Exercise 32.69. Show that the formulas for both the chart and the parameterization of stereographic projection here are as follows:

$$\phi(x, y, z) = (X, Y) = \left(\frac{Rx}{R - z}, \frac{Ry}{R - z}\right)$$

$$\psi(X, Y) = (x, y, z) = \left(\frac{2R^2X}{X^2 + Y^2 + R^2}, \frac{2R^2Y}{X^2 + Y^2 + R^2}, R\frac{X^2 + Y^2 - R^2}{X^2 + Y^2 + R^2}\right)$$

(It might help to look back at Proposition 24.1, and attempt Exercise 24.1).

Running through the same arguments as in the chapter above (which you don't have to write down), its straightforward to check that this new map is a conformal map between $\overset{2}{R}$ minus N, and the plane. This means its parameterization ψ both preserves angles and stretches all vectors by a uniform length: we can use this fact to compute the dot product for this map.

Exercise 32.70. At a point $p = (X, Y)$ on the plane, what is the factor by which a vector $v \in T_p\mathbb{E}^2$ is stretched when mapped onto $\overset{2}{R}$ by the parameterization of stereographic projection? *Hint: we know the factor is the same for all vectors: so pick an easy vector to calculate with and find its length*!

Once you know this, follow the argument style of Theorem 24.3 to compute the map-dot product on the plane, and show that it is equal to

$$(v \cdot w)_{\text{map}} = \frac{4R^4}{(R^2 + X^2 + Y^2)^2}(v \cdot w)$$

Problem Set XII

Getting Used to Hyperbolic Space

Exercise 32.71 (Hyperbolic Circle Area)**.** In this exercise you will go through the transition-style arguments we use to turn formulas on the sphere into their analogs in hyperbolic geometry, much like we did in class.

The starting point is the formula for the area of a circle of radius r on the sphere of radius R:

$$A(r) = \int_0^r C(r)dr = \int_0^r 2\pi \sin\left(\frac{r}{R}\right) dr = 2\pi R^2 \left(1 - \cos\frac{r}{R}\right)$$

- Re-express this in terms of curvature
- Convert to the Taylor series
- Plug in $\kappa = -1$
- Convert back to hyperbolic trigonometric functions

What is the correct hyperbolic formula? (We wrote it down without doing the full derivation in class, so you can confirm)

Exercise 32.72 (Hyperbolic Pizza)**.** One way to try and develop intuition for the strange behavior of circles is to think about the type of circles we see in daily life: pizzas! One major factor determining how good a pizza is is its *crust percentage* which we will define as

$$\text{CrustPercent} = \frac{\text{area(Crust)}}{\text{area(Pizza)}}$$

In this probelm we will consider pizzas which have 1 inch crusts: meaning a 10 inch (radius) pizza has a 9inch radius center of toppings, surrounded by a 1 inch thick circle of crust.

- Show the CrustPercent for Euclidean pizza is

  $$\frac{2}{r} - \frac{1}{r^2}.$$

  From this we see that as $r \to \infty$ the crust percent drops to zero: this makes sense, if you imagine an extremely large pizza with only a 1inch thick crust, it's totally reasonable that *most of the pizza is not crust*!
- What is the CrustPercent for a hyperbolic pizza of radius r? Show that when r is large, this limits to the constant

  $$\text{CrustPercent} \to 1 - \frac{1}{e} \approx 63\%$$

  Thus crust is an inevitable part of life in hyperbolic space: no matter what size pizza you make it will always be well over half crust!

Exercise 32.73 (Hyperbolic Pizza II)**.** In this problem, we will imagine our unit to be inches (so, the radius appearing in formulas for space of curvature -1 is measured in inches).

You are at a pizzerias and are trying to decide if the 5 inch radius pizza they sell is large enough for you and your friends. They also sell a six inch (radius) pizza, but it

costs twice as much. At first, you think this sounds crazy! But is this actually a good deal, or not?

Your friend is feeling very hungry, and jokingly asks the pizza-maker how large of a pizza he would need to order so that its areas is the same as an american football field (100×50 yards). The pizza-maker says "I think I have room for that in my oven, coming right up!" How big of a pizza is he going to make?

Hint: invert the formula for area in terms of radius, to get radius in terms of area, then plug into a calculator!

Working with the Models

Exercise 32.74 (Hyperbolic space is homogeneous)**.** We proved in class that the horizontal translations $T(x, y) = (x + t, y)$ are isometries of the Half Plane model, and we also proved that the similarities $S(x, y) = (sx, sy)$ are also isometries.

Combining these two, prove that $\mathbb{H}^2$ is homogeneous: that is, that for any two points $p, q \in \mathbb{H}^2$, there exists an isometry that takes p to q.

Hint: can you first show that you can build an isometry that takes $(0, 1)$ *to any point in the plane? Then combine two of these to get what you want?*

Exercise 32.75 (The Circumference of Circles)**.** In the Disk Model, if $a < 1$ the Euclidean circle $x^2 + y^2 = a^2$ centered at O is also a hyperbolic circle. In the text, we compute that its hyperbolic circumference is

$$C = \frac{4\pi a}{1 - a^2}$$

and that its hyperbolic radius is

$$r = 2\text{arctanh}(a)$$

Using these two, prove that $C(r) = 2\pi \sinh(r)$. *Hint: solve for a in terms of r and plug into circumference. Then use hyperbolic trigonometric identities to simplify!*

Together these two arguments prove that the geometry modeled by the Half Plane and the Disk has the circumference function $C(r) = 2\pi \sinh(r)$ for circles based at any point (the second problem establishes this for circles at a special point, and hte first problem establishes that space is homogeneous, so its the same at all points). Thus, this space truly *is* hyperbolic geometry, and has curvature -1 (we proved in class, any space with this circumference function has constant curvature -1).