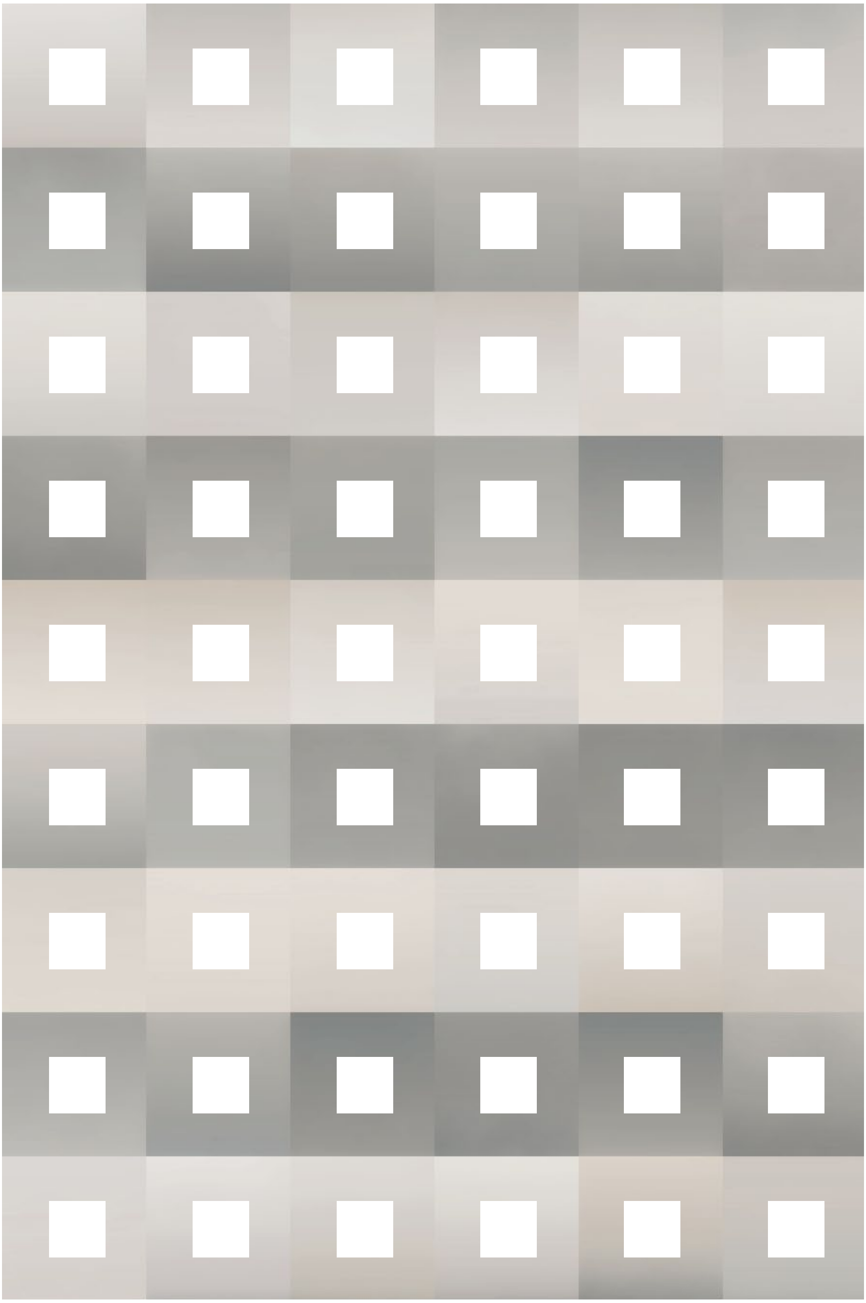
3

EUCLIDEAN TRANSFORMATIONS

In the third part of this book, we will look at Euclidean geometry from a different perspective, that of Euclidean transformations. It is a point of view that has been most closely associated with Felix Klein– that the way to study some property (such as congruence) is to study the maps that preserve it. The first lesson sets the scene with a quick development of analytic geometry. Then it is on to Euclidean isometries– bijections of the Euclidean plane which preserve distance. Over several lessons we will study these isometries, and ultimately we will classify all Euclidean isometries into four types: reflections, rotations, translations, and glide reflections. Then it is time to loosen the restriction a bit to consider bijections which preserve congruence, but not necessarily distance. Finally, we will look at inversion, a type of bijection of the punctured plane (the Euclidean plane minus a point). As luck would have it, inversion provides a convenient bridge into non-Euclidean geometry.

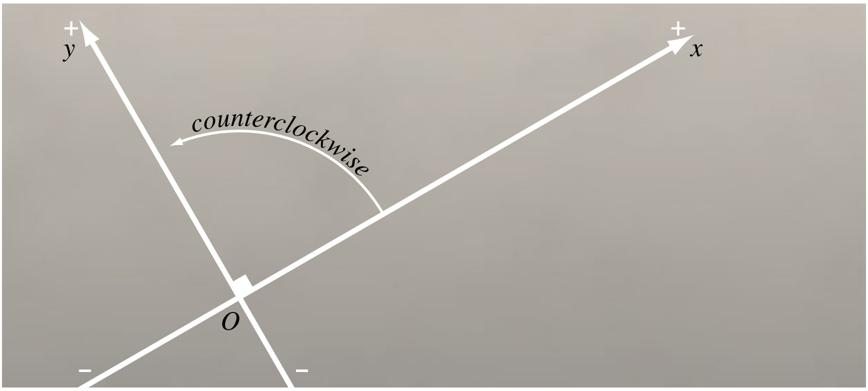


23 BACK ON THE GRID
ANALYTIC GEOMETRY

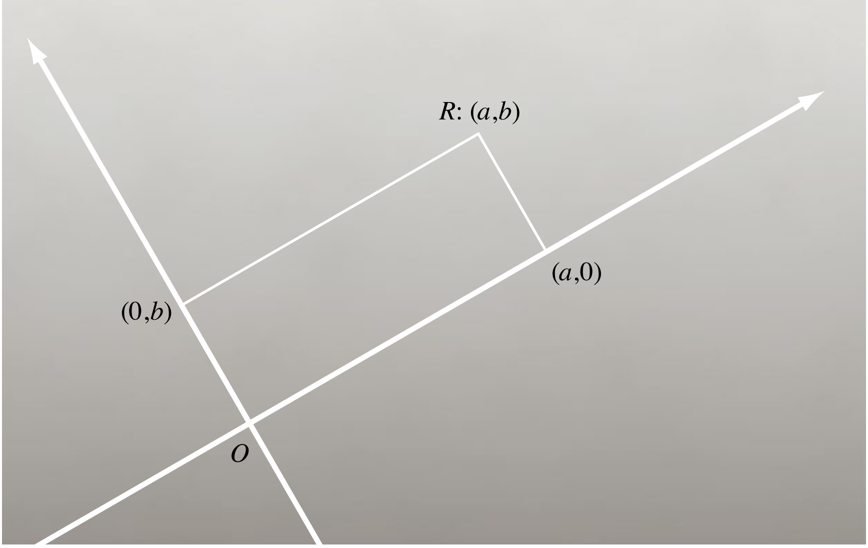
This lesson is just a quick development of analytic geometry and trigonometry in the language of Euclidean geometry. I feel an obligation to provide the connection between traditional Euclidean geometry (as I have developed it in these lessons) and more contemporary analytic geometry, but you should already be comfortable with this material, so feel free to skim through it.

Analytic geometry

At the heart of analytic geometry, there is a correspondence between points and coordinates, ordered pairs of real numbers. The Cartesian approach to that correspondence is a familiar one, but let me quickly run through it. Begin with two perpendicular lines (the choice is arbitrary). These are the x- and y-axes. Their intersection is the origin O. We will want to measure signed distances from O along these axes, and that means we have to assign a positive direction to each axis. From a geometric point of view, the choice of those directions is arbitrary, but there is an established convention as follows. Once directions have been chosen, each axis will be divided into two rays that share O as their common vertex: a positive axis consisting of points whose signed distance from O is positive, and a negative axis consisting of points whose signed distance from O is negative. The convention is that the axes are assigned positive directions so that the positive y-axis is a 90° *counterclockwise* turn from the positive x-axis. Now here's the catch: the geometry itself provides no way to distinguish which direction is the counterclockwise direction. So this is a convention that must be passed along by way of illustrations (and clocks).



A point P on the x-axis is assigned the coordinates $(p,0)$, where p is the signed distance from O to P. A point Q on the y-axis is assigned the coordinates $(0,q)$ where q is the signed distance from O to Q. Most points will not lie on either axis. For these points, we must consider their projections onto the axes. If R is such a point, then we draw the two lines that pass through R and are perpendicular to the two axes. If the points where these perpendiculars cross the axes have coordinates $(a,0)$ and $(0,b)$, then the coordinates of R are (a,b). With this correspondence, every point corresponds to a unique coordinate pair, and every coordinate pair corresponds to a unique point.



The next step is to figure out how to calculate the distance between points in terms of their coordinates. This is pretty much essential for everything else that we are going to do. Let's begin with two special cases.

LEM: VERTICAL DISTANCE

For points that share an x-coordinate, $P_1 = (x,y_1)$ and $P_2 = (x,y_2)$,

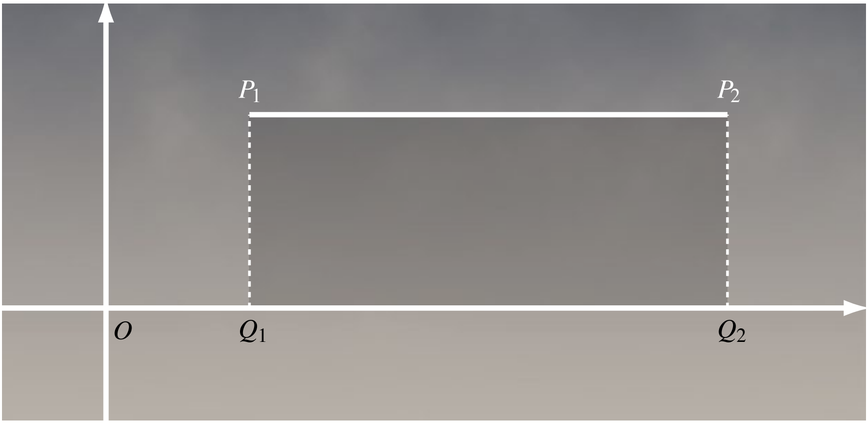
$$|P_1P_2| = |y_1 - y_2|.$$

HORIZONTAL DISTANCE

For points that share a y-coordinate, $P_3 = (x_3,y)$ and $P_4 = (x_4,y)$,

$$|P_3P_4| = |x_3 - x_4|.$$

Proof. I will just prove the first statement. Label two more points, $Q_1 = (0, y_1)$ and $Q_2 = (0, y_2)$. The resulting quadrilateral $P_1P_2Q_2Q_1$ is a rectangle, so its opposite sides P_1P_2 and Q_1Q_2 have to be the same length.



This is where we make the direct connection between coordinates and distance– the coordinates along each axis were chosen to reflect their signed distance from the origin O. To be thorough, though, there are several cases to consider:

$$\begin{aligned}
O*Q_1*Q_2: &\quad |Q_1Q_2| = |OQ_2| - |OQ_1| = y_2 - y_1 = |y_1 - y_2|\\
O*Q_2*Q_1: &\quad |Q_1Q_2| = |OQ_1| - |OQ_2| = y_1 - y_2 = |y_1 - y_2|\\
Q_1*O*Q_2: &\quad |Q_1Q_2| = |OQ_1| + |OQ_2| = -y_1 + y_2 = |y_1 - y_2|\\
Q_2*O*Q_1: &\quad |Q_1Q_2| = |OQ_2| + |OQ_1| = -y_2 + y_1 = |y_1 - y_2|\\
Q_1*Q_2*O: &\quad |Q_1Q_2| = |OQ_1| - |OQ_2| = -y_1 - (-y_2) = |y_1 - y_2|\\
Q_2*Q_1*O: &\quad |Q_1Q_2| = |OQ_2| - |OQ_1| = -y_2 - (-y_1) = |y_1 - y_2|
\end{aligned}$$

No matter the case, $|P_1P_2| = |Q_1Q_2| = |y_1 - y_2|$. □

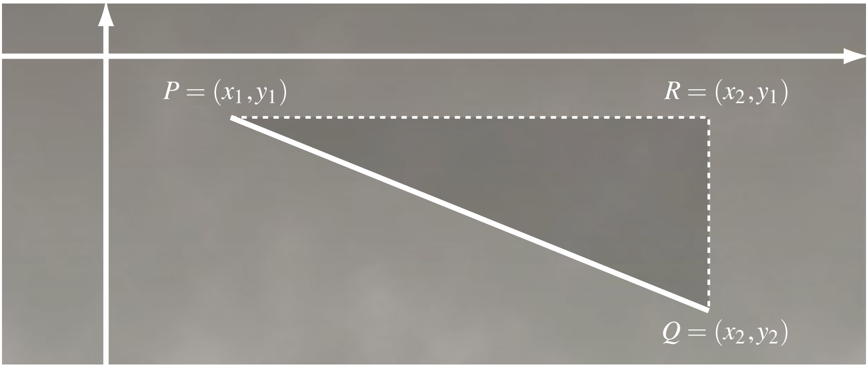
The general distance formula is now an easy consequence of the Pythagorean Theorem.

THM: THE DISTANCE FORMULA

For any two points $P=(x_1,y_1)$ and $Q=(x_2,y_2)$,

$$|PQ| = \sqrt{(x_1-x_2)^2+(y_1-y_2)^2}.$$

Proof. If P and Q share either x-coordinates or y-coordinates, then this formula reduces down to the special case in the previous lemma (because $\sqrt{a^2}=|a|$). If not, mark one more point: $R=(x_2,y_1)$.



Then $|PR|=|x_1-x_2|$, and $|RQ|=|y_1-y_2|$, and $\triangle PRQ$ is a right triangle. By the Pythagorean theorem,

$$|PQ|^2 = |PR|^2+|QR|^2 = (x_1-x_2)^2+(y_1-y_2)^2$$

Now take the square root to get the formula. □

COR: THE EQUATION OF A CIRCLE

The equation of a circle C with center at $P=(h,k)$ and radius r is

$$(x-h)^2+(y-k)^2=r^2.$$

Proof. By definition, the points of C are all those points that are a distance of r from P. Therefore (x,y) is on C if and only if

$$\sqrt{(x-h)^2+(y-k)^2}=r.$$

Square both sides of the equation to get the standard form. □

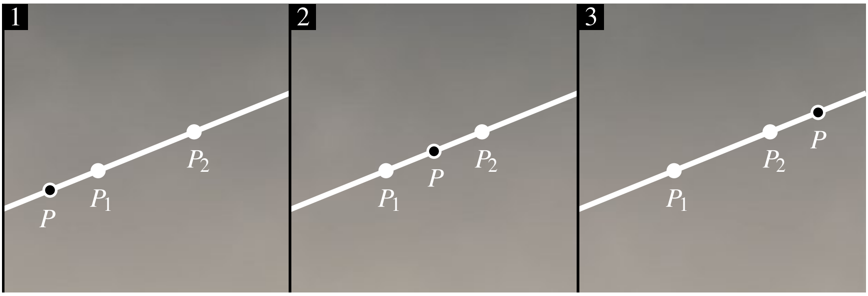
Moving along, lines are next. Intuitively, the key is the idea that a line describes the shortest path between points. That is captured more formally in the triangle inequality, which you should recall states that $|AB|+|BC| \geq |AC|$, but that the equality only happens when $A*B*C$.

PARAMETRIC FORM FOR THE EQUATION OF A LINE
Given two distinct points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ on a line ℓ, a third point $P = (x, y)$ lies on ℓ if and only if its coordinates can be written in the form

$$x = x_1 + t(x_2 - x_1) \quad \& \quad y = y_1 + t(y_2 - y_1)$$

for some $t \in \mathbb{R}$.

Proof. The different possible orderings of P, P_1, and P_2 on the line create several scenarios



Let me just take the middle case, where t is between 0 and 1 and P is between P_1 and P_2. It is representative of the other two cases.

$\Longrightarrow$ Show that if $P = (x_1 + t(x_2 - x_1), y_1 + t(y_2 - y_1))$ for some value of t between 0 and 1, then P is between P_1 and P_2.

We can directly calculate $|P_1P|$ and $|PP_2|$:

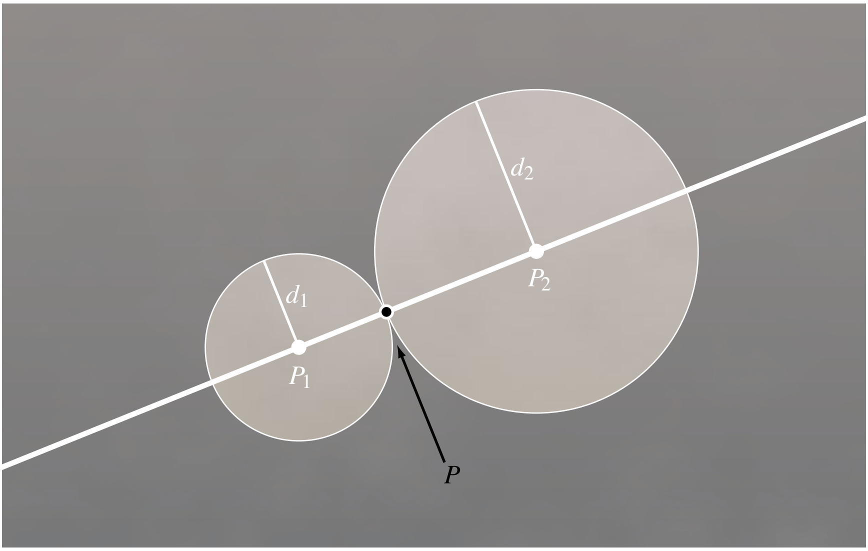
$$\begin{aligned}|P_1P| &= [(x-x_1)^2 + (y-y_1)^2]^{1/2}\\ &= [(x_1 + t(x_2-x_1) - x_1)^2 + (y_1 + t(y_2-y_1) - y_1)^2]^{1/2}\\ &= [(tx_2 - tx_1)^2 + (ty_2 - ty_1)^2]^{1/2}\\ &= t[(x_2-x_1)^2 + (y_2-y_1)^2]^{1/2}\\ &= t|P_1P_2|.\end{aligned}$$

$$\begin{aligned}|PP_2| &= [(x_2-x)^2+(y_2-y)^2]^{1/2}\\ &= [(x_2-(x_1+t(x_2-x_1)))^2+(y_2-(y_1+t(y_2-y_1)))^2]^{1/2}\\ &= [((1-t)x_2-(1-t)x_1)^2+((1-t)y_2-(1-t)y_1)^2]^{1/2}\\ &= (1-t)[(x_2-x_1)^2+(y_2-y_1)^2]^{1/2}\\ &= (1-t)|P_1P_2|.\end{aligned}$$

According to the Triangle Inequality, then, P is between P_1 and P_2, since

$$|P_1P|+|PP_2| = t|P_1P_2|+(1-t)|P_1P_2| = |P_1P_2|.$$

$\Longleftarrow$ Show that if P is between P_1 and P_2, then the coordinates of P can be written in the parametric form $(x_1+t(x_2-x_1), y_1+t(y_2-y_1))$ for some value of t between 0 and 1.



Point P is the only point in the plane which is a distance $d_1 = |P_1P|$ from P_1 *and* a distance $d_2 = |PP_2|$ from P_2. Because of that uniqueness, we just need to find a point in parametric form that is also those respective distances from P_1 and P_2. The point that we are looking for is the one where $t = d_1/(d_1+d_2)$. The two calculations, that the distance from this point to P_1 is d_1, and that the distance from this point to P_2 is d_2, are both straightforward, so I will leave them to you. □

From the parametric form it is easy to get to standard form, and from there to point-slope form, slope-intercept form, and so on. The latter steps are standard fare for a pre-calculus course, so I will only go one step further.

STANDARD FORM FOR THE EQUATION OF A LINE
The coordinates (x,y) of the points of a line all satisfy an equation of the form $Ax+By=C$ where A, B, and C are real numbers.

Proof. Suppose that (x_1,y_1) and (x_2,y_2) are distinct points on the line. As we saw in the last theorem, the other points on the line have coordinates (x,y) that satisfy the equations

$$\begin{cases} x = x_1 + t(x_2 - x_1) \\ y = y_1 + t(y_2 - y_1). \end{cases}$$

Now it is just a matter of combining the equations to eliminate the parameter t.

$$\begin{cases} x - x_1 = t(x_2 - x_1) \\ y - y_1 = t(y_2 - y_1). \end{cases}$$

At this point, you could divide the second equation by the first. That eliminates the t variable and also serves as a definition of the slope of a line (in particular, it shows that the slope is constant). But it also presents a potential "divide by zero" scenario, so instead let's multiply:

$$\begin{cases} (x - x_1)(y_2 - y_1) = t(x_2 - x_1)(y_2 - y_1) \\ (y - y_1)(x_2 - x_1) = t(y_2 - y_1)(x_2 - x_1). \end{cases}$$

Set the two equations equal and simplify

$$\begin{aligned} (x - x_1)(y_2 - y_1) &= (y - y_1)(x_2 - x_1) \\ x(y_2 - y_1) - x_1(y_2 - y_1) &= y(x_2 - x_1) - y_1(x_2 - x_1) \\ x(y_2 - y_1) - y(x_2 - x_1) &= x_1(y_2 - y_1) - y_1(x_2 - x_1). \end{aligned}$$

This equation has the proper form, with

$$A = y_2 - y_1 \quad B = -(x_2 - x_1) \quad \& \quad C = x_1(y_2 - y_1) - y_1(x_2 - x_1).$$

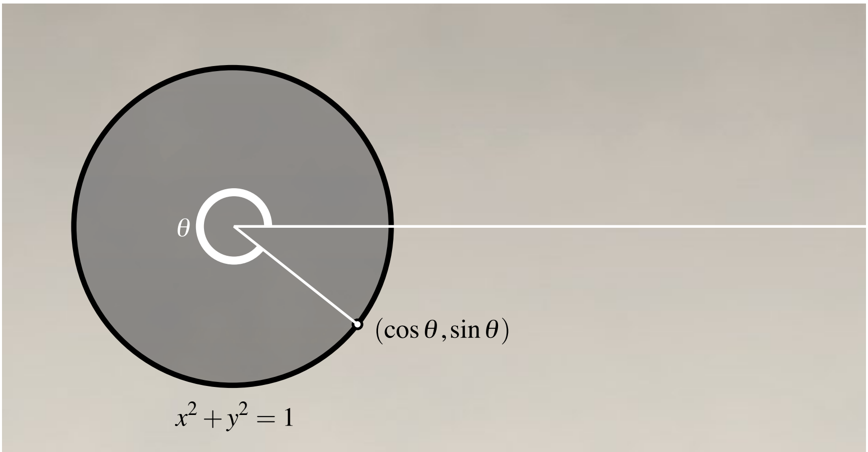
□

Finally, it should be noted that any three real numbers A,B,C do describe a line, so long as A and B are not both zero.

The unit circle approach to trigonometry

At the end of the lesson on similarity, in the exercises, we defined the six trigonometric functions. At that time, we defined them in terms of the angles of a right triangle, which means that they were restricted to values in the interval $(0,\pi/2)$. As you know, there is also a "unit circle approach" that extends these definitions beyond that narrow window. You have seen this before, so I will be as brief as I can be. A point with two *positive* coordinates (x,y) on the unit circle corresponds to a right triangle whose vertices are $(0,0)$, $(x,0)$ and (x,y). If θ is the measure of the angle at the origin, then $\cos\theta = x$ and $\sin\theta = y$ (because the hypotenuse has length one). Now just continue that: any ray from the origin forms an angle θ measured in the counterclockwise direction from the x-axis. That ray intersects the unit circle at a point (x,y) and we define

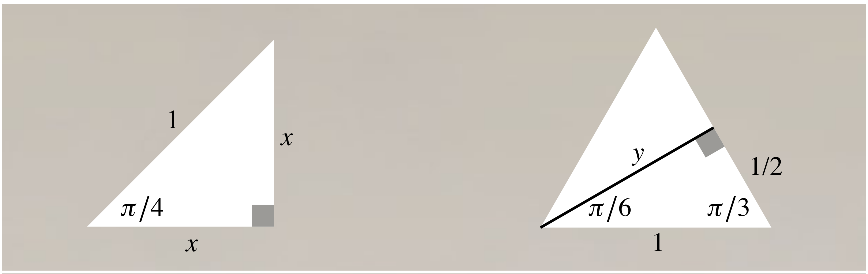
$$\cos(\theta) = x \quad \sin(\theta) = y.$$



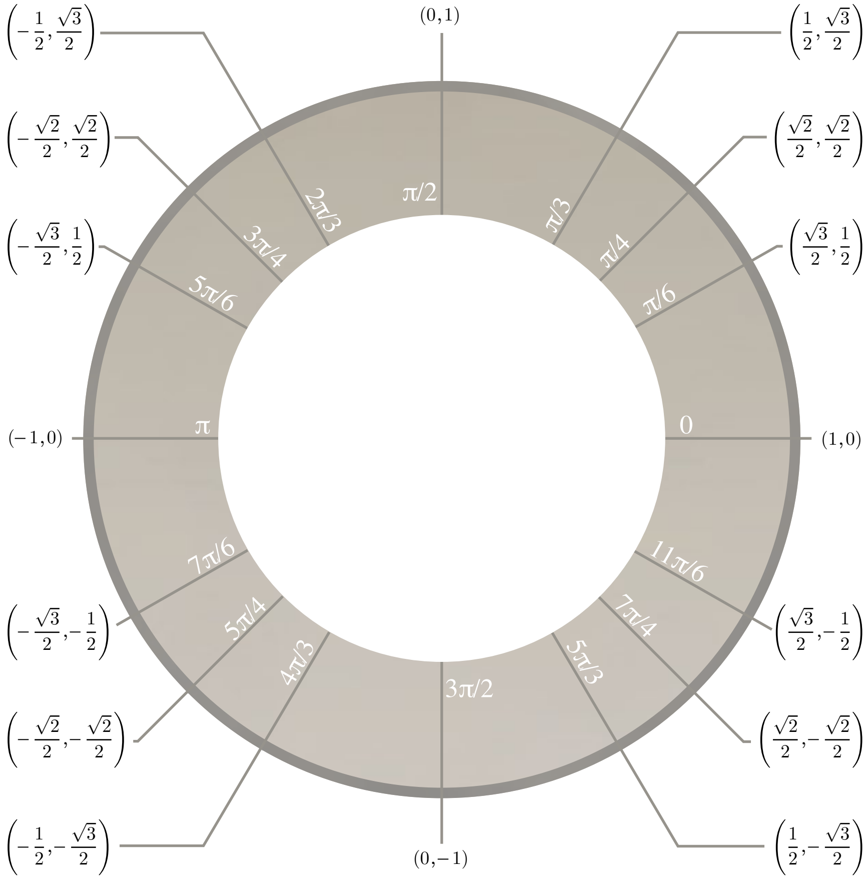
Allowing for both proper and reflex angles, that extends the domains of sine and cosine to $[0,2\pi)$, but we can go farther. Informally, we need to allow the ray to spin around the circle more than once (for θ values greater than 2π) or in the counterclockwise direction (for negative θ). Formally, this can be done by imposing periodicity:

$$\cos(\theta+2n\pi) = \cos(\theta) \quad \sin(\theta+2n\pi) = \sin(\theta) \quad \forall n \in \mathbb{N}.$$

Use an isosceles and equilateral triangle to find sine and cosine values for π/3, π/4, and π/6. Use the symmetry of the circle to extend outside of quadrant I.



$$x^2+x^2=1 \implies x=\sqrt{2}/2 \qquad (1/2)^2+y^2=1 \implies y=\sqrt{3}/2$$


$\left(-\frac{1}{2},\frac{\sqrt{3}}{2}\right)$
$(0,1)$
$\left(\frac{1}{2},\frac{\sqrt{3}}{2}\right)$
$\left(-\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right)$
$\left(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right)$
$\left(-\frac{\sqrt{3}}{2},\frac{1}{2}\right)$
$\left(\frac{\sqrt{3}}{2},\frac{1}{2}\right)$
$\pi/2$
$2\pi/3$
$\pi/3$
$3\pi/4$
$\pi/4$
$5\pi/6$
$\pi/6$
$(-1,0)$
π
0
$(1,0)$
$7\pi/6$
$11\pi/6$
$5\pi/4$
$7\pi/4$
$4\pi/3$
$5\pi/3$
$3\pi/2$
$\left(-\frac{\sqrt{3}}{2},-\frac{1}{2}\right)$
$\left(\frac{\sqrt{3}}{2},-\frac{1}{2}\right)$
$\left(-\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2}\right)$
$\left(\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2}\right)$
$\left(-\frac{1}{2},-\frac{\sqrt{3}}{2}\right)$
$(0,-1)$
$\left(\frac{1}{2},-\frac{\sqrt{3}}{2}\right)$

The other four trigonometric functions (tangent, cotangent, secant, cosecant) are defined similarly as the ratios

$$\tan(\theta) = y/x \quad \cot(\theta) = x/y \quad \sec(\theta) = 1/x \quad \csc(\theta) = 1/y.$$

There are a lot of relationships between the trigonometric functions, some easy and some subtle. Let's get the easy ones out of the way. From the very definitions of the functions, we get the reciprocal identities

$$\sec\theta = \frac{1}{\cos\theta} \quad \csc\theta = \frac{1}{\sin\theta} \quad \cot\theta = \frac{1}{\tan\theta},$$

and identities that relate tangent and cotangent to sine and cosine

$$\tan\theta = \frac{\sin\theta}{\cos\theta} \quad \cot\theta = \frac{\cos\theta}{\sin\theta}.$$

From the equation of the circle $x^2 + y^2 = 1$, we get the Pythagorean identities:

$$\sin^2\theta + \cos^2\theta = 1 \quad \tan^2\theta + 1 = \sec^2\theta \quad 1 + \cot^2\theta = \csc^2\theta.$$

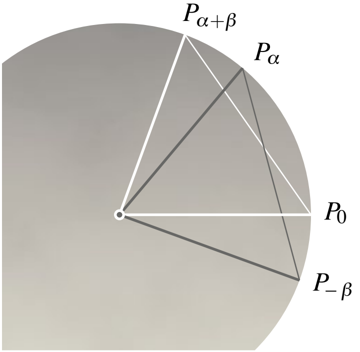
By comparing angles taken in the counterclockwise and clockwise directions, we see that cosine and secant are even functions (where $f(-x) = f(x)$) and that the other four are odd functions (where $f(-x) = -f(x)$).

Beyond these, there is a second tier of identities– double angle, half angle, power reduction, etc – that are not so immediately clear. They can all be derived from two big identities, the addition formulas for sine and cosine, but the proofs of those two formulas require a more careful look at the geometry of the unit circle. To close out this lesson, I will prove the two addition formulas.

ADDITION RULE FOR COSINE

$$\cos(\alpha + \beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$$

Proof. The key to the proof is to compare two distances which we know to be the same– one distance expressed in terms of the angle $\alpha + \beta$, the other in terms of the individual angles α and β. The real trick to this is to make the right choice of distances. In particular, you have to be careful so that you don't get stuck with a $\sin(\alpha+\beta)$ term in the first calculation. On the unit circle, label the following points:



$$\begin{aligned}
P_0 &= (1,0)\\
P_\alpha &= (\cos\alpha, \sin\alpha)\\
P_{-\beta} &= (\cos(-\beta), \sin(-\beta))\\
&= (\cos\beta, -\sin\beta)\\
P_{\alpha+\beta} &= (\cos(\alpha+\beta), \sin(\alpha+\beta))
\end{aligned}$$

If O is the origin, then the triangles $\triangle OP_0P_{\alpha+\beta}$ and $\triangle OP_{-\beta}P_\alpha$ are congruent (S·A·S: in each triangle, two of the sides are radii, and the angle between them measures $\alpha+\beta$). That means that the two segments $P_0P_{\alpha+\beta}$ and $P_{-\beta}P_\alpha$ have to be congruent, and so we can compare their lengths (it is actually easier to work with the squares of those lengths). Throughout these calculations, we make repeated use of the Pythagorean Identity $\sin^2 x + \cos^2 x = 1$.

$$\begin{aligned}
|P_0P_{\alpha+\beta}|^2 &= (\cos(\alpha+\beta)-1)^2 + (\sin(\alpha+\beta)-0)^2\\
&= \cos(\alpha+\beta)^2 - 2\cos(\alpha+\beta) + 1 + \sin^2(\alpha+\beta)\\
&= 2 - 2\cos(\alpha+\beta).
\end{aligned}$$

$$\begin{aligned}
|P_{-\beta}P_\alpha|^2 &= (\cos\alpha - \cos\beta)^2 + (\sin\alpha + \sin\beta)^2\\
&= \cos^2\alpha - 2\cos\alpha\cos\beta + \cos^2\beta\\
&\qquad + \sin^2\alpha + 2\sin\alpha\sin\beta + \sin^2\beta\\
&= 2 - 2\cos\alpha\cos\beta + 2\sin\alpha\sin\beta.
\end{aligned}$$

Set these two expressions equal to each other, subtract 2 and divide by -2 to get the desired formula

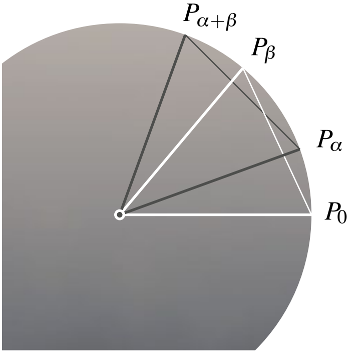
$$\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta.$$

□

ADDITION RULE FOR SINE

$$\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta$$

Proof. For this proof, one approach would be to use the cofunction identity $\sin(x) = \cos(\pi/2 - x)$ followed by the addition rule for cosine that we just derived. That is pretty easy, but you would have to verify the cofunction identity first. That too is easy for x values between 0 and $\pi/2$, but gets to be a nuisance once you have to consider all the other possible values of x. I think it is easier to do something like the last proof– compare some distances and then do a little algebra. On the unit circle, label the following points



$$\begin{aligned}
P_0 &= (1,0)\\
P_\alpha &= (\cos\alpha, \sin\alpha)\\
P_\beta &= (\cos\beta, \sin\beta)\\
P_{\alpha+\beta} &= (\cos(\alpha+\beta), \sin(\alpha+\beta)).
\end{aligned}$$

By S·A·S, the segments $P_\alpha P_{\alpha+\beta}$ and $P_0 P_\beta$ are congruent. Let's compare those two distances. Here we go (note the use of the addition rule for cosine midway through the first distance calculation).

$$\begin{aligned}
|P_\alpha P_{\alpha+\beta}|^2 &= (\cos(\alpha+\beta) - \cos(\alpha))^2 + (\sin(\alpha+\beta) - \sin(\alpha))^2\\
&= \cos^2(\alpha+\beta) - 2\cos\alpha\cos(\alpha+\beta) + \cos^2\alpha\\
&\qquad + \sin^2(\alpha+\beta) - 2\sin\alpha\sin(\alpha+\beta) + \sin^2\alpha\\
&= 2 - 2\cos\alpha\cos(\alpha+\beta) - 2\sin\alpha\sin(\alpha+\beta)\\
&= 2 - 2\cos\alpha(\cos\alpha\cos\beta - \sin\alpha\sin\beta) - 2\sin\alpha\sin(\alpha+\beta)\\
&= 2 - 2\cos^2\alpha\cos\beta + 2\sin\alpha\cos\alpha\sin\beta - 2\sin\alpha\sin(\alpha+\beta)
\end{aligned}$$

and

$$\begin{aligned}|P_0P_\beta|^2 &= (\cos\beta - 1)^2 + (\sin\beta - 0)^2 \\ &= \cos^2\beta - 2\cos\beta + 1 + \sin^2\beta \\ &= 2 - 2\cos\beta.\end{aligned}$$

Now set these two expressions equal, subtract 2 from both sides and divide through by -2 to get

$$\cos^2\alpha\cos\beta - \sin\alpha\cos\alpha\sin\beta + \sin\alpha\sin(\alpha+\beta) = \cos\beta.$$

In this equation solve for the $\sin(\alpha+\beta)$ term

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

As long as $\sin\alpha$ is not zero, we can divide both sides by that, and what's left over is what we want. What if $\sin\alpha$ is zero? Well, that happens when α is any multiple of π, and those cases are easy enough to handle on their own. On the left side, adding $n\pi$ corresponds to a half-turn or a whole turn around the unit circle, so

$$\sin(n\pi+\beta) = \begin{cases}\sin\beta & \text{if } n \text{ is even} \\ -\sin\beta & \text{if } n \text{ is odd.}\end{cases}$$

Compare that to the right side

$$\begin{aligned}\sin(n\pi)\cos\beta + \cos(n\pi)\sin\beta &= 0\cdot\cos\beta + \cos(n\pi)\sin\beta \\ &= \begin{cases}\sin\beta & \text{if } n \text{ is even} \\ -\sin\beta & \text{if } n \text{ is odd}\end{cases}\end{aligned}$$

They are the same. □

Exercises

1. Prove the midpoint formula. Let $P = (a,b)$ and $Q = (c,d)$. Verify that the coordinates of the midpoint of PQ are

$$\left(\frac{a+c}{2}, \frac{b+d}{2}\right).$$

2. Show that the points on the circle with center (h,k) and radius r can be described by the parametric equations

$$\begin{cases} x(\theta) = h + r\cos\theta \\ y(\theta) = k + r\sin\theta \end{cases}.$$

3. Let ℓ_1 and ℓ_2 be perpendicular lines, neither of which is a vertical line. Show that the slopes of ℓ_1 and ℓ_2 are negative reciprocals of one another.

4. Verify that the triangle with vertices at $(0,0)$, $(2a,0)$, and $(a, a\sqrt{3})$ is equilateral.

5. Find the equation of the circle which passes through the three points: $(0,0)$, $(4,2)$ and $(2,6)$.

6. Let $\triangle ABC$ be the triangle with vertices at the coordinates $A = (0,0)$, $B = (1,0)$, $C = (a,b)$. Find the coordinates of its circumcenter, orthocenter, and centroid (in terms of a and b).

7. All of the special values on the unit circle can be written in the form $n\pi/12$, but not all values of that form are represented. Find the coordinates on the unit circle for the angles $\theta = \pi/12, 5\pi/12, 7\pi/12$, and $11\pi/12$.

The remaining exercises verify some common trigonometric identities that we will need to for later calculations. You don't need to do them all– I really just want to have all of these identities together in one place.

8. Use the addition formulas to derive the cofunction identities.

$$\sin\left(\frac{\pi}{2}-\theta\right)=\cos\theta \qquad \cos\left(\frac{\pi}{2}-\theta\right)=\sin\theta$$

$$\tan\left(\frac{\pi}{2}-\theta\right)=\cot\theta \qquad \cot\left(\frac{\pi}{2}-\theta\right)=\tan\theta$$

$$\sec\left(\frac{\pi}{2}-\theta\right)=\csc\theta \qquad \csc\left(\frac{\pi}{2}-\theta\right)=\sec\theta$$

9. Use the addition formulas to derive the double angle formulas

$$\sin(2\theta)=2\sin\theta\cos\theta$$

$$\begin{aligned}\cos(2\theta)&=\cos^2\theta-\sin^2\theta\\&=2\cos^2\theta-1\\&=1-2\sin^2\theta\end{aligned}$$

$$\tan(2\theta)=\frac{2\tan\theta}{1-\tan^2\theta}$$

10. Use the double angle formulas for cosine to derive the power-reduction formulas

$$\sin^2\theta=\frac{1-\cos(2\theta)}{2}$$

$$\cos^2\theta=\frac{1+\cos(2\theta)}{2}$$

$$\tan^2\theta=\frac{1-\cos(2\theta)}{1+\cos(2\theta)}$$

11. Use the power-reduction formulas to derive the half-angle formulas

$$\sin\frac{\theta}{2}=\pm\sqrt{\frac{1-\cos\theta}{2}}$$

$$\cos\frac{\theta}{2}=\pm\sqrt{\frac{1+\cos\theta}{2}}$$

$$\tan\frac{\theta}{2}=\frac{1-\cos\theta}{\sin\theta}=\frac{\sin\theta}{1+\cos\theta}$$

12. Verify the product-to-sum formulas

$$\sin\alpha\sin\beta = \frac{1}{2}[\cos(\alpha-\beta)-\cos(\alpha+\beta)]$$

$$\cos\alpha\cos\beta = \frac{1}{2}[\cos(\alpha+\beta)+\cos(\alpha-\beta)]$$

$$\sin\alpha\cos\beta = \frac{1}{2}[\sin(\alpha+\beta)+\sin(\alpha-\beta)]$$

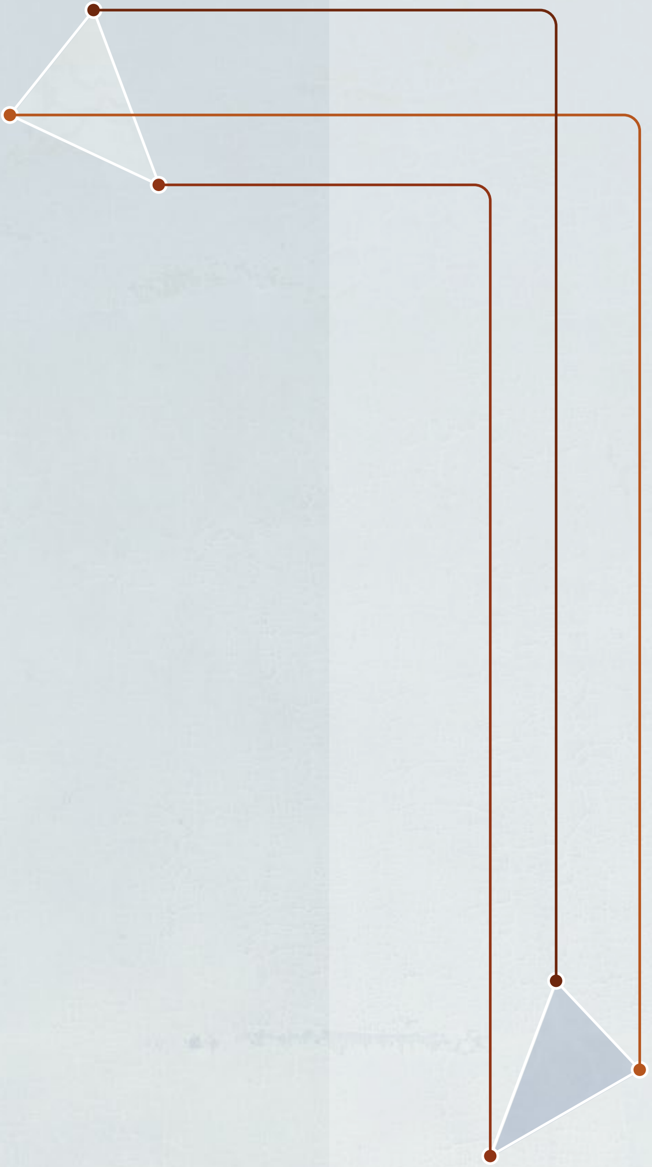
13. Verify the sum-to-product formulas

$$\sin\alpha+\sin\beta = 2\sin\left(\frac{\alpha+\beta}{2}\right)\cos\left(\frac{\alpha-\beta}{2}\right)$$

$$\sin\alpha-\sin\beta = 2\cos\left(\frac{\alpha+\beta}{2}\right)\sin\left(\frac{\alpha-\beta}{2}\right)$$

$$\cos\alpha+\cos\beta = 2\cos\left(\frac{\alpha+\beta}{2}\right)\cos\left(\frac{\alpha-\beta}{2}\right)$$

$$\cos\alpha-\cos\beta = -2\sin\left(\frac{\alpha+\beta}{2}\right)\sin\left(\frac{\alpha-\beta}{2}\right)$$



24 ISOMETRIES

One of the prevailing philosophies of modern mathematics is that in order to study something, you need to study the types of maps that preserve it– that is, the types of maps that leave it invariant. For instance, in group theory we study group homomorphisms because they preserve the group operation (in the sense that $f(a \cdot b) = f(a) \cdot f(b)$). In Euclidean geometry there are several structures that might be worth preserving– incidence, order, congruence– but in the next few lessons our focus will be on mappings that preserve distance.

Definitions

Let's start with a review of some basic terminology associated with maps from one set to another.

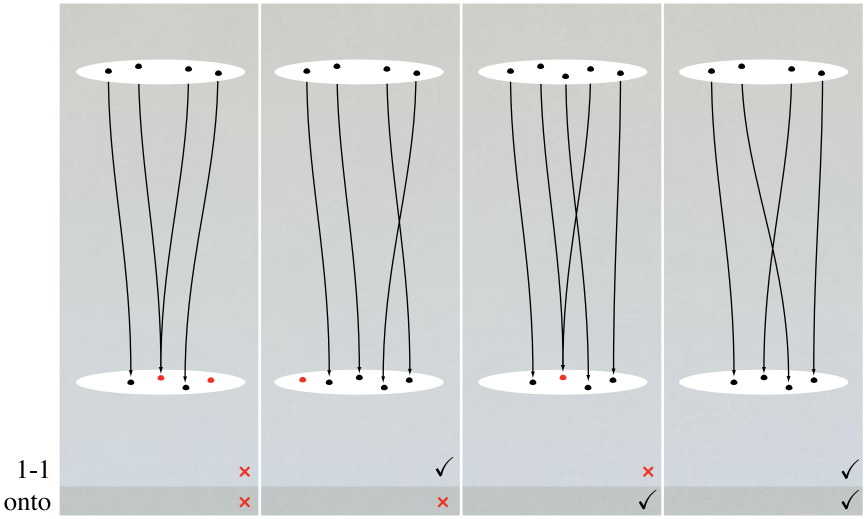
DEF: ONE-TO-ONE, ONTO, AND BIJECTIVE MAPPINGS

A map $f : X \to Y$ is:

· *one-to-one* if $f(x) = f(y) \implies x = y$;

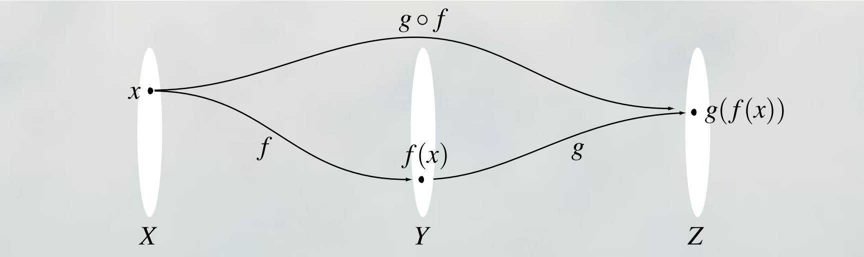
· *onto* if for every $y \in Y$ there is an $x \in X$ such that $f(x) = y$;

· *bijective* if it is both one-to-one and onto.



Under the right circumstances, two mappings may be chained together: the composition of $f : X \to Y$ and $g : Y \to Z$ is

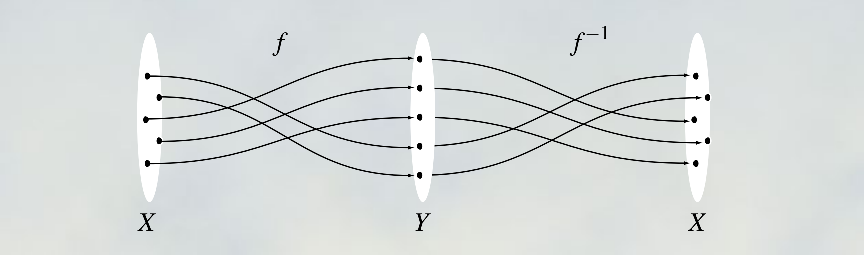
$$g \circ f : X \to Z : g \circ f(x) = g(f(x)).$$



This type of composition is usually not commutative– in fact, $f \circ g$ may not even be defined. It is associative, though, and that is a very essential property. For any space X the map

$$id : X \to X : id(x) = x$$

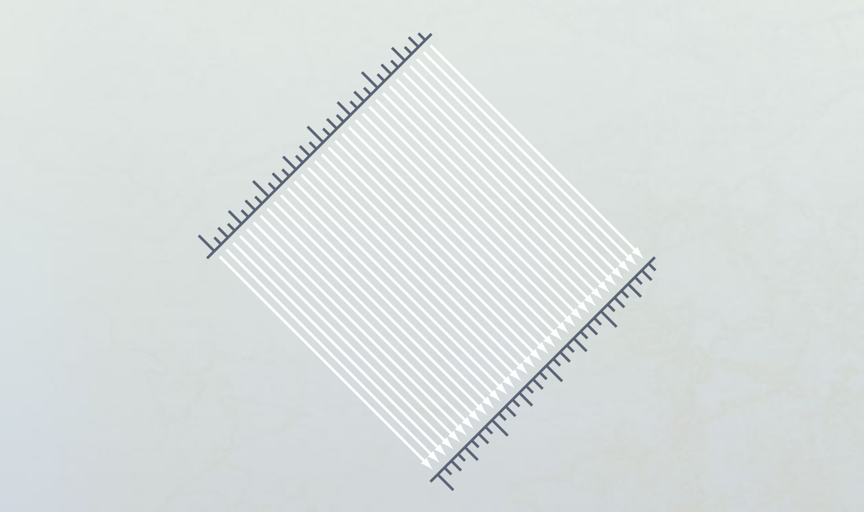
is called the identity map. Two maps $f : X \to Y$ and $g : Y \to X$ are inverses of one another if $f \circ g$ is the identity map on Y and $g \circ f$ is the identity map on X. In order for a map to have an inverse, it must be bijective (and conversely, any bijection is invertible).



DEF: AUTOMORPHISM

An automorphism is a bijective mapping f from a space to itself.

We are interested in automorphisms of the Euclidean plane, but not just any automorphisms. We want the ones that do not distort the distances between points. These are called Euclidean isometries.



DEF: ISOMETRY

Let $\mathbb{E}$ denote the set of points of the Euclidean plane. A Euclidean isometry is an automorphism $f : \mathbb{E} \to \mathbb{E}$ that preserves the distance between points: for all A, B in $\mathbb{E}$, $|f(A)f(B)| = |AB|$.

I will leave the proof of the following basic properties of isometries to you. If you are familiar with the concept of a group, these properties mean that the set of Euclidean isometries is a group.

LEM: BASIC PROPERTIES OF ISOMETRIES

The composition of two isometries is an isometry. The identity map is an isometry. The inverse of an isometry is an isometry.

Recall that everything we have done in Euclidean geometry floats on five undefined terms: point, line, on, between, and congruence. An isometry is defined in terms of its behavior on points, but the distance preservation condition has implications for the remaining undefined terms as well.

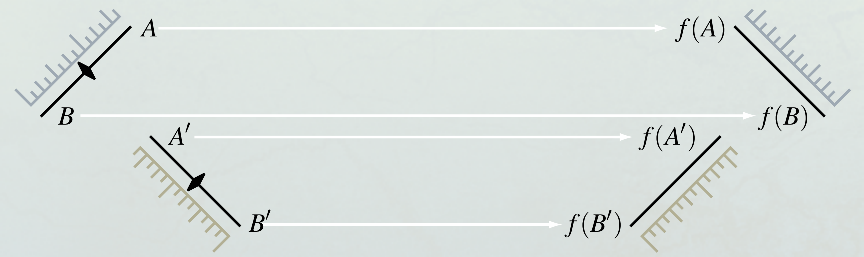
LEM: ISOMETRIES AND CONGRUENCE

An isometry preserves both segment and angle congruence. That is,

$$AB \simeq A'B' \implies f(A)f(B) \simeq f(A')f(B')$$
$$\angle ABC \simeq \angle A'B'C' \implies \angle f(A)f(B)f(C) \simeq \angle f(A')f(B')f(C')$$

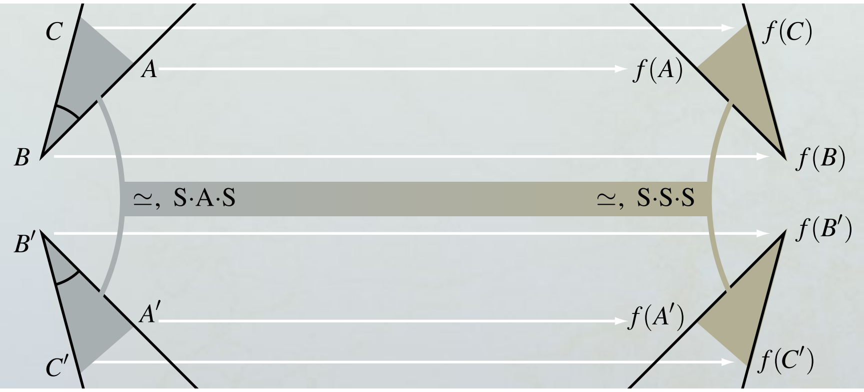
Proof. The segment congruence part is easy, because isometries preserve distance and hence segment length, and it is those lengths that determine whether or not segments are congruent: if $AB \simeq A'B'$, then

$$|f(A)f(B)| = |AB| = |A'B'| = |f(A')f(B')|$$



and so $f(A)f(B) \simeq f(A')f(B')$. The angle congruence part is not that hard either, but we will need to use a few of the triangle congruence theorems. Relocate, if necessary, A' and C' on their respective rays so that $BA \simeq B'A'$ and $BC \simeq B'C'$. By S·A·S, the triangles $\triangle ABC$ and $\triangle A'B'C'$ are congruent. The corresponding sides of these two triangles are congruent, and from the first part of the proof, the congruences are transferred by f:

$$\begin{aligned} AB \simeq A'B' &\implies f(A)f(B) \simeq f(A')f(B') \\ BC \simeq B'C' &\implies f(B)f(C) \simeq f(B')f(C') \\ CA \simeq C'A' &\implies f(C)f(A) \simeq f(C')f(A') \end{aligned}$$



By S·S·S, triangles $\triangle f(A)f(B)f(C)$ and $\triangle f(A)f(B)f(C)$ are congruent, and so the corresponding angles $\angle f(A)f(B)f(C)$ and $\angle f(A')f(B')f(C')$ are congruent. □

If you were paying attention in the last proof, you may have noticed that it could easily be tweaked to say a bit more: an isometry doesn't preserve just distance– it also preserves angle measure, in the sense that

$$(\angle ABC) = (\angle f(A)f(B)f(C)).$$

This is useful. In fact, we will use it in the last proof of this lesson.

LEM: ISOMETRIES, INCIDENCE AND ORDER

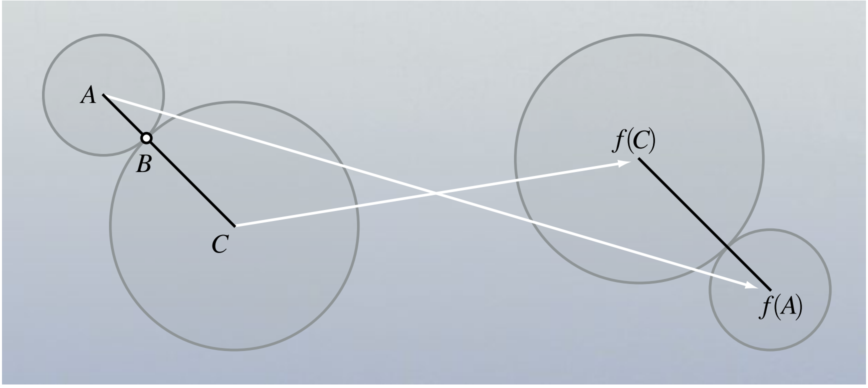
If A, B, and C are collinear, in the order $A*B*C$, and f is an isometry, then $f(A)$, $f(B)$, and $f(C)$ are collinear, in the order $f(A)*f(B)*f(C)$.

Proof. Suppose $A*B*C$. Then, by segment addition

$$|AC| = |AB| + |BC|.$$

Distance is invariant under f, so we can make the substitutions

$$|f(A)f(B)| = |AB|, \quad |f(B)f(C)| = |BC|, \quad |f(A)f(C)| = |AC|,$$



to get

$$|f(A)f(C)| = |f(A)f(B)| + |f(B)f(C)|.$$

This is the degenerate case of the Triangle Inequality: the only way this equation can be true is if $f(A)$, $f(B)$, and $f(C)$ are collinear, and that $f(B)$ is between $f(A)$ and $f(C)$. □

In the last result we were talking about three points, but by extension, this means that all the points on a line are mapped again to collinear points. In other words, an isometry, which is defined as a bijection of points, is also a bijection of the lines of the geometry. Further, an isometry maps segments to segments, rays to rays, angles to angles, and circles to circles. Well, here's an opportunity to simplify notation. When I apply an isometry f to a segment AB, for example, instead of writing $f(A)f(B)$, I will go with the more streamlined $f(AB)$. For an angle $\angle ABC$, instead of $\angle f(A)f(B)f(C)$, I will write $f(\angle ABC)$. And so on.

Fixed points

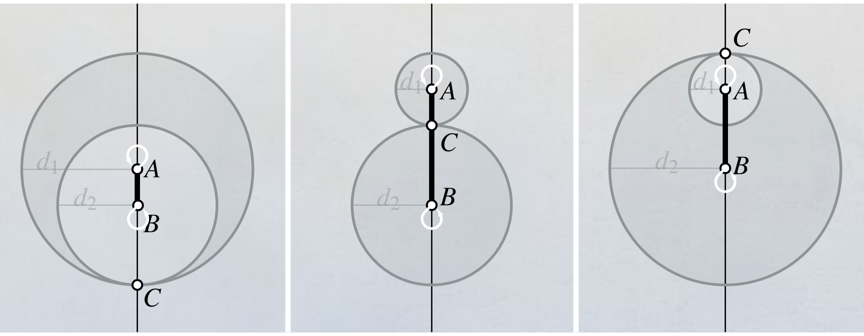
The overarching goal of the next few lessons is to classify all Euclidean isometries. It turns out that one of the keys to this is *fixed points*.

> DEF: FIXED POINT
> A point P is a fixed point of an isometry f if $f(P) = P$.

The first big step towards a classification is to answer the following question:

> Given isometries f_1 and f_2, which may be described in very different ways, how do we figure out if they are really the same?

Showing that they are *not* the same is usually easy– you just need to find one point P where $f_1(P) \neq f_2(P)$. Showing that they *are* the same seems like a more difficult task. At the most basic level, isometries are functions of the Euclidean plane. Without any additional structure, the only way to show two functions are equal is to show that they agree on the value of all points. This is because the behavior of an arbitrary function is quite unconstrained. Fortunately, the bijection and distance-preserving properties of an isometry impose significant constraints on its behavior. Those constraints mean that we can determine whether or not two isometries are the same by looking at just a few points.



f(C) must still be on both of these circles.

THM: TWO FIXED POINTS

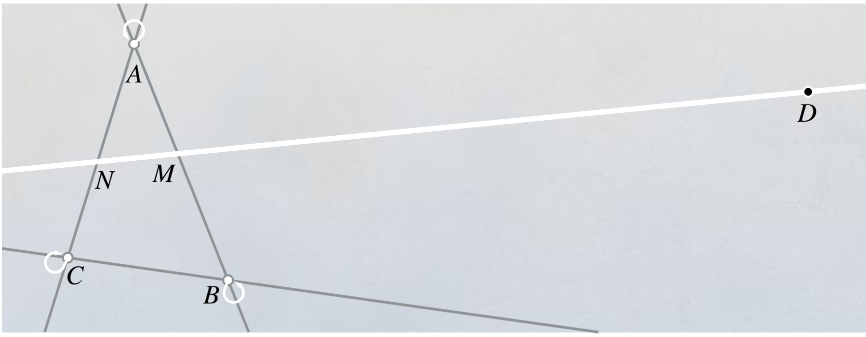
If an isometry f fixes two distinct points A and B, then it fixes all the points of the line $\leftarrow AB \rightarrow$.

Proof. Let C be a third point on this line. Label its distances from A as d_1 and from B as d_2. The key here is that C is the only point that is a distance d_1 from A *and* a distance d_2 from B (I think this is intuitively clear, but for a more formal point of view, you can look back at our investigation of the possible intersections of circles in Lesson 16). Now hit these three points with the isometry f. Distances stay the same, so $f(C)$ is still a distance d_1 from $f(A) = A$, and $f(C)$ is still a distance d_2 from $f(B) = B$. That means that $f(C)$ must be C. □

THM: THREE (NON-COLLINEAR) FIXED POINTS

If an isometry f fixes three non-collinear points A, B, and C, then it fixes all points (it is the identity isometry).

Proof. By the last result, f must fix all the points on each of the lines $\leftarrow AB \rightarrow$, $\leftarrow AC \rightarrow$, and $\leftarrow BC \rightarrow$. Now suppose that D is a point that is not on any of those lines. We need to show that D is a fixed point as well. Choose a point M that is between A and B. It is fixed by f. According to Pasch's lemma, the line $\leftarrow DM \rightarrow$ must intersect at least one other side of $\triangle ABC$. Call this intersection N. It too is fixed by f. Therefore D is on a line $\leftarrow MN \rightarrow$ with two fixed points. According to the previous result, it is a fixed point. □



A line through D intersecting two fixed lines.

Now we can answer the question I posed at the start of this section: how much do we need to know about two isometries before we can say they are the same?

THM: THREE NON-COLLINEAR POINTS ARE ENOUGH

If two isometries f_1 and f_2 agree on three non-collinear points, then they are equal.

Proof. Suppose that A, B, and C are three non-collinear points, and that

$$f_1(A) = f_2(A) \quad f_1(B) = f_2(B) \quad f_1(C) = f_2(C).$$

Applying f_2^{-1} to both sides of each of these equations,

$$\begin{aligned} f_2^{-1} \circ f_1(A) &= f_2^{-1} \circ f_2(A) = id(A) = A, \\ f_2^{-1} \circ f_1(B) &= f_2^{-1} \circ f_2(B) = id(B) = B, \\ f_2^{-1} \circ f_1(C) &= f_2^{-1} \circ f_2(C) = id(C) = C. \end{aligned}$$

Therefore $f_2^{-1} \circ f_1$ has three non-collinear fixed points– it must be the identity, and so

$$\begin{aligned} & f_2^{-1} \circ f_1 = id \\ & f_2 \circ f_2^{-1} \circ f_1 = f_2 \circ id \\ & id \circ f_1 = f_2 \\ & f_1 = f_2. \end{aligned}$$

□

The analytic viewpoint

To wrap up this lesson, let's look at isometries from the analytical point of view. Any isometry defines a function on the coordinate pairs. As we have seen, isometries themselves are fairly structured, so it makes sense, then, that the functions they define on the coordinate pairs would have to be similarly inflexible. That is indeed the case.

GENERAL FORM FOR AN ISOMETRY

Any Euclidean isometry T has analytic equations that can be written in one of two matrix forms

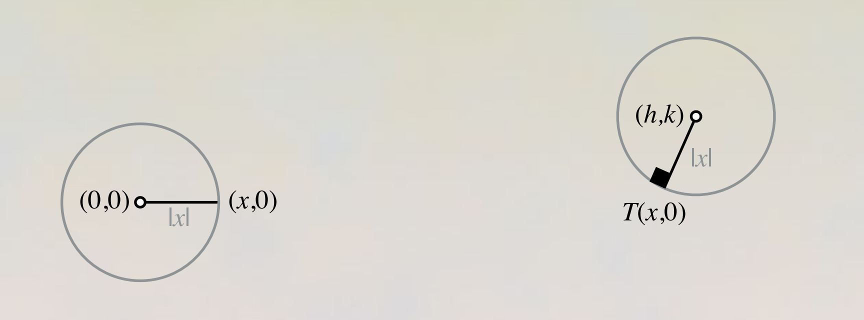
$$(1)\quad T\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}h\\k\end{pmatrix}+\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}$$

$$(2)\quad T\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}h\\k\end{pmatrix}+\begin{pmatrix}\cos\theta & \sin\theta\\ \sin\theta & -\cos\theta\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}$$

where h, k, and θ are real numbers.

Proof. Let T be an isometry. Ultimately, we want to know about $T(x,y)$, but it will take a few steps to get there, starting with the origin, moving to the point $(x,0)$, and then finally to (x,y).

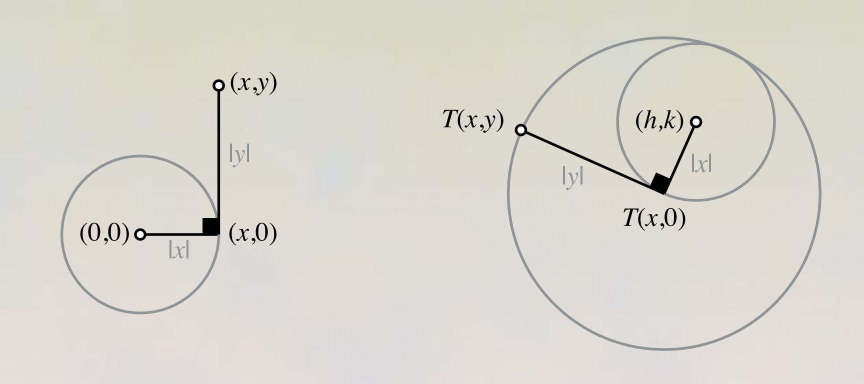
The origin $(0,0)$. This is the easy one. Since the origin is our first point of consideration, there are no limitations on where it goes (we don't know it yet, but there are isometries that take any point to any other point of the plane). Set h and k by looking at what happens to the origin: set $(h,k)=T(0,0)$.



The point $(x,0)$. An isometry preserves distances, and the distance from $(x,0)$ to the origin is $|x|$. Applying the isometry to both of those points, the distance from $T(x,0)$ to (h,k) also has to be $|x|$. In other words, $T(x,0)$ is on the circle with center (h,k) and radius $|x|$. If you did the exercise in the last lesson on parametrizing circles (or if you have worked with parametrized circles in calculus), then you know this means that $T(x,0)$ has to have the form

$$(h+|x|\cos\theta,\ k+|x|\sin\theta)$$

for some value of θ. In fact (and I will leave it to you to figure out why), the absolute value signs around the x are not needed.



The point (x,y). Likewise, since the distance from $(x,0)$ to (x,y) is $|y|$, $T(x,y)$ has to be on the circle centered at $T(x,0)$ with radius $|y|$. That means its coordinates can be written in the form

$$(h+x\cos\theta+|y|\cos\phi,\ k+x\sin\theta+|y|\sin\phi)$$

for some value of ϕ. The possibilities are more limited than that, though: the three points $(0,0)$, $(x,0)$ and (x,y) form a right angle at $(x,0)$. Since an isometry preserves angle measures, the images of these three points must also form a right angle. This can only happen if $\phi=\theta+\pi/2$ or $\phi=\theta-\phi/2$. As before, the absolute value signs around the y can be dropped and that gets us to:

$$\left(h+x\cos\theta+y\cos\left(\theta\pm\frac{\pi}{2}\right),\ k+x\sin\theta+y\sin\left(\theta\pm\frac{\pi}{2}\right)\right).$$

Now use the addition formulas for sine and cosine

$$\begin{aligned}\cos(\theta \pm \pi/2) &= \cos\theta\cos(\pm\pi/2) - \sin\theta\sin(\pm\pi/2) = \mp\sin\theta\\ \sin(\theta \pm \pi/2) &= \sin\theta\cos(\pm\pi/2) + \cos\theta\sin(\pm\pi/2) = \pm\cos\theta\end{aligned}$$

and the coordinates for $T(x,y)$ take on the form

$$\begin{aligned}&(1)\quad T(x,y) = (h + x\cos\theta - y\sin\theta, k + x\sin\theta + y\cos\theta)\\ &(2)\quad T(x,y) = (h + x\cos\theta + y\sin\theta, k + x\sin\theta - y\cos\theta).\end{aligned}$$

Written in matrix form, these are

$$(1)\quad T\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}h\\k\end{pmatrix} + \begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}$$

$$(2)\quad T\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}h\\k\end{pmatrix} + \begin{pmatrix}\cos\theta & \sin\theta\\ \sin\theta & -\cos\theta\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}.$$

□

Exercises

1. Let T be an isometry and let r be a ray with endpoint O. Prove that $T(r)$ is also a ray, with endpoint $T(0)$.

2. Verify that if ℓ_1 and ℓ_2 are parallel lines and T is an isometry, then $T(\ell_1)$ and $T(\ell_2)$ will be parallel.

3. Let T be an isometry and let A and B be two points that are on the same side of a line ℓ. Prove that $T(A)$ and $T(B)$ are on the same side of $T(\ell)$.

4. Let T be an isometry and let D be a point in the interior of angle $\angle ABC$. Prove that $T(D)$ is a point in the interior of $T(\angle ABC)$.

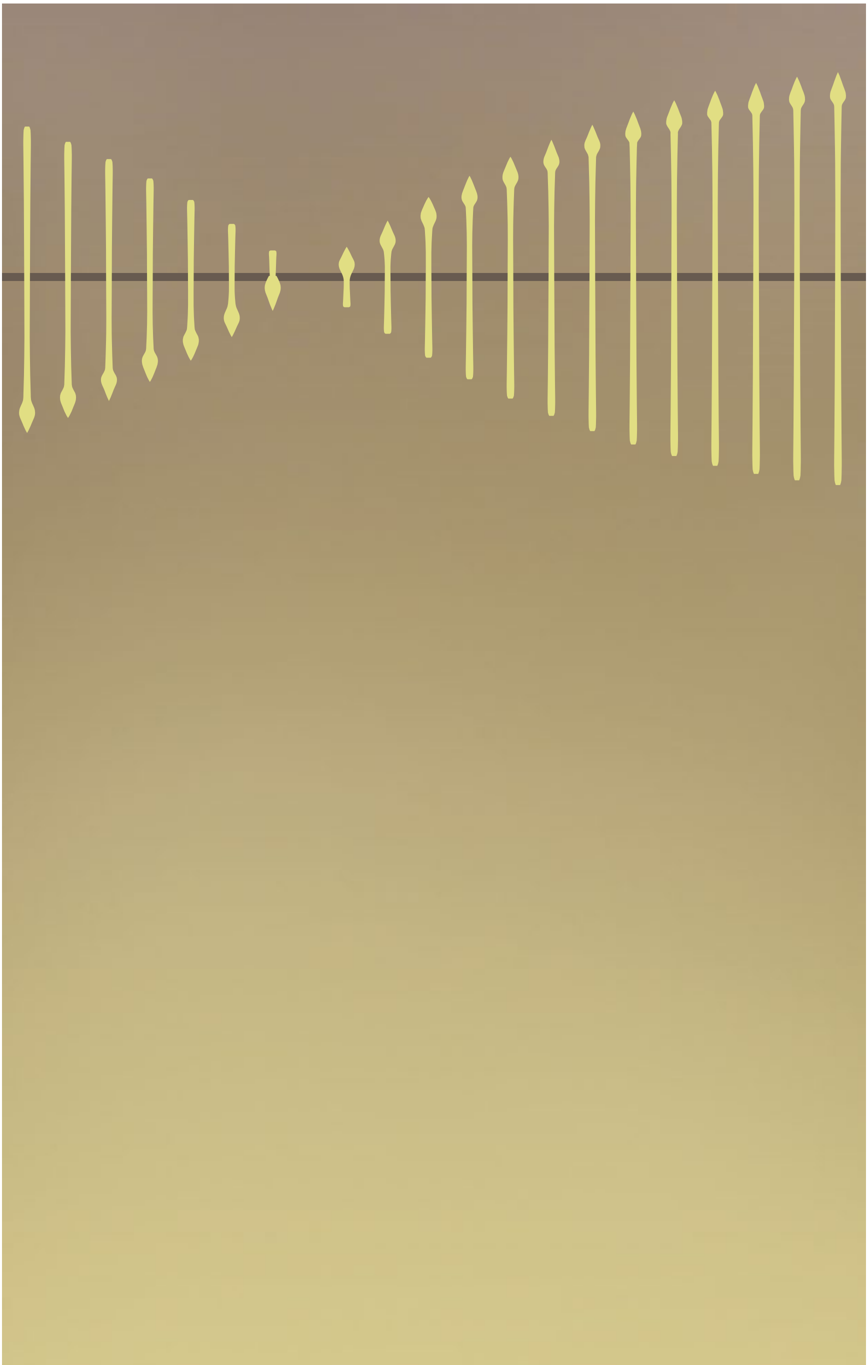
5. Let M be the midpoint of a segment AB, and let T be an isometry so that $T(A) = B$ and $T(B) = A$. Prove that M is a fixed point of this isometry.

6. Given a proper angle $\angle ABC$ and an isometry T such that

$$(1)\ T(BA\rightarrow) = BC\rightarrow \quad \& \quad (2)\ T(BC\rightarrow) = BA\rightarrow,$$

show that T fixes all the points of the angle bisector of $\angle ABC$.

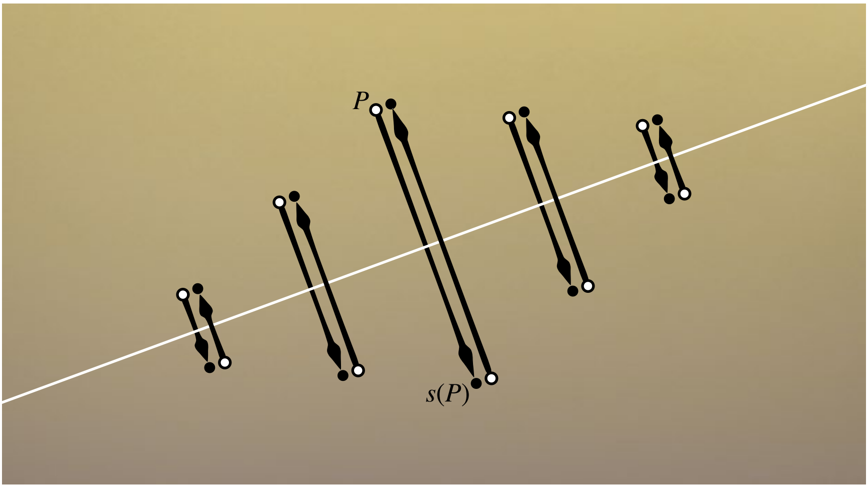
7. In the final theorem of this lesson I showed that every isometry can be written in one of two forms. Prove the converse, that any mapping of that form is an isometry.



25 REFLECTIONS

This lesson introduces the first type of isometry– reflection across a line. As it turns out, reflections are the building blocks for all isometries. In this lesson we will see why, in a theorem that I don't believe has a formal name, but that I call the "Three Reflections Theorem". This theorem provides the strategy that we will use over the next few lessons to classify all isometries.

> DEF: REFLECTION ACROSS A LINE
> Define the reflection s across a line ℓ as follows. For any point P on ℓ, set $s(P) = P$. For any point P that is not on ℓ, there is a unique line passing through P that is perpendicular to ℓ. On this line, there is one other point that is the same distance from ℓ as P– it is on the opposite side of ℓ from P. Set $s(P)$ to be this point.



Of course, the first agenda item is to verify that a reflection really is an isometry.

THM
A reflection is an isometry.

Proof. It is easy to see that any reflection s is a bijection. Just look at the composition $s \circ s$: the swap of points done by the first application of s is immediately undone by the second application of s, so that $s^2 = id$. Therefore s is its own inverse, and in order for a mapping to have an inverse, it must be a bijection.

The other step is to show that s preserves distances– that $|s(PQ)| = |PQ|$ for any points P and Q. The only thing that makes this part difficult is that there are so many possible positions of P and Q relative to each other and to ℓ, the line of reflection:

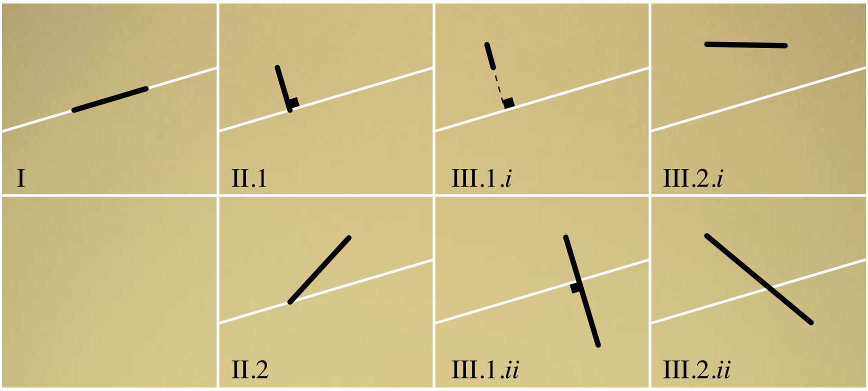
I. P and Q are both on ℓ.

II. One of P and Q is on ℓ, while the other is not.

1. the line $\leftarrow PQ \rightarrow$ is perpendicular to ℓ
2. the line $\leftarrow PQ \rightarrow$ is not perpendicular to ℓ

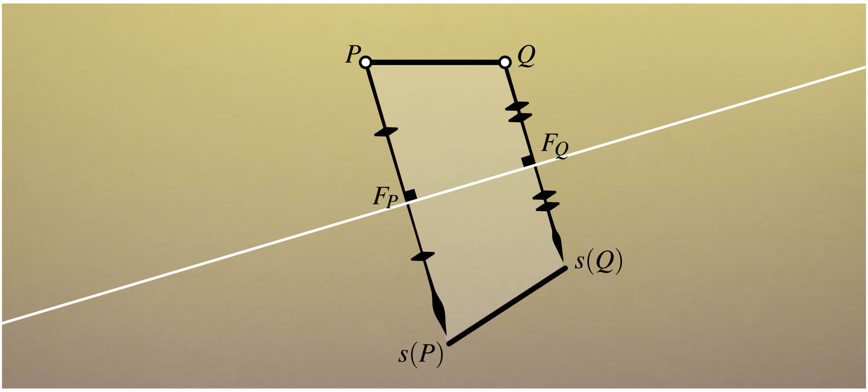
III. Neither P nor Q is on ℓ.

1. the line $\leftarrow PQ \rightarrow$ is perpendicular to ℓ
    - *i*. P and Q are on the same side of ℓ
    - *ii*. P and Q are on opposite sides of ℓ
2. the line $\leftarrow PQ \rightarrow$ is not perpendicular to ℓ
    - *i*. P and Q are on the same side of ℓ
    - *ii*. P and Q are on opposite sides of ℓ



At this point, none of these cases should cause any trouble. Let me look at just one, Case III.2.*i*, which is, I feel, the archetypal case in this proof. To verify this case, first label two more points (both fixed by s).

F_P: the foot of the perpendicular to ℓ through P, and
F_Q: the foot of the perpendicular to ℓ through Q.



From the very definition of a reflection,

$$PF_P \simeq s(PF_P) \quad \& \quad QF_Q \simeq s(QF_Q)$$

and the angles at F_P and F_Q are right angles. Of course F_PF_Q is congruent to itself, so by S·A·S·A·S, the quadrilaterals PF_PF_QQ and $s(PF_PF_QQ)$ are congruent, and therefore PQ and $s(PQ)$ are the same length. □

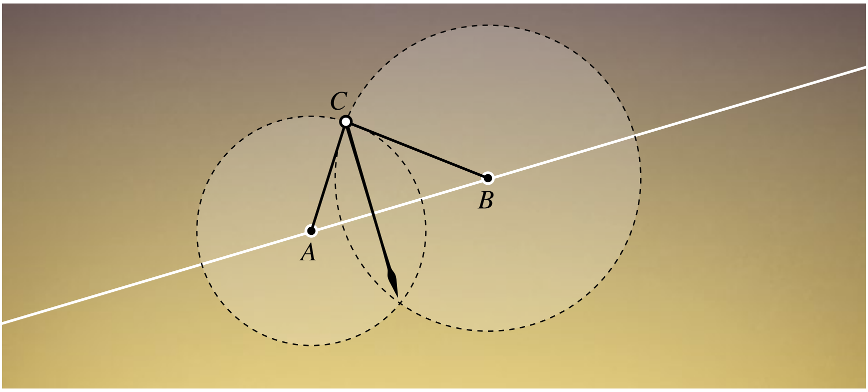
We saw in the last lesson that if an isometry fixes two points, it must fix all the points on the line through those points. Of course, every reflection fixes all the points of a line. A good question to ask, then, is how common is this "line-fixing" behavior? Not that common, as it turns out, and so this is a useful characterization of a reflection.

THM

If an isometry fixes all the points of a line, but is not the identity, then it must be a reflection.

Proof. Let f be an isometry that fixes all the points on a line ℓ but that is not the identity. Let s be the reflection across that line. We already know that f and s agree on all the points of ℓ, so we just need to show that they agree on one point that isn't on ℓ. Take two points A and B on ℓ, and a third point C that is not on ℓ. Since an isometry preserves distance, and since both A and B are fixed

$$\begin{aligned}
&AC \simeq f(AC) \simeq Af(C)\\
&\qquad \Longrightarrow f(C) \text{ is on the circle with center } A \text{ and radius } |AC|\text{, and}\\
&BC \simeq f(BC) \simeq Bf(C)\\
&\qquad \Longrightarrow f(C) \text{ is on the circle with center } B \text{ and radius } |BC|.
\end{aligned}$$

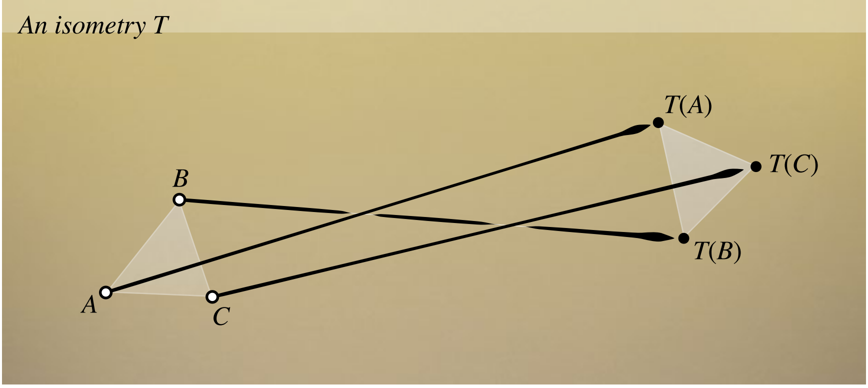


We are triangulating in on the location of $f(C)$: it has to be at an intersection of these two circles, and there are only two such intersections (distinct circles intersect at most twice). Furthermore, one of those intersections is C itself, and if $f(C) = C$, then f would fix three non-collinear points and would have to be the identity. We excluded that possibility at the outset, so $f(C)$ has to be the other intersection of the circles. For all the same reasons, $s(C)$ must also be that second intersection. Therefore $f(C) = s(C)$, the two isometries agree on three non-collinear points, A, B, and C, and so they must be equal. □

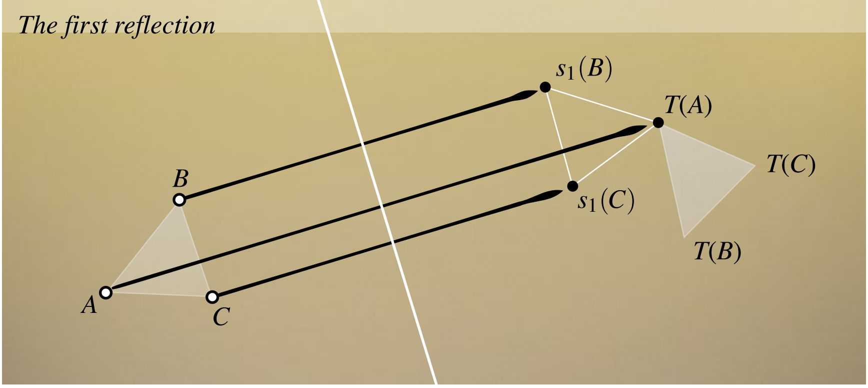
THE THREE REFLECTIONS THEOREM

Any isometry can be written as a reflection, as a composition of two reflections, or as a composition of three reflections.

Proof. Let A, B, and C be three non-collinear points and let T be an isometry. We saw in the last lesson that when isometries agree on three non-collinear points, they have to be the same. That is how we will proceed. We just need to find a composition of up to three reflections $s_3 \circ s_2 \circ s_1$ that agrees with T on each of A, B, and C. There are three steps to this. At each step we want to get one of the three points into the right position, without moving any of the previously set points.


An isometry T

Step One. With the first isometry, s_1, we are going to get A into position. If $A = T(A)$, let s_1 be the identity isometry. If $A \neq T(A)$, let s_1 be the reflection across the perpendicular bisector of $AT(A)$. Either way, $s_1(A) = T(A)$.

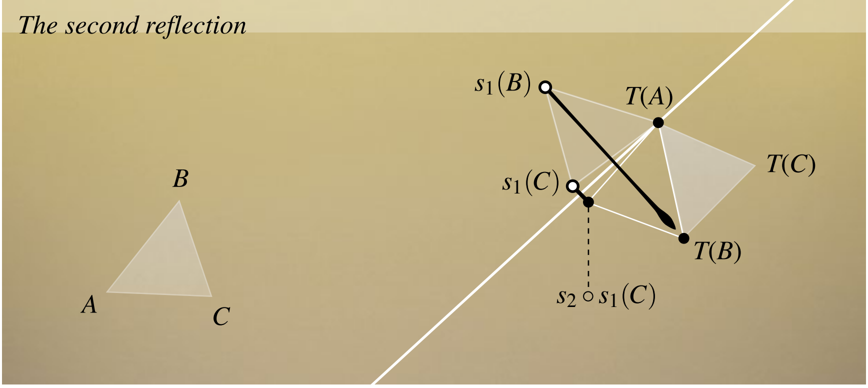

The first reflection

Step Two. With the second isometry, s_2, we put B into position. In order to do this, we need to look at where $s_1(B)$ ended up after step one. It is possible (but unlikely) that $s_1(B)$ ended up on the line $T(\leftarrow AB\rightarrow)$. If that is the case, then because

$$|s_1(AB)| = |AB| = |T(AB)|,$$

there are only two possible spots for $s_1(B)$, one on either side of $T(A)$. If $s_1(B)$ is on the same side of $T(A)$ as $T(B)$, then $s_1(B) = T(B)$ already, so we can just let s_2 be the identity isometry. If $s_1(B)$ is on the opposite side of $T(A)$ from $T(B)$, then let s_2 be the reflection across the line that passes through $T(A)$ and is perpendicular to $s_1(B)T(B)$. That reflection fixes $T(A)$ and maps $s_1(B)$ to $T(B)$.

The more likely possibility is that $s_1(B)$ is not on $T(\leftarrow AB\rightarrow)$. In that case, let s_2 be the reflection across the bisector of $\angle s_1(B)T(A)T(B)$. Then $T(A)$ is on the line of reflection, so it will be fixed by s_2. Furthermore, the reflecting line cuts the triangle $\triangle s_1(B)T(A)T(B)$ in two pieces, that, by S·A·S, are congruent. Therefore the reflecting line is the perpendicular bisector to $s_1(B)T(B)$– and that means s_2 will map $s_1(B)$ to $T(B)$.

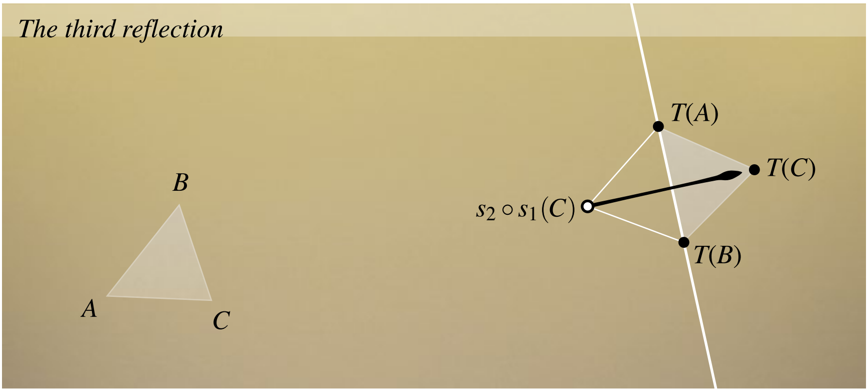

The second reflection

Here is where we stand after step two:

$$s_2 \circ s_1(A) = s_2 \circ T(A) = T(A),$$
$$s_2 \circ s_1(B) = T(B).$$

Step Three. That just leaves point C. As in the previous step, what we do next depends upon where $s_2 \circ s_1(C)$ is. There aren't that many possibilities at this point though. We know that $s_2 \circ s_1(AB) = T(AB)$, and we know that $s_2 \circ s_1(\triangle ABC)$ is congruent to $\triangle ABC$, which is, in turn, congruent to $T(\triangle ABC)$. There are only two ways to build that triangle on the given side $T(AB)$– one on either side of it. If $s_2 \circ s_1(C)$ is on the same side of $T(AB)$ as $T(C)$, then $s_2 \circ s_1(C) = T(C)$ already, so just let s_3 be the identity map. If $s_2 \circ s_1(C)$ is on the opposite side of $T(AB)$ from $T(C)$, then let s_3 be the reflection across the line $T(AB)$. That fixes both $T(A)$ and $T(B)$, but maps $s_2 \circ s_1(C)$ onto $T(C)$.

The third reflection



Putting it all together,

$$\begin{aligned} s_3 \circ s_2 \circ s_1(A) &= T(A) \\ s_3 \circ s_2 \circ s_1(B) &= T(B) \\ s_3 \circ s_2 \circ s_1(C) &= T(C). \end{aligned}$$

Since the two isometries agree on three non-collinear points, they must be the same. As long as at least one of s_1, s_2, and s_3 is a reflection, we have met the requirements of the theorem. What if all of them are the identity map though? In that case, T is the identity map, and the identity can be written as the composition of any reflection s with itself: $T = s \circ s$. □

Over the next few lessons, we will use this result to classify all isometries. In the next lesson, we will look at what happens when you compose two reflections. Then, after a little diversion, we will look at what happens when you tack on a third reflection.

The analytic viewpoint

It is a little messy to try to work out an equation for an arbitrary reflection at this point. We can, however, work out an equation for a reflection across a line that passes through the origin. Let's close out this lesson by doing so.

EQN: REFLECTION ACROSS A LINE THROUGH THE ORIGIN
Let ℓ be a line through the origin, and let (a,b) be the coordinates of an intersection of ℓ with the unit circle. Then the reflection s across this line is given by the equation

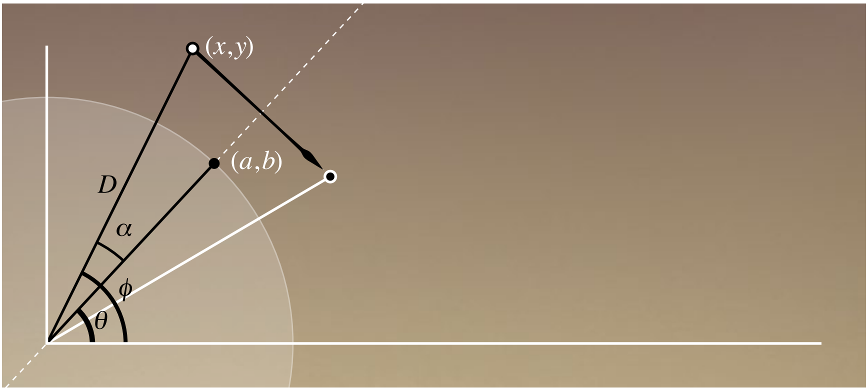
$$s\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}a^2-b^2 & 2ab\\ 2ab & b^2-a^2\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}$$

Proof. Since (a,b) is on the unit circle, it can be written as $(\cos\theta,\sin\theta)$. Let D be the distance from the point (x,y) to the origin and let ϕ be its angle measure as measured from the x-axis, in the counterclockwise direction, so that

$$\begin{cases}\cos\phi=x/D\\ \sin\phi=y/D\end{cases}\implies\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}D\cos\phi\\ D\sin\phi\end{pmatrix}.$$

If α is the angle between ϕ and θ, $\alpha=\phi-\theta$, then $s(x,y)$ will still be at a distance D from the origin, but at an angle

$$\phi-2\alpha=\phi-2(\phi-\theta)=2\theta-\phi.$$



Therefore

$$s\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}D\cos(2\theta-\phi)\\D\sin(2\theta-\phi)\end{pmatrix}$$

and we can use the addition rules for sine and cosine

$$s\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}D\cos(2\theta)\cos(-\phi)-D\sin(2\theta)\sin(-\phi)\\D\sin(2\theta)\cos(-\phi)+D\cos(2\theta)\sin(-\phi)\end{pmatrix}$$

$$=\begin{pmatrix}D\cos(2\theta)\cos\phi+D\sin(2\theta)\sin\phi\\D\sin(2\theta)\cos\phi-D\cos(2\theta)\sin\phi\end{pmatrix}.$$

This can factored into a matrix form, and from there, the double angle formulas will take us the rest of the way.

$$s\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}\cos(2\theta)&\sin(2\theta)\\\sin(2\theta)&-\cos(\theta)\end{pmatrix}\begin{pmatrix}D\cos\phi\\D\sin\phi\end{pmatrix}$$

$$=\begin{pmatrix}\cos^2\theta-\sin^2\theta&2\sin\theta\cos\theta\\2\sin\theta\cos\theta&\sin^2\theta-\cos^2\theta\end{pmatrix}\begin{pmatrix}D\cos\phi\\D\sin\phi\end{pmatrix}$$

$$=\begin{pmatrix}a^2-b^2&2ab\\2ab&b^2-a^2\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}.$$

□

There are two special cases worth noting. The equation for reflecting across the x-axis is

$$s\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}1&0\\0&-1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}$$

and the equation for reflecting across the y-axis is

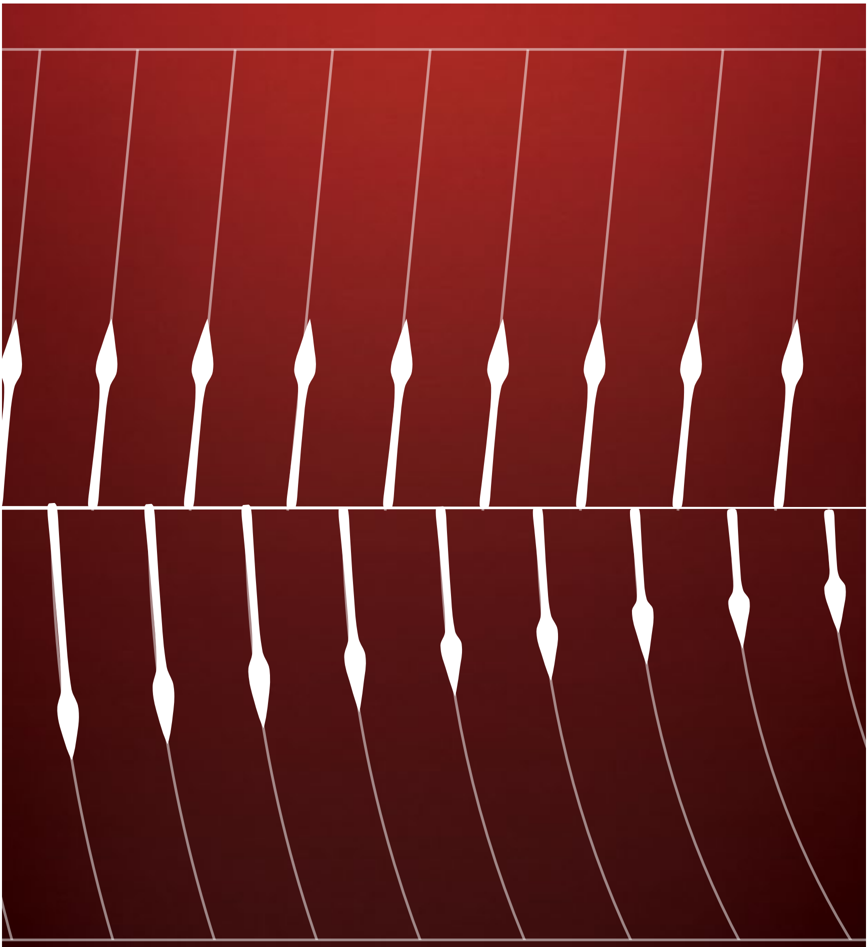
$$s\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}-1&0\\0&1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}.$$

Exercises

1. What is the matrix equation for a reflection across the line $y = x$?

2. What is the matrix equation for a reflection across the horizontal line $y = k$?

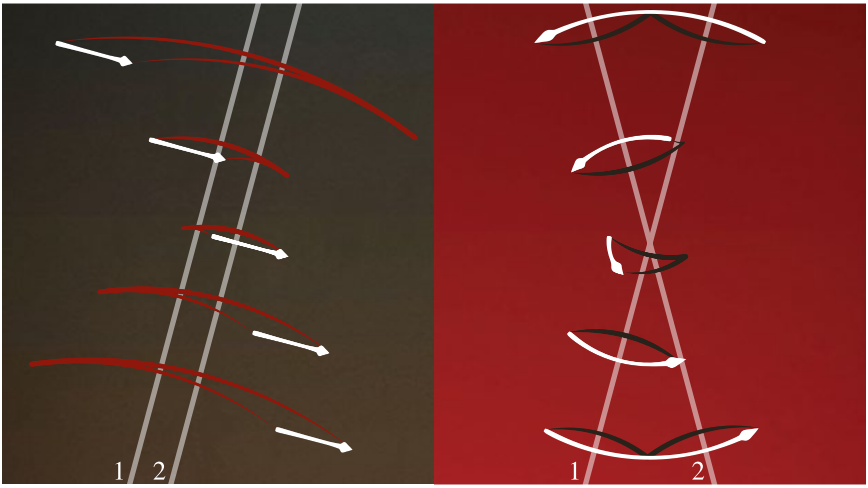
3. Let s_1 and s_2 be reflections across perpendicular lines ℓ_1 and ℓ_2 that intersect at a point P. Show that if Q is any other point, then P is the midpoint of the segment connecting Q to $s_2 \circ s_1(Q)$.



26 TRANSLATIONS & ROTATIONS

The big result of the last lesson was that every isometry can be written as a reflection, or as a composition of two or three reflections. In this lesson we will look at the types of isometries that you can get by composing two reflections. Of course, any reflection composed with itself results in the identity, so we are really interested in compositions of two *distinct* reflections. In that case, there are essentially two scenarios.

Scenario 1: the reflecting lines are parallel
Scenario 2: the reflecting lines are intersecting



The two scenarios do describe two fundamentally different types of isometries. In the second scenario, the intersection point of the two lines is fixed by the composition of isometries. This doesn't happen in the first scenario, since there is no intersection point, and in fact, this type of composition does not have any fixed points.

DEF: TRANSLATION AND ROTATION
A translation is a composition of reflections across parallel lines. A rotation is a composition of reflections across intersecting lines.

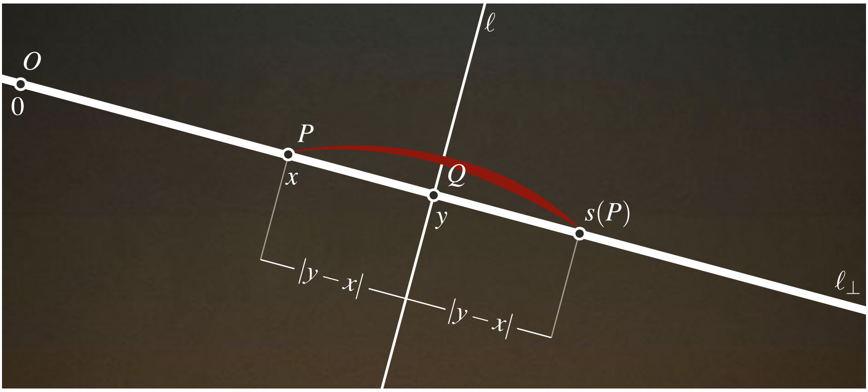
These are strategic definitions– by defining translations and rotations as compositions of isometries, it is automatically true that they will be isometries as well. But these definitions do not do a good job of revealing what a translation or rotation actually looks like. That is the purpose of this lesson.

Translation

First, let's tackle the case of the translation. To do that, I think it is helpful to back up a little bit, and to take a more measured look at the behavior of a single reflection. Consider a reflection s across a line ℓ. Let P be a point that is not on ℓ and let $\ell_\perp$ be the line through P that is perpendicular to ℓ. Now let's set up $\ell_\perp$ as a number line. That is, choose an arbitrary point O to be the origin, and a ray from O that points in the positive direction; then every point on $\ell_\perp$ has a "coordinate"– its signed distance from O. Suppose that P is at coordinate x, and that ℓ and $\ell_\perp$ intersect at the point Q with coordinate y. Given the definition of a reflection, $s(P)$ has to be somewhere on $\ell_\perp$ as well, and so it too must correspond to some coordinate. Well, what is that coordinate? The distance from P to Q is $|y-x|$. Since s is an isometry and Q is a fixed point, the distance from $s(P)$ to Q is $|y-x|$ too. That limits the possible coordinates for $s(P)$ to:

$$y+|y-x| = \begin{cases} y+(y-x) = 2y-x & \text{if } y-x \geq 0 \\ y-(y-x) = x & \text{if } y-x < 0 \end{cases}$$

$$y-|y-x| = \begin{cases} y-(y-x) = x & \text{if } y-x \geq 0 \\ y-(-(y-x)) = 2y-x & \text{if } y-x < 0. \end{cases}$$



Since P is not on ℓ, it is not a fixed point, so $s(P)$ is not at the coordinate x. The only other possibility, then, is that $s(P)$ is at the coordinate $2y-x$. Note that this formula still works even if P is on ℓ. In that case P is fixed, so $s(P)$ should also be at coordinate x. And that is what the formula reveals: if P is on ℓ, $y=x$, and so $2y-x=x$. Having this little formula in hand will make it a little easier to compose parallel reflections.

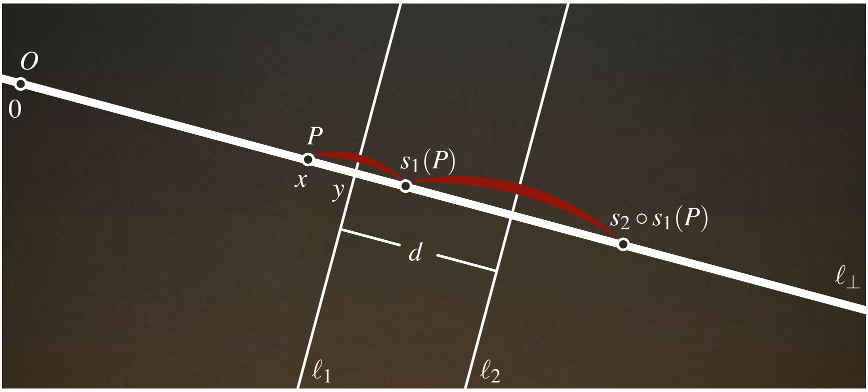
THM: PROPERTIES OF A TRANSLATION

Suppose that t is the translation $s_2 \circ s_1$ where s_1 and s_2 are reflections across parallel lines ℓ_1 and ℓ_2 that are separated by a distance d. Then for any point P, $t(P)$ is located

1) on the line through P that is perpendicular to both ℓ_1 and ℓ_2,
2) in the direction of the ray that points from ℓ_1 to ℓ_2,
3) at a distance $2d$ from P.

Proof. Take a point P, and let $\ell_\perp$ be the line through P that is perpendicular to ℓ_1 and ℓ_2. By definition, $s_1(P)$ will still be on $\ell_\perp$, and then so will $s_2(s_1(P))$. Let's just look along this line then, and, as in the preceding discussion, lay out a number line along it. It does not matter where you put the origin on the line, but it does help the discussion to choose the positive direction so that going from ℓ_1 to ℓ_2 moves in the positive direction. Then mark these coordinates:

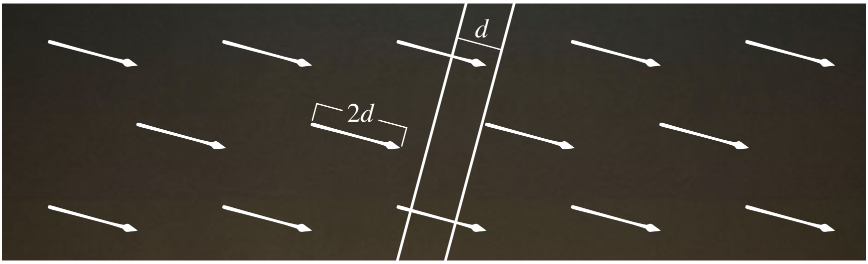
x: coordinate of P
y_1: coordinate for the intersection of $\ell_\perp$ and ℓ_1
y_2: coordinate for the intersection of $\ell_\perp$ and ℓ_2



According to our previous calculations, $s_1(x)$ will be at coordinate $2y_1 - x$ and $s_2 \circ s_1(x)$ will be at coordinate

$$2y_2 - (2y_1 - x) = x + 2(y_2 - y_1) = x + 2d.$$

Therefore $s_2 \circ s_1(P)$ will be $2d$ farther along the line $\ell_\perp$ than P, in the direction pointing from ℓ_1 to ℓ_2. □



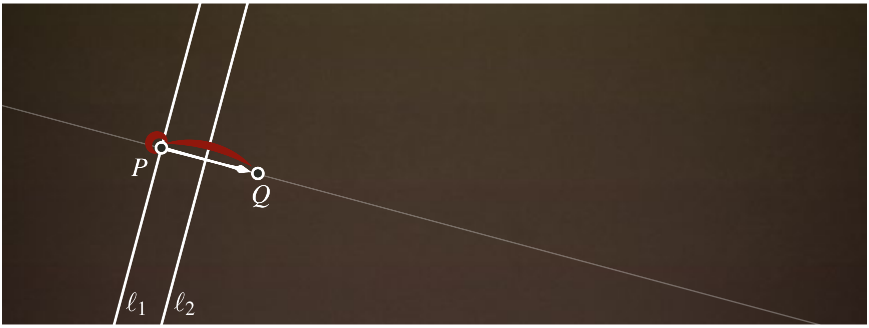
So you see, a translation moves all points along lines that are perpendicular to ℓ_1 and ℓ_2. They all move in parallel, in the same direction, over the same distance. All of that– the parallel lines, the direction, and the distance– can be determined by looking at the effect of the translation on a single point. That means that a translation is completely determined by its behavior on a single point. And because of that, we can get a very precise idea of how many translations there are.

THM: THERE ARE JUST ENOUGH TRANSLATIONS

Given any two distinct points P and Q, there is exactly one translation t so that $t(P) = Q$.

Proof. Existence: Let's just take the most straightforward approach and describe a translation that maps P to Q. The two reflections, s_1 and s_2, will be across lines that are perpendicular to PQ (and hence are parallel to one another). Let s_1 be the reflection across the line through P. Let s_2 be the reflection across the line through the midpoint of PQ. Then $s_2 \circ s_1$ is a translation and

$$s_2 \circ s_1(P) = s_2(P) = Q.$$

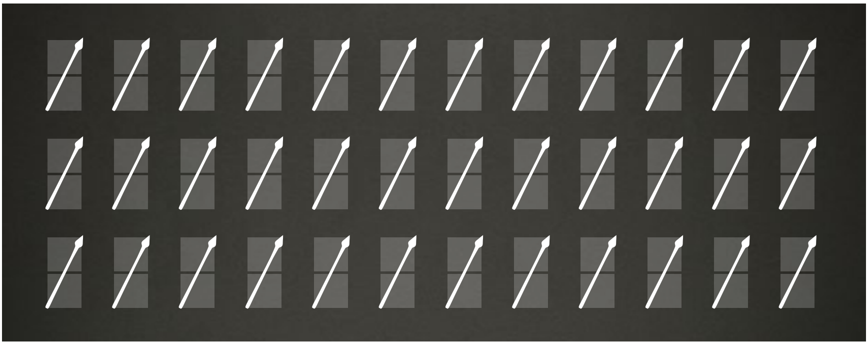


Uniqueness: Since a translation is completely determined by its behavior on one point, there can be only one translation taking P to Q. □

In the long run, it is cumbersome to try to think of a translation as a composition of reflections. The properties derived above give a much better sense of the effects of a translation, and those properties can be formalized as follows. A *directed segment* is a line segment that distinguishes between the two ends: one is called the initial endpoint, the other the terminal endpoint. We can define an equivalence relation on the set of all directed segments as follows: two directed segments σ_1 and σ_2 are equivalent if there is a translation t mapping σ_1 to σ_2, so that initial point is mapped to initial point, and terminal point is mapped to terminal point.

DEF: VECTOR

A *vector* is an equivalence class of directed segments.



Some equivalence class representatives of the vector $\langle 1,2\rangle$ *(one over, two up).*

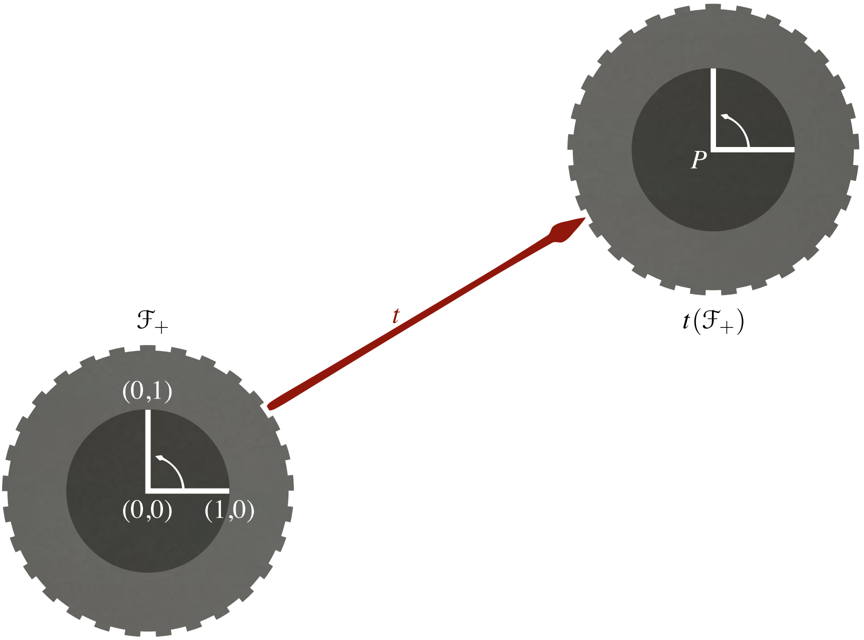
Associated to any transformation t is the vector that is represented by directed segments of the form $Pt(P)$ with initial point P and terminal point $t(P)$. That vector is both defined by and defines t. It is called the *translation vector* of t. It is almost always more convenient and natural to think about a translation in terms of its translation vector rather than as a composition of reflections. For instance, if you think of a translation t as a composition of reflections, it might not be that clear that t has no fixed points. If you think of that translation in terms of its translation vector, it is clear that no point P can be fixed by t, since $Pt(P)$ is always a directed segment with two distinct endpoints.

The transport of orientation

An *orthonormal frame* $\mathcal{F} = \{PP_x, PP_y\}$ is an ordered pair of perpendicular, unit length segments that share a common endpoint. One such frame, $\mathcal{F}_+$, centered at the origin with

$$P = (0,0) \quad P_x = (1,0) \quad P_y = (0,1),$$

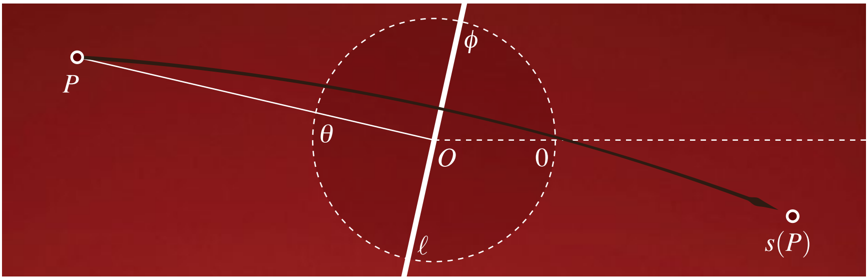
is at the very heart of the coordinate system. There is another such frame, $\mathcal{F}_-$, that shares the same first segment as $\mathcal{F}_+$, but that has $P_y = (0,-1)$. In general, any frame can be viewed as a way to represent information about orientation, the distinction between clockwise and counterclockwise. To this point, we have only made that choice at the origin: in $\mathcal{F}_+$, the directed minor arc from P_x to P_y points in the counterclockwise direction; in $\mathcal{F}_-$, the directed minor arc from P_x to P_y points in the clockwise direction. But translation now provides a vehicle to propagate that choice consistently across the rest of the plane. For any point P, let t be the translation that maps the origin to P. Then $t(\mathcal{F}_+)$ is a frame centered at P indicating the counterclockwise direction and $t(\mathcal{F}_-)$ is a frame centered at P indicating the clockwise direction.



Rotations

The illustrations at the start of this lesson suggest that when ℓ_1 and ℓ_2 intersect, the rotation $r = s_2 \circ s_1$ acts by turning points around the intersection point O. To measure the effect of this turning, we need to establish an angular coordinate system around O (just as we established a linear coordinate system on $\ell_\perp$ when ℓ_1 and ℓ_2 were parallel). Choose a ray with endpoint O– this marks the "zero angle"– and an orientation (clockwise or counter-clockwise). After making those choices, every ray from O will form an angle with r and we can then associate each point on the ray with that angle measure. Before attempting two reflections, let's back up and try to understand how the angular coordinates of a point behave when hit with just one reflection s across a line ℓ. Pick a point O on ℓ, and set up an angular coordinate system as described. Let P be an arbitrary point that is not on ℓ. Then label

θ: the angular coordinate at P
ϕ: the angular coordinate of one of the rays from O that make up ℓ.



The two choices of ϕ will be of the form θ and $\pi + \theta$, but as far as this calculation goes, it makes no difference which one you pick. The angle between ℓ and OP has a measure of $|\phi - \theta|$. Since isometries preserve angle measure and the whole line ℓ is fixed by s, the angle between ℓ and $Os(P)$ also has a measure of $|\phi - \theta|$. That severly limits the possibilities for the angular coordinates of $s(P)$:

$$\phi + |\phi - \theta| = \begin{cases} 2\phi - \theta & \text{if } \phi - \theta \geq 0 \\ \theta & \text{if } \phi - \theta < 0 \end{cases}$$

$$\phi - |\phi - \theta| = \begin{cases} \theta & \text{if } \phi - \theta \geq 0 \\ 2\phi - \theta & \text{if } \phi - \theta < 0. \end{cases}$$

Since P is not on ℓ, it is not fixed, and therefore $s(P)$ will not be at angle θ. The only other possibility, then, is that $s(P)$ is at angle $2\phi - \theta$. Furthermore, this formula still holds when P is on ℓ. In that case, P is fixed, so $s(P)$ should also be at angle θ. That is indeed what the formula indicates: if P is on ℓ, then $\phi = \theta$, and so $2\phi - \theta = \theta$. Now let's take that formula and use it to figure out what happens when we compose two intersecting reflections.

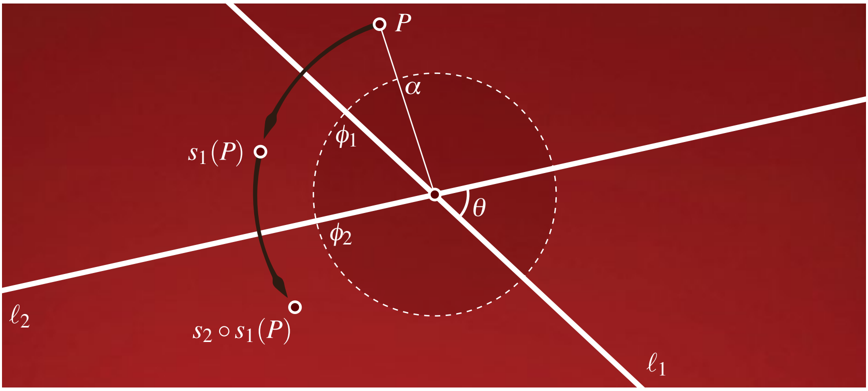
THM: PROPERTIES OF A ROTATION

Suppose that r is the rotation $s_2 \circ s_1$ where s_1 and s_2 are reflections across lines ℓ_1 and ℓ_2 that intersect at a point O at an angle of θ to one another. For any point P, $r(P)$ is located

1) on the circle centered at O that passes through P
2) so that OP and $Os(P)$ form an angle with measure 2θ,
3) in the direction indicated by the arc from ℓ_1 to ℓ_2.

Proof. Since $s_2 \circ s_1$ preserves distances and O is a fixed point, the distance from O to $s(P)$ is the same as the distance from O to P. That places $s(P)$ on the circle centered at O passing through P. Now where precisely is it on that circle? As in the discussion above, set up an angular coordinate system centered at O. Mark these coordinates:

α: the angular coordinate for P,
ϕ_1: the angular coordinate for ℓ_1,
ϕ_2: the angular coordinate for ℓ_2.

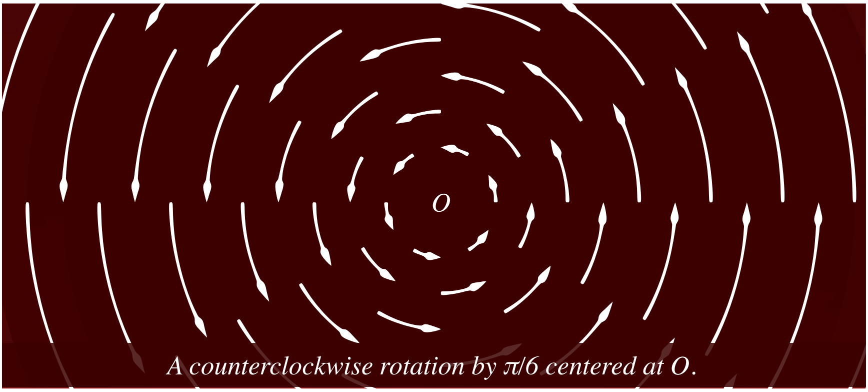


The intersection of ℓ_1 and ℓ_2 forms two vertical angle pairs. It is helpful to make the clockwise/counterclockwise choice so that the directed angle from ℓ_1 to ℓ_2 is the smaller of those two pairs (if ℓ_1 and ℓ_2 intersect at right angles, then it doesn't matter which orientation you choose). According to the previous discussion, $s_1(P)$ will have the coordinate $2\phi_1 - \alpha$. Then $s_2(s_1(P))$ will have the angular coordinate

$$2\phi_2 - (2\phi_1 - \alpha) = \alpha + 2(\phi_2 - \phi_1) = \alpha + 2\theta.$$

Therefore OP and $Os(P)$ do form an angle of 2θ, measured in the direction from ℓ_1 to ℓ_2. □

It is generally just a lot more convenient to think of a rotation in terms of the angle 2θ, the *rotation angle*, and the fixed point, the *center of rotation*, rather than as a composition of reflections. For instance, by thinking of a rotation in terms of its rotation angle and center, it is clear that a rotation only has one fixed point– the center of rotation.



A counterclockwise rotation by π/6 centered at O.

This viewpoint also gives a good perspective on just how common rotations are. The proof of the following result is left to the reader.

THM: THERE ARE JUST ENOUGH ROTATIONS

For a given point O and angle measure $0 < \theta < 2\pi$, there is exactly one clockwise rotation and exactly one counterclockwise with rotation center O and rotation angle θ. When $\theta = \pi$, the clockwise and counterclockwise rotations coincide (this is called a half-turn).

The analytic viewpoint

From the analytic point of view, translations are the simplest of the isometries. If we break the translation vector of a translation T down into a horizontal component h and a vertical component k, then

$$T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x+h \\ y+k \end{pmatrix}.$$

The equations for rotations are a little more challenging. In fact, for now, let's restrict our attention to rotations that are centered at the origin.

EQN: ROTATION AROUND THE ORIGIN

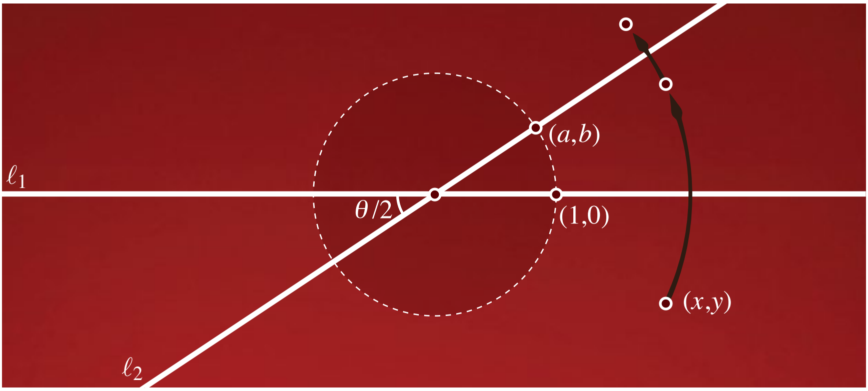
The analytic equation for a rotation r around the origin by an angle θ in the counterclockwise direction is

$$r\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}.$$

Proof. We can realize this rotation as a composition of two reflections across lines through the origin. For convenience sake, let's choose the reflections s_1 across ℓ_1 and s_2 across ℓ_2, where:

ℓ_1 is the x-axis and

ℓ_2 forms an angle of $\theta/2$ (counterclockwise) with the x-axis



Then $s_2 \circ s_1$ will be a rotation by an angle of $2 \cdot \theta/2 = \theta$. In the last lesson, we found out that equations for these types of reflections take the form

$$s\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a^2 - b^2 & 2ab \\ 2ab & b^2 - a^2 \end{pmatrix}$$

where (a,b) marks the intersection of the line and the unit circle. We can put that equation to good use now. The first line intersects the unit circle at $(1,0)$, so

$$s_1\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}.$$

The second line intersects the unit circle at $(\cos\theta/2, \sin\theta/2)$, and so

$$s_2\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos^2(\theta/2) - \sin^2(\theta/2) & 2\cos(\theta/2)\sin(\theta/2) \\ 2\cos(\theta/2)\sin(\theta/2) & \sin^2(\theta/2) - \cos^2(\theta/2) \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}.$$

We can use the double angle formulas to rewrite

$$\begin{aligned} \cos^2(\theta/2) - \sin^2(\theta/2) &= \cos(\theta), \\ 2\cos(\theta/2)\sin(\theta/2) &= \sin(\theta), \end{aligned}$$

which simplifies the matrix considerably to

$$s_2\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}.$$

To compute the composition of the transformations, just multiply the matrices:

$$\begin{aligned} r\begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \\ &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}. \end{aligned}$$

Exercises

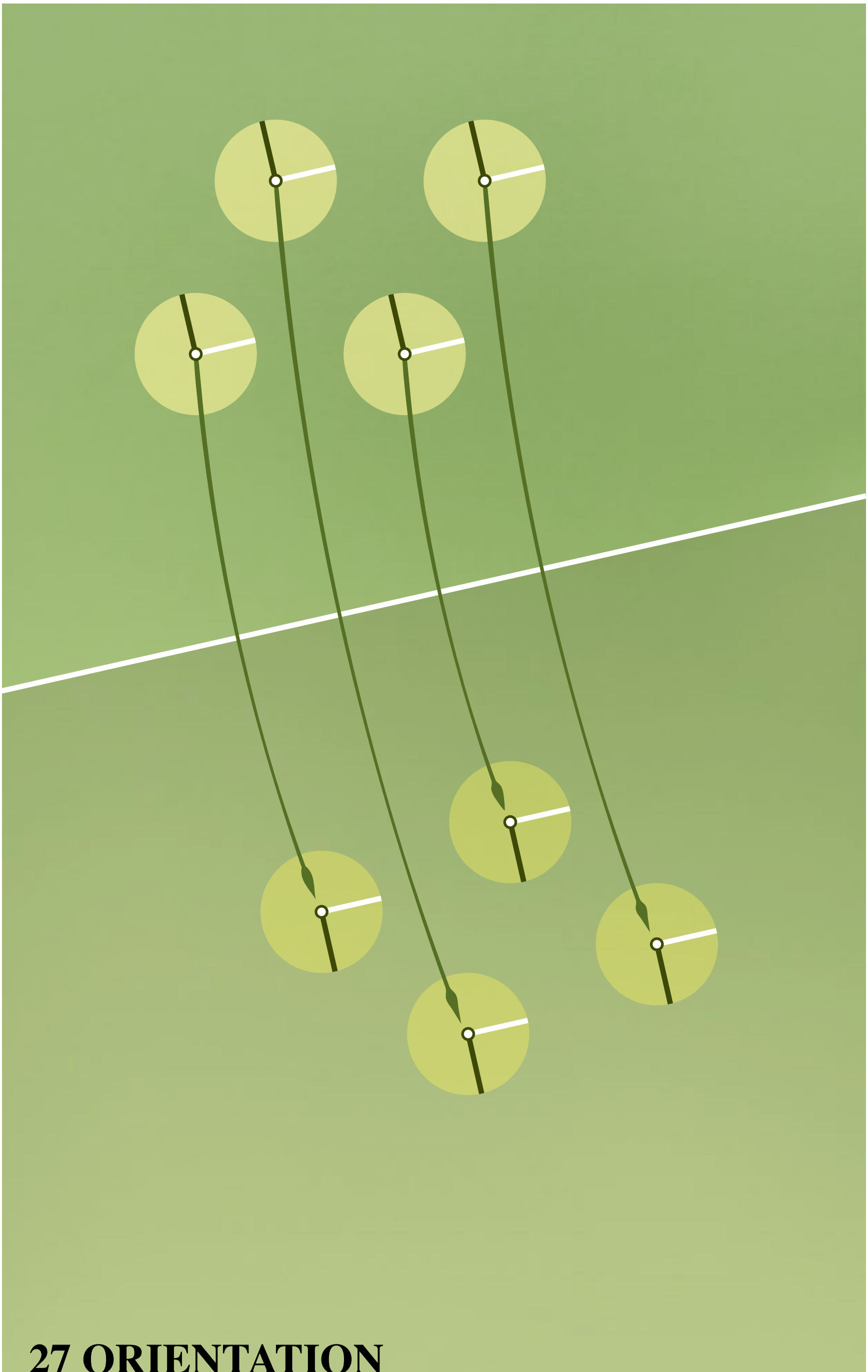
1. Prove that every isometry T can be written as a composition $t_1 \circ t_2$ where t_1 is a translation (or the identity) and t_2 is an isometry that fixes the origin.

2. Find the analytic equations for reflections across lines $x = a$ and $x = b$. Then verify that the composition of those reflections has the form of a translation.

3. Suppose that r_1 and r_2 are counterclockwise rotations about the origin, by angles of θ_1 and θ_2 respectively. Working from the matrix equations for r_2 and r_1, show that the matrix equations for $r_2 \circ r_1$ have the form of a rotation or the identity.

4. Suppose that ℓ is an invariant line of a rotation r. That is, if P is any point on ℓ, then $r(P)$ is also on ℓ. Show that ℓ passes through the center of rotation and the angle of rotation is π (r is then called a half-turn).

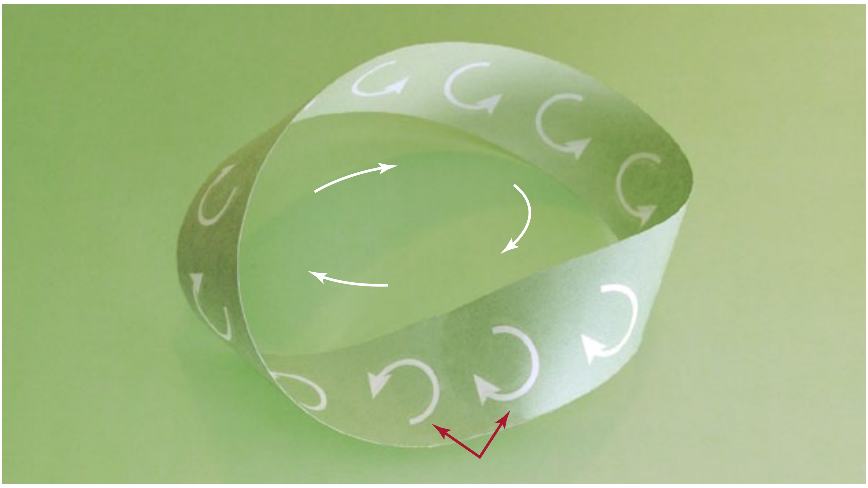
5. Take a vector $\langle a,b \rangle$. Let S be the set consisting of the identity isometry and all translations whose translation vectors have the form $\langle ma,nb \rangle$. Show that the composition of two elements of S is an element of S. Show that the inverse of an element of S is an element of S. This makes S a subgroup of the group of isometries.



27 ORIENTATION

MIND YOUR p'S AND q'S.

Earlier, we used translations to transport orientation (clockwise versus counterclockwise) from the origin to the rest of the plane. This is not a completely trivial issue because not all surfaces can be oriented consistently like this. The most famous non-orientable surface is the Möbius strip. It is formed by taking a strip, giving it a half-twist, and then joining the two ends. A frame F on the Möbius strip can be translated from one point to another in two different ways, t_1 and t_2, and the resulting frames $t_1(F)$ and $t_2(F)$ are *not* oriented the same way. Fortunately, we do not have this problem in the plane because there is only one translation from one point to another.



One lap aroud the Möbius strip flips orientation.

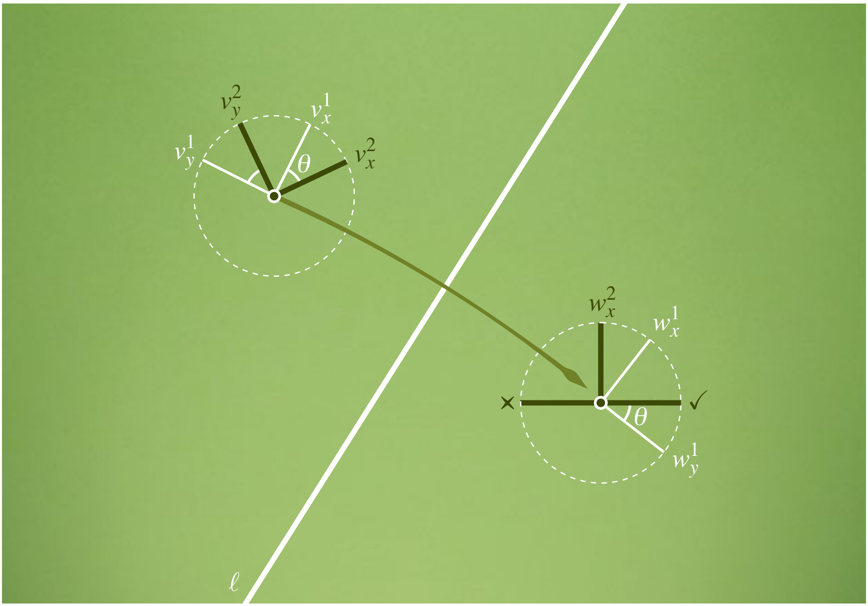
In this lesson we look at how isometries interact with orientation. Since all isometries are compositions of reflections, we can begin the process by looking at reflections. Once we understand their effect on orientation, the rest is pretty easy.

LEM: CONSISTENCY OF ORIENTATION

Let s be a reflection. If F_1 and F_2 are frames at a point P that are oriented in the same direction, then $s(F_1)$ and $s(F_2)$ are frames at a point $s(P)$ that are oriented in the same direction.

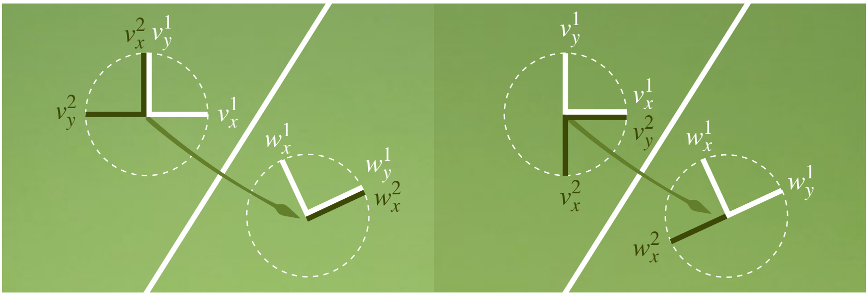
Proof. A frame is composed of two length one segments that form a right angle. Since a reflection changes neither the length of a segment, nor the angle between a pair of them, the reflection of a frame is a frame. The issue of orientation is a bit more delicate. Let's suppose that F_1 and F_2 are oriented in the same direction, and then compare the orientations of $s(F_1)$ and $s(F_2)$. To do that, label the components of each frame:

$$\begin{aligned} F_1 &= \{v_x^1, v_y^1\} &\qquad s(F_1) &= \{w_x^1, w_y^1\} \\ F_2 &= \{v_x^2, v_y^2\} &\qquad s(F_2) &= \{w_x^2, w_y^2\} \end{aligned}$$



The right and wrong choice for w_y^2

Let θ denote the angle between v_x^1 and v_x^2. Since F_1 and F_2 are oriented in the same direction, θ is also the angle between v_y^1 and v_y^2. Now move on over to the frames after the reflection. The angle between w_x^1 and w_x^2 still has to be θ. And the angle between w_y^1 and w_y^2 has to be θ. Remember that the orientation of a frame is essentially a choice: given the first segment, there will always be two directions perpendicular to it. If we make the wrong choice for w_y^2 (that is, we orient $s(F_1)$ and $s(F_2)$ oppositely), then the angle between w_y^1 and w_y^2 will be $\pi - \theta$, not θ. Generally, speaking, that cannot happen, and that is sufficient to show that $s(F_1)$ and $s(F_2)$ must be oriented in the same direction.



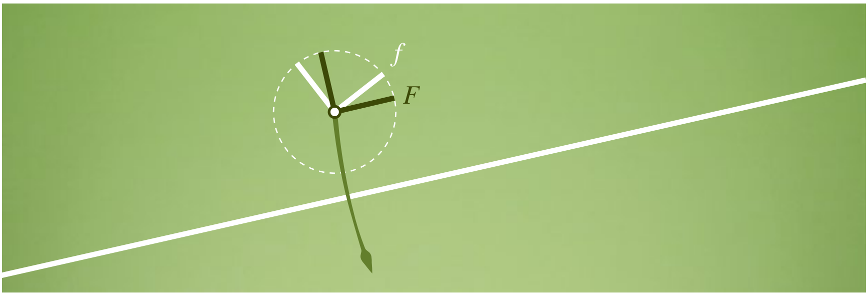
There is still ambiguity in one case though: the previous argument hinged upon the fact that the angle between w_y^1 and w_y^2 must be θ, not $\pi - \theta$, but those two angle measures could be the same, when $\theta = \pi - \theta$, so when $\theta = \pi/2$. The illustrations above show the possible scenarios. In the first, the wrong choice of w_y^2 maps two distinct segments, v_x^1 and v_y^2, to the same segment, which is not permitted since a reflection is one-to-one. In the second, the wrong choice of w_y^2 maps one segment, $v_x^1 = w_y^2$ to two different segments– again not permitted since a reflection is well-defined. □

THM: REFLECTIONS REVERSE ORIENTATION
A reflection s reverses the orientation of any frame F.

Proof. Let s be any reflection and ℓ be the fixed line of that reflection. While the theorem itself claims that s reverses *any* frame, the previous lemma gives us a way to focus on a particularly well-suited subset. That subset consists of frames of the form $F = \{v_x, v_y\}$ where

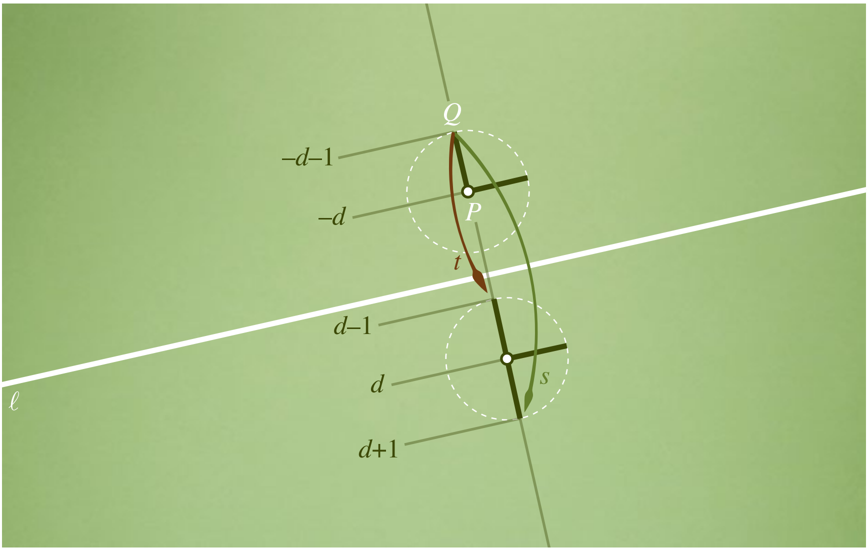
1) v_x is parallel to ℓ (or runs along ℓ), and
2) v_y is perpendicular to ℓ, pointing away from it.

At any point P, there are two frames that meet these conditions, one oriented in the clockwise direction, the other in the counterclockwise direction. Therefore, for every frame f we can find a frame F of the form described above which has the same orientation as f. According to the previous lemma, $s(f)$ and $s(F)$ must have the same orientation, so if we can show that F and $s(F)$ are oriented oppositely, then that will mean that f and $s(f)$ are too. Essentially, the previous lemma lets us rotate f into the more convenient position of F.



To see whether s really does reverse orientation, we need to compare $s(F)$ to $t(F)$, where t is the translation from the point P to its reflection $s(P)$. Note that t will map v_x to $s(v_x)$– that is the advantage of this particular subset of frames– so the question of whether $s(F)$ and $t(F)$ have the same orientation really just comes down to a comparison of $s(v_y)$ and $t(v_y)$. To that end, label the endpoint of v_y as Q. Let d be the distance from ℓ to P so that t is a translation by $2d$. Set up a coordinate axis on the line through v_y and $s(v_y)$ so that ℓ intersects it at the origin and the ray $v_ys(v_y)\rightarrow$ points in the positive direction. Compare coordinates:

$$\begin{aligned}
&Q : -d-1\\
&s(Q) : d+1\\
&t(Q) : (-d-1)+2d = d-1.
\end{aligned}$$



If s were to preserve the orientation of F, then $s(Q)$ and $t(Q)$ would be the same so $s(Q)$ and $t(Q)$ would have the same coordinates:

$$d+1=d-1 \implies 2=0.$$

This cannot happen. □

DEF: ORIENTATION PRESERVING/REVERSING

An isometry is *orientation-preserving* if it maps clockwise frames to clockwise frames and counterclockwise frames to counterclockwise frames. A isometry is *orientation-reversing* if it swaps clockwise and counterclockwise frames.

Because reflections are orientation reversing, and because every isometry is a composition of reflections, determining what an isometry does to orientation is essentially just a matter of counting flips.

COR: ORIENTATION AND COMPOSITION

A composition of two orientation-preserving maps is orientation preserving; a composition of two orientation-reversing maps is orientation preserving; a composition of one orientation-preserving map and one orientation-reversing map is orientation-reversing.

COR: CLASSIFICATION OF ISOMETRIES BY ORIENTATION

Translations, rotations and the identity map are orientation-preserving. They are the only orientation-preserving isometries.

Let's now recap our progress in the classification of isometries.

# of ref$^{\text{n}}$s	isometry	orientation	fixed pts
1	reflection	reversing	line
2	identity	preserving	all
	translation	..	none
	rotation	..	point
3	?	reversing	?

In the next lesson we find what goes in place of those questions marks.

Exercises

1. Show that if τ is an orientation-preserving isometry which fixes two points, then it must be the identity. Show that if τ is an orientation-preserving isometry which has at least one fixed one point, and at least one non-fixed point, then it must be a rotation.

2. Let τ_1 be a counterclockwise rotation by $\pi/2$ about the origin. Let τ_2 be a counterclockwise rotation by $\pi/2$ about the point $(1,0)$. Show that $\tau_1 \circ \tau_2$ is a rotation.

28 GLIDE REFLECTIONS

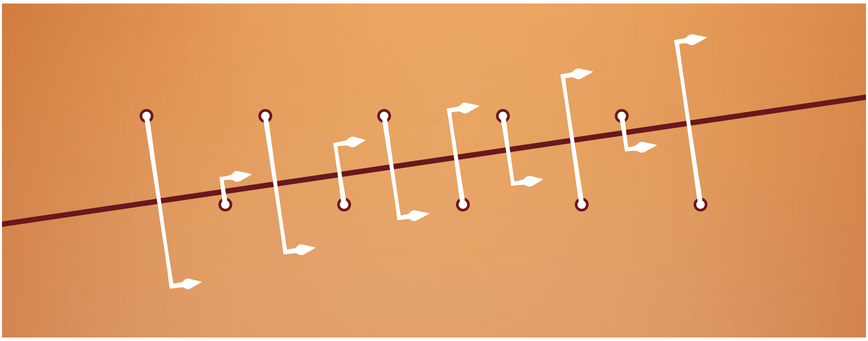
So now let's look at a composition of three reflections. The first two reflections will get us to either the identity map, a translation, or a rotation. We are going to add another reflection to that. Composing a reflection with the identity will, of course, give a reflection. What about composing a reflection with a translation or a rotation? That is the subject of this lesson.

Glide reflections

Straight off, we can see that, yes, there is a fundamentally new type of isometry here. Just take a reflection s across a line ℓ followed by a translation t whose translation vector is parallel to ℓ. The composition $t \circ s$ swaps the two sides of ℓ and translates along ℓ. Therefore it has no fixed points. We have only seen one type of isometry that has no fixed points so far– a translation. But this isometry, a composition of three reflections, will be orientation-reversing, so it can't be a translation.

DEF: GLIDE REFLECTION

A glide reflection is a composition of a translation t followed by a reflection s across a line that is parallel to the translation vector.



The path of a few points under a glide reflection.

In general, you can't just switch the order that you compose isometries and expect to get the same answer. But the s and t that make up a glide reflection are interchangeable.

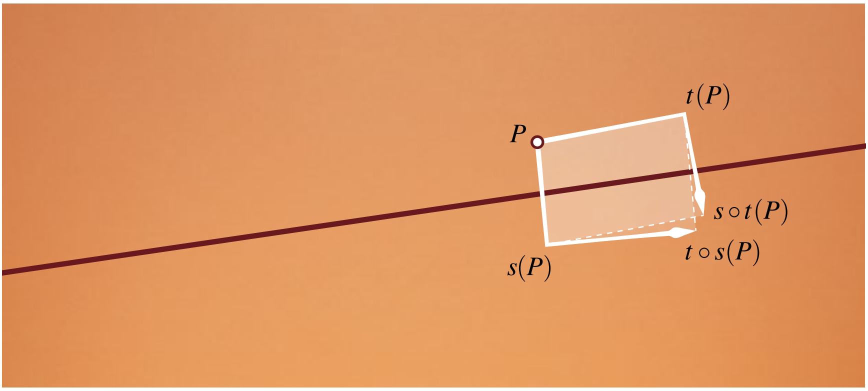
LEM: SWAPPING GLIDE COMPONENTS
Let s be a reflection across a line ℓ and let t be a translation parallel to ℓ. Then $s \circ t = t \circ s$.

Proof. If P is a point on the reflecting line ℓ, then so is its translation $t(P)$, and in that case, the reflection has no effect on either one of them, so

$$s \circ t(P) = t(P) = t \circ s(P).$$

Now suppose that P is not on ℓ. In that case, let's compare the two quadrilaterals

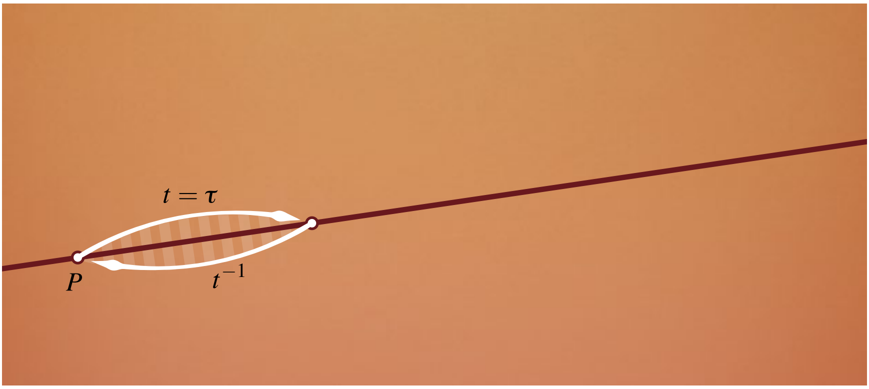
1) with vertices P, $s(P)$, $t(P)$, and $s \circ t(P)$;
2) with vertices P, $s(P)$, $t(P)$, and $t \circ s(P)$.



The quadrilaterals share two sides, $Ps(P)$ and $Pt(P)$, and since the reflection and translation are perpendicular motions, both quadrilaterals have right angles at three of the four vertices, at P, $s(P)$, and at $t(P)$. That of course means that the fourth angle must also be a right angle, and so the two quadrilaterals are in fact rectangles. Well, there is only way to build a rectangle given two of its adjacent sides. Therefore $s \circ t(P)$ and $t \circ s(P)$ must be the same. □

For what we are going to do, we need an easy way to recognize glide reflections in the field. The key is that if you narrow your focus down to just the reflecting line, a glide reflection looks just like a translation. I call this line of reflection the "glide line", and the distance of translation along that line the "glide distance".

LEM: RECOGNIZING GLIDE REFLECTIONS I
Let τ be an isometry, and suppose that there is a line ℓ and a translation t so that $\tau(P) = t(P)$ for all points P on ℓ. If $\tau \neq t$, then τ is a glide reflection.



Proof. Look at the effect of the composition of τ and t^{-1} on the points of the line ℓ:

$$t^{-1} \circ \tau(P) = t^{-1} \circ t(P) = P.$$

It fixes all the points on ℓ. Assuming $\tau \neq t$, $t^{-1} \circ \tau$ cannot be the identity map. The only other isometry that fixes an entire line is a reflection. Therefore $t^{-1} \circ \tau = s$ where s is the reflection across the line ℓ, and so $\tau = t \circ s = s \circ t$ is a glide reflection. □

By itself, that lemma is already useful, but we can punch it up even more by combining it with the next one.

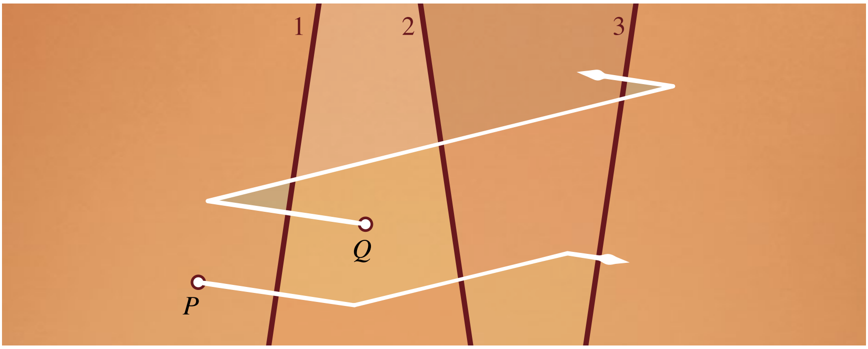
LEM: RECOGNIZING GLIDE REFLECTIONS II
Let τ be an isometry and let t be a translation. Suppose that for two distinct points P and Q, $\tau(P) = t(P)$ and $\tau(Q) = t(Q)$. Then $\tau = t$ for all points on the line $\leftarrow PQ \rightarrow$.

Proof. Again look at the composition $t^{-1} \circ \tau$:

$$t^{-1} \circ \tau(P) = P \qquad t^{-1} \circ \tau(Q) = Q.$$

Since $t^{-1} \circ \tau$ fixes these two points, it must fix all points on $\leftarrow PQ \rightarrow$. That is, $t^{-1} \circ \tau(R) = R$ for all points R on $\leftarrow PQ \rightarrow$. Composing t with both sides of this equation, $\tau(R) = t(R)$ for all R points on $\leftarrow PQ \rightarrow$. Therefore τ and t agree for all points on $\leftarrow PQ \rightarrow$. □

By combining those two lemmas we get: an orientation-reversing isometry that agrees with a translation on two distinct points must be a glide reflection.



A composition of three reflections

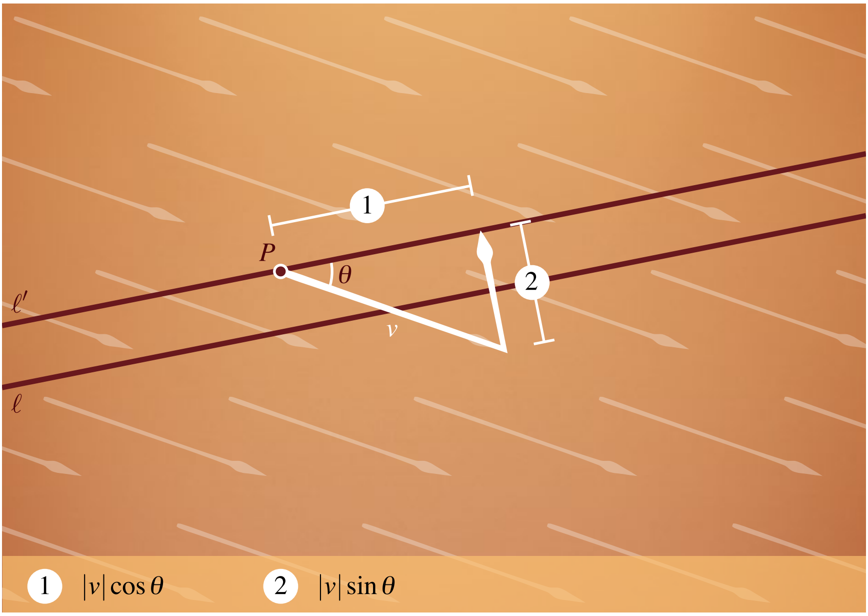
Compositions of three reflections

Let's start the hunt by looking at what happens when we compose a translation and a reflection. If the translation is parallel to the line of reflection, of course, then that is the very definition of a glide reflection. But what if the translation is not along the reflecting line?

THM: TRANSLATION AND REFLECTION

Let t be a translation with translation vector v, let s be a reflection across line ℓ, and let θ be the angle between v and ℓ. Then $s \circ t$ is a glide reflection whose glide line is parallel to ℓ, at a distance $(|v| \sin\theta)/2$ from ℓ, and whose glide distance is $|v| \cos\theta$.

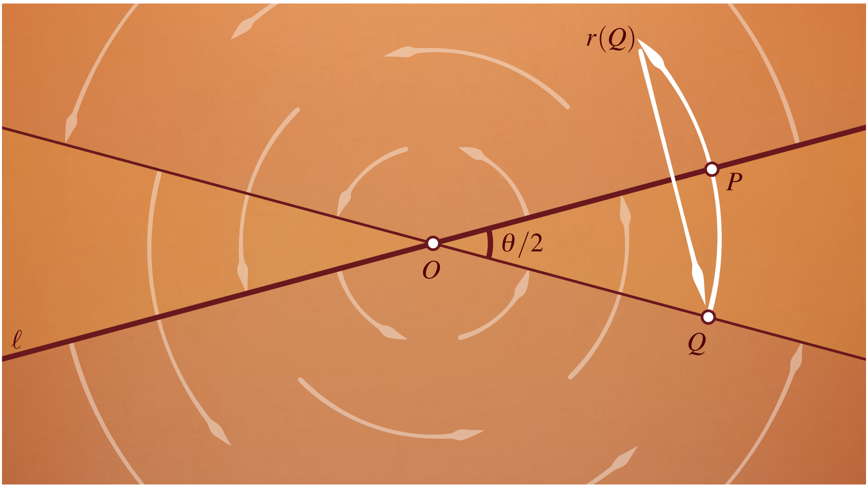
Proof. As the previous lemmas suggest, if we want to show that $s \circ t$ is a glide reflection, then we need to find its glide line. The best way to do that is to experiment around with the translation-reflection combination. You are looking for a line along which $s \circ t$ acts as a translation– first t will move the points off the glide line, and then s will move them back, shifted from their original location.



It turns out that the glide line ℓ' is a line that runs parallel to ℓ. It is on the opposite side of ℓ from the direction that v points, and is separated from ℓ by a distance of $(|v|\sin\theta)/2$. Let's verify that ℓ really is the glide line, and therefore that $s \circ t$ is a glide reflection. Take a point P on ℓ. We can break the translation $t(P)$ down into two steps: first a translation by $|v|\cos\theta$ along ℓ', and then a translation by $|v|\sin\theta$ perpendicular to ℓ. The second translation means that $t(P)$ is located on the opposite side of ℓ from P, at a distance of $(|v|\sin\theta)/2$ from ℓ. Therefore, when we apply the reflection s to $t(P)$, the result $s \circ t(P)$ is back on the line ℓ, but shifted up from P a distance of $|v|\cos\theta$. All the points on ℓ exhibit this behavior, so $s \circ t$ acts as a translation along ℓ'. Since $s \circ t$ is orientation-reversing, it cannot be a translation. According to the lemma above, it must be a glide reflection. □

We have taken care of combinations of a translation with a reflection–what happens when we combine a reflection and a rotation? There really are two scenarios, depending upon whether or not the reflecting line passes through the center of rotation. The scenario where the reflecting line *does* pass through the rotation center is a little bit easier, so let's start with that one.

THM: ROTATION AND REFLECTION I
Let r be a rotation by an angle of θ centered at a point O, and let s be a reflection across a line ℓ that passes through O. Then $s \circ r$ is a reflection across a line that passes through O and forms a (signed) angle of $-\theta/2$ with ℓ.



Proof. First notice that O is a fixed point of $s \circ r$. If we can find just one other fixed point, then that will mean that the entire line between them is fixed. As a result, $s \circ r$ will either be the identity or a reflection, and $s \circ r$ can't be the identity since it is orientation-reversing. So really, this is just a matter of finding a point where the action of the reflection undoes the action of the rotation. Take a point P on ℓ other than O and rotate it by $-\theta/2$ about O (that is, rotate it in the opposite direction from r) to a point Q. This point Q is the one we want: $Or(Q)\rightarrow$ will form an angle of $\theta/2$ with ℓ. Reflecting back across ℓ, $Os \circ r(Q)\rightarrow$ will again form an angle of $-\theta/2$ with ℓ. Since its distance from O remains unchanged throughout this whole operation, that means $s \circ r(Q) = Q$. □

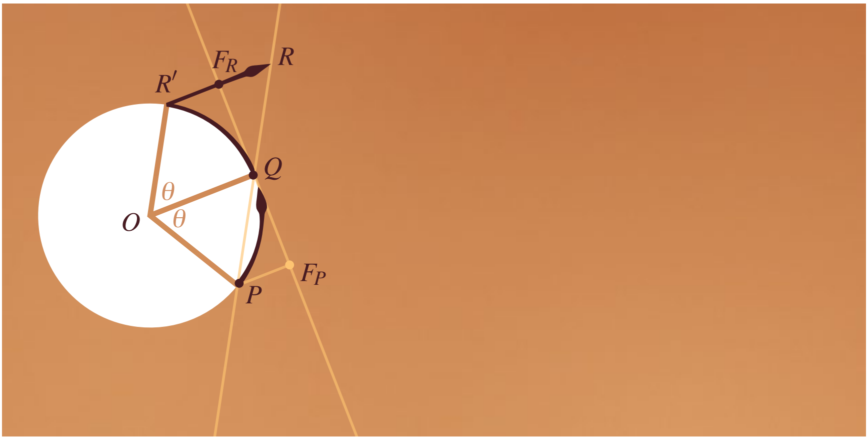
If the reflecting line *does not* pass through the center of rotation, then the situation is more complicated.

THM: ROTATION AND REFLECTION II

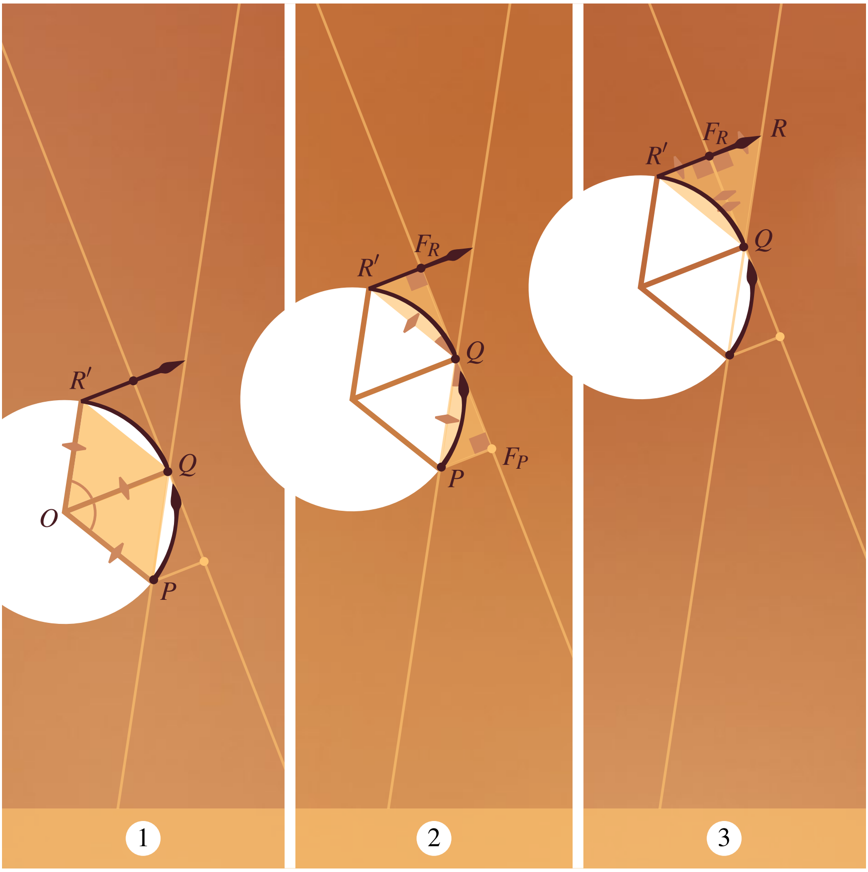
Let r be a rotation by an angle of θ centered at a point O, let s be a reflection across a line ℓ that does not pass through O, and let Q be the closest point on ℓ to O. Then $s \circ r$ is a glide reflection along a line that passes through Q at an angle of $\theta/2$ to ℓ.

Proof. This theorem claims that $s \circ r$ is a glide reflection, and if that is the case, then we need to find its glide line. Let's use the following labels:

$P = (s \circ r)^{-1}(Q)$
$R = (s \circ r)(Q)$
$R' = r(Q)$
F_P: the foot of perpendicular from P to ℓ
F_R: the foot of perpendicular from R to ℓ



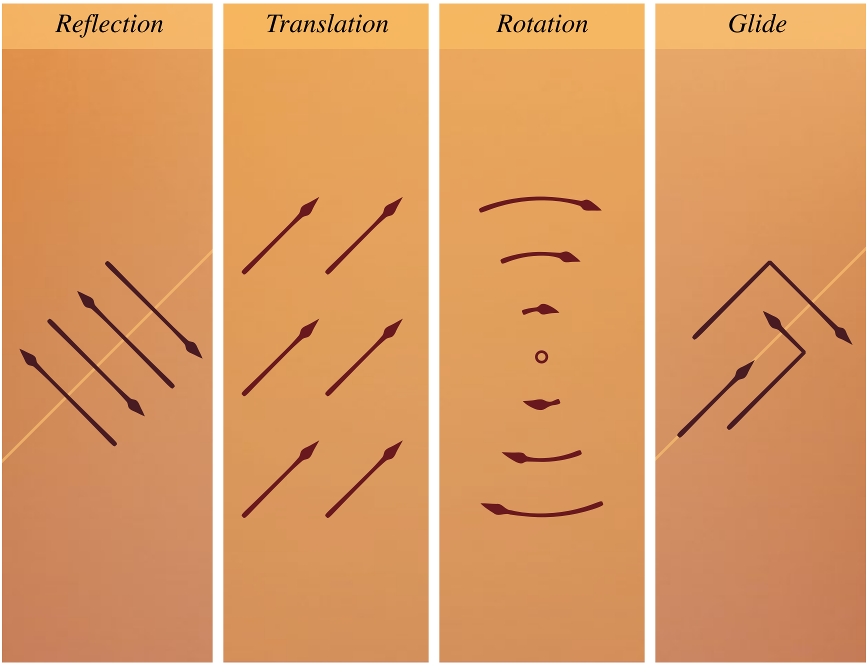
Note that the labels are set up so that $s \circ r$ will move P to Q and Q to R. It turns out that the glide line is the line through P, Q, and R. Now, ultimately there are a two things to do to show that. First, we need to show that P, Q, and R are in fact collinear. Second, we need to show that $s \circ r$ moves P and Q in the same way that a translation does– that it moves P and Q the same distance in the same direction– essentially this means we need to show that $|PQ| = |QR|$. If we can show both of those things, then that means $s \circ r$ will have to be a glide reflection.



We *can* do it– we just need to use some congruent triangles.

1. By S·A·S, $\triangle OQP \simeq \triangle OQR'$.
2. By A·A·S, $\triangle PQF_P \simeq \triangle R'QF_R$.
3. By S·A·S, $\triangle R'QF_R \simeq \triangle RQF_R$.

Therefore, $\triangle PQF_P$ and $\triangle RQF_R$ are congruent. Their corresponding angles $\angle PQF_P$ and $\angle RQF_R$ are congruent, and since F_P, Q, and F_R are collinear, that means that P, Q, and R must be collinear too. Furthermore, by comparing the lengths of the hypotenuses of these congruent triangles, $|PQ| = |QR|$. Therefore $s \circ r$ acts like a translation for the two points P and Q. It follows that $s \circ r$ acts like a translation for all points on that line. Since $s \circ r$ is not a translation (it is orientation-reversing), it must be a glide reflection. □



The four non-identity Euclidean isometries.

That's it! We have looked at all possible combinations of at most three reflections and seen the following types of isometries: the identity, reflections, translations, rotations, and glide reflections. Let's put it all together in a convenient table.

The Isometries of the Euclidean Plane.

# of ref^{n}s	isometry	orientation	fixed pts
1	reflection	reversing	line
2	identity	preserving	all
	transation	..	none
	rotation	..	point
3	glide reflection	reversing	none

Exercises

1. Give analytic equations for the glide reflection formed by reflecting across the line $y = mx$ and then translating a distance d along this line (choose the translation vector so that it points from the origin into the *first* quadrant).

2. We saw that the composition of a rotation and a reflection is a glide reflection if the center of rotation is not on the line of reflection. What is the glide distance in this case (in terms of the rotation center, the rotation angle, and the line of reflection)?

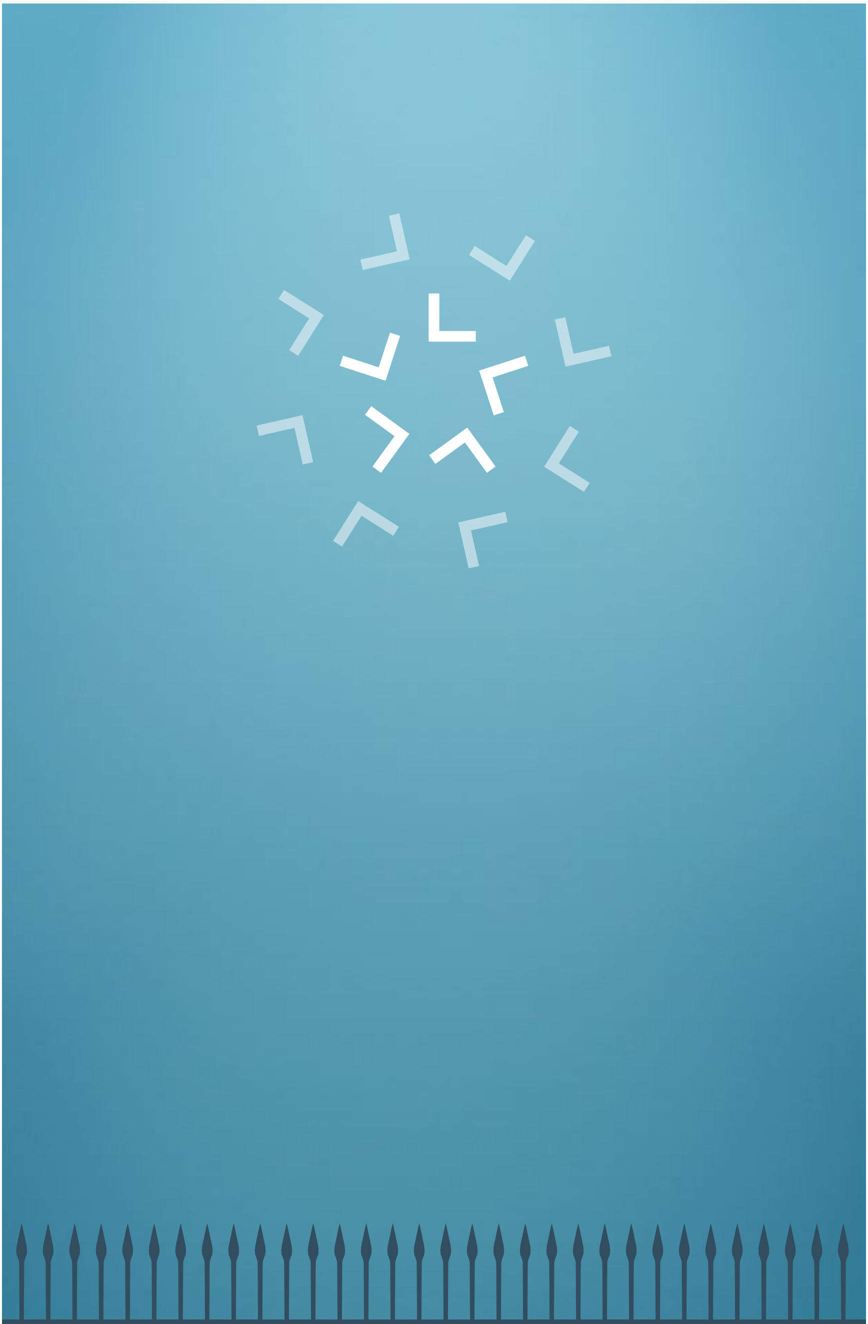
3. Let r be a counterclockwise rotation by $\pi/4$ about the origin. Let s be the reflection across the line $y = x + 1$. What is the equation of the glide line of the glide reflection $s \circ r$?

4. Let g be a glide reflection. What is the minimum number of points required to completely determine g (to find both its glide line and glide distance)?

5. Describe the isometries τ that satisfy the condition $\tau^2 = \mathrm{id}$. Describe the isometries that satisfy the condition $\tau^n = \mathrm{id}$ for $n > 2$.

6. Show that the composition of a glide reflection and reflection is either a rotation or a translation. Give specific examples in which each outcome occurs.

7. Show that the composition of two glide reflections is either the identity, a rotation, or a translation. Give specific examples in which each of these outcomes occurs.

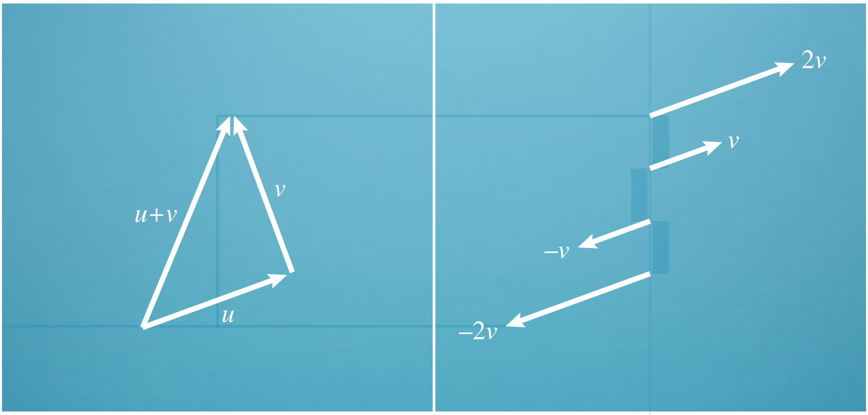


29 CHANGE OF COORDINATES

Vector arithmetic

In the lesson on translation and rotation, I introduced vectors, but did little more than define them. Let's take a more detailed look at vectors now. In general, a vector holds two pieces of information: a length and a direction. It is represented by a directed segment, and it is common to distinguish the two endpoints of that segment with the names "tail" and "head", so that the segment points from the tail to the head. There is one exception: the zero vector is a vector with length zero and no direction. You can think of it as the degenerate case that occurs when a segment shrinks all the way down to a point and the head and tail merge. It is common practice to conflate a vector with one of its representative directed segments, and there is generally no problem with that. For now, I think it is probably a good idea to maintain a little distance between the two: for this section I will write $\vec{v}$ for a vector, and v for one of its representative directed segments. Once we are out of this section, I will do as everyone else does, and just mix up the two notions.

One of the strengths of vectors is that they have an inherent arithmetic that points do not. Any two vectors can be added together using a "head-to-tail" procedure as follows. Given any two vectors $\vec{u}$ and $\vec{v}$, their sum $\vec{u}+\vec{v}$ is the vector which is represented by a directed segment $u+v$ that is defined as follows. Let u be any representative of $\vec{u}$ and let v be the representative of $\vec{v}$ whose tail is located at the head of u. Then $u+v$ is the directed segment from the tail of u to the head of v.



Vector addition

Scalar multiplication

Any vector $\vec{v}$ can be multiplied by any real number r. The resulting vector $r \cdot \vec{v}$ is represented by a directed segment that

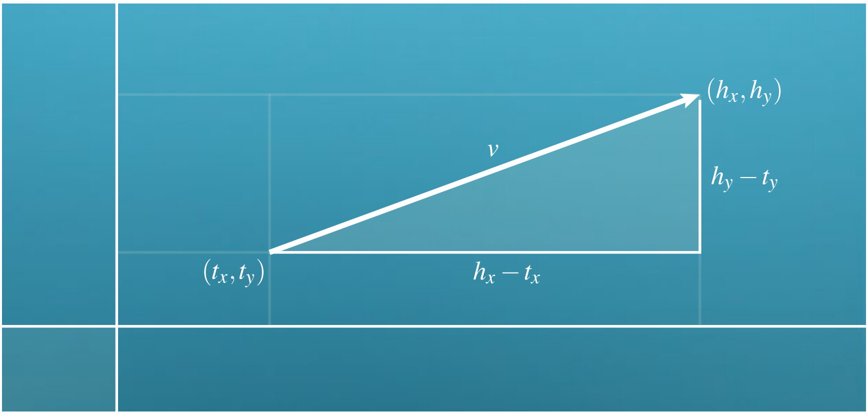
1) has the same tail as v and is on the same line as v,
2) has length $|r| \cdot |v|$, and
3) is in the same direction as v if $r > 0$ and in the opposite if $r < 0$.

Note that each of these calculations requires a choice of representatives. This raises a potential issue: these operations may not be well-defined– different choices for the representatives could conceivably lead to different answers. It's not too hard to see that this is not the case. I will leave it as an exercise.

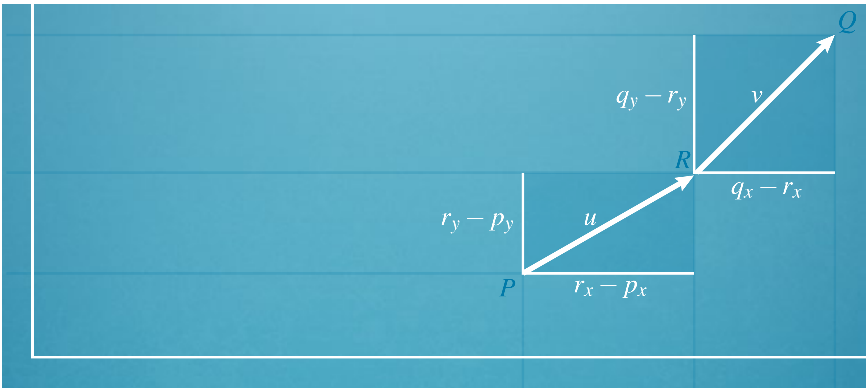
There is an analytic side of the story too. Let $\vec{v}$ be a vector represented by a directed segment v, and mark:

(t_x, t_y): the coordinates of the tail of v;
(h_x, h_y): the coordinates of the head of v.

Then $h_x - t_x$ is called the horizontal component or x-component of $\vec{v}$, and $h_y - t_y$ is called the vertical component or y-component of $\vec{v}$. Note that these values do not depend upon the choice of v. We write the vector $\vec{v}$ in terms of its components as $\vec{v} = \langle h_x - t_x, h_y - t_y \rangle$.



The horizontal and vertical components of a vector.



LEM: ADDITION

Let $\vec{u} = \langle u_x, u_y \rangle$ and $\vec{v} = \langle v_x, v_y \rangle$. Then $\vec{u} + \vec{v} = \langle u_x + v_x, u_y + v_y \rangle$.

Proof. Position u and v head-to-tail. Label the coordinates of the tail of u as (p_x, p_y), of the head of v as (q_x, q_y), and of the head of u, which is the tail of v, as (r_x, r_y). Then the horizontal component of $\vec{u} + \vec{v}$ is

$$q_x - p_x = (q_x - r_x) + (r_x - p_x) = u_x + v_x,$$

and the vertical component of $\vec{u} + \vec{v}$ is

$$q_y - p_y = (q_y - r_y) + (r_y - p_y) = u_y + v_y.$$

□

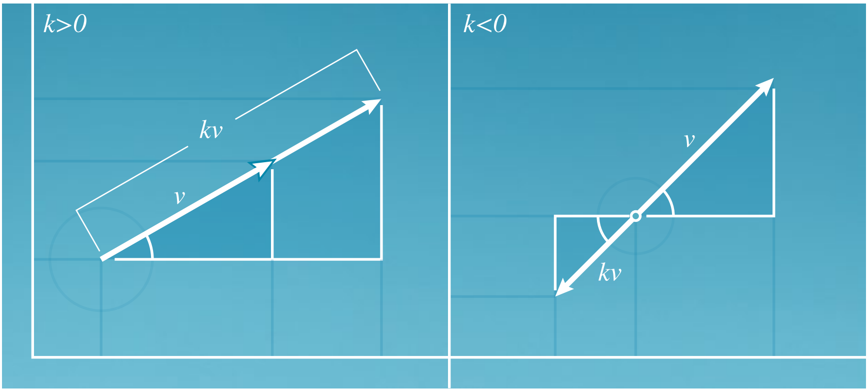
LEM: SCALAR MULTIPLICATION

Let $\vec{v} = \langle v_x, v_y \rangle$ and k be a real number. Then

$$k \cdot \vec{v} = \langle kv_x, kv_y \rangle.$$

Proof. From the previous part, we can break $\vec{v}$ down into two vectors, one containing the horizontal component, the other the vertical:

$$\vec{v} = \langle v_x, 0 \rangle + \langle 0, v_y \rangle.$$



These two vectors, together with $\vec{v}$ itself, form a right triangle. Similarly, we can form a right triangle from $k \cdot \vec{v}$ and its horizontal and vertical components. Now note that these two triangles are similar. Comparing the two hypotenuses, the (signed) scaling factor between those triangles is k. Scaling the legs by the same amount, $k \cdot \vec{v}$ has a horizontal component of kv_x and a vertical component of kv_y. □

THM: PROPERTIES OF VECTOR ARITHMETIC

The following are true for all vectors $\vec{u}, \vec{v}$ and $\vec{w}$ and for all real number k and l:

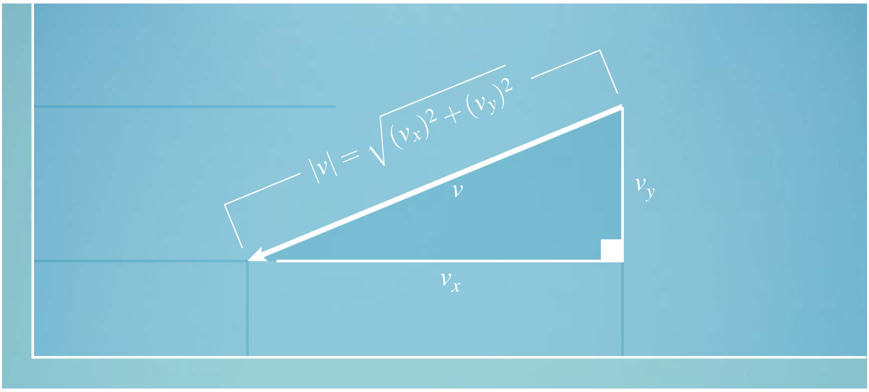
1. Additive associativity: $(\vec{u}+\vec{v})+\vec{w}=\vec{u}+(\vec{v}+\vec{w})$
2. Additive commutativity: $\vec{u}+\vec{v}=\vec{v}+\vec{u}$
3. Additive identity: the sum of the zero vector and $\vec{v}$ is $\vec{v}$
4. Additive inverse: every vector $\vec{v}$ has an additive inverse $\vec{w}$ so that $\vec{v}+\vec{w}$ is the zero vector
5. Distributive 1: $k(\vec{u}+\vec{v})=k\vec{u}+k\vec{v}$
6. Distributive 2: $(k+l)\vec{v}=k\vec{v}+l\vec{v}$
7. Multiplicative associativity: $kl(\vec{v})=k(l\vec{v})$
8. Multiplicative identity: $1(\vec{v})=\vec{v}$

These properties are really more linear algebra than geometry, so I will not take the time to verify them.

Vectors and points are not the same thing, so point coordinates (x,y) should not be equated with vector components $\langle x,y\rangle$. There is, however, a useful conduit between the two. If $\vec{v}=\langle x,y\rangle$, then the representative of $\vec{v}$ that has its tail at the origin will have its head at the point with coordinates (x,y). In fact, I have already used this correspondence: to be proper, the input of a matrix equation for an isometry is a vector $\langle x,y\rangle$, not a point's coordinates (x,y).

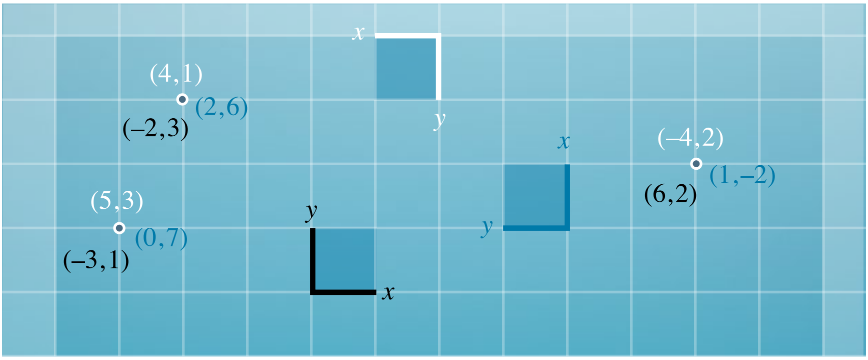
Before moving on, there is one more term to define. The *norm* (or length, or size, or magnitude) of a vector $\vec{v}$, written $|\vec{v}|$, is the length of any of its representative segments. Using the distance formula, the norm of a vector may be calculated from its components to be

$$|\langle v_x,v_y\rangle|=\sqrt{(v_x)^2+(v_y)^2}.$$



Change of coordinates

Our study of the analytic side of geometry began with choices about where to put the origin, and how to point the x- and y-axes. A frame provides that same information– the vertex of the frame is the origin, and the two segments v_x and v_y point in the directions of the positive x- and y-axes. In essence, then, each frame F determines a coordinate system C_F. In practice, there are times when it is convenient to switch from one coordinate system, say C_F, to another coordinate system, say C_G. To do that, we need to understand how a point's C_F coordinates are related to its C_G coordinates. As you might expect, the key to this is an isometry that maps the frame F to the frame G.



Coordinates of three points in three systems

There are a few more things that we need to know before we can proceed. First, we need to know that there is an isometry from F to G. Second, in the course of the proof, we will need to use a linearity property of matrices (that you may have seen in, say, a linear algebra course).

THM: THERE ARE JUST ENOUGH ISOMETRIES

There is a unique isometry from any frame to any other frame.

THM: THE LINEARITY OF MATRIX OPERATIONS

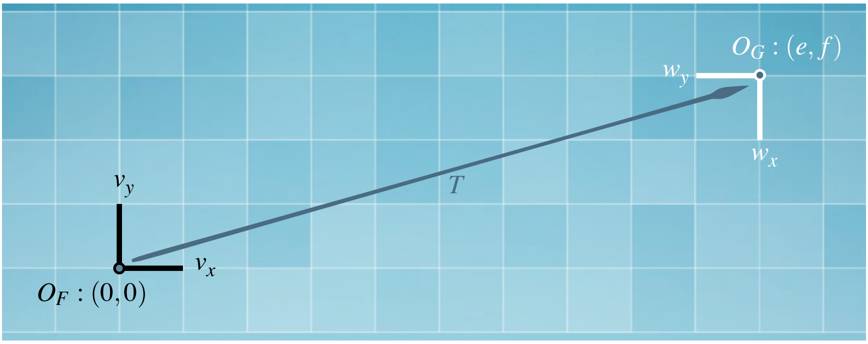
If M is a matrix, v_1 and v_2 are vectors, and k is a constant, then

1. $M(v_1+v_2) = Mv_1 + Mv_2$
2. $M(kv_1) = kM(v_1)$

I will leave the proofs of both of these results to you. For the first, you should be able to model your proof on the argument I gave in Lesson 24 where I computed a general form of the analytic equations of all isometries. For the second, you are only really obligated to deal with 2×2 matrices (since that is all we will be using), in which case the calculations are not hard at all. Now back to business.

THM: CHANGE OF COORDINATES

Let C_F and C_G be the coordinate systems determined by the frames F and G respectively, and let T be the isometry from F to G. Then the C_G coordinates of a point P are the same as the C_F coordinates of $T^{-1}(P)$.



Proof. Start by taking a look at the isometry T and its inverse. From our work on analytic isometries, we know quite specifically what forms the equations of T can have. In general, we can write

$$T\begin{pmatrix}x\\y\end{pmatrix} = M\begin{pmatrix}x\\y\end{pmatrix} + \begin{pmatrix}e\\f\end{pmatrix}$$

where M is some 2×2 matrix. Note that since we want to know about the C_F coordinates of $T^{-1}(P)$, this matrix M must be set up in the C_F coordinate system (for instance, if M represents a rotation about the origin, then it is the C_F origin). This equation for T is a matrix manifestation of an equation of the form $Y = MX + B$. To find the inverse of such an equation, you switch the X and Y, then solve for the Y:

$$X = MY + B \quad \Longrightarrow \quad Y = M^{-1}(X - B).$$

Thus T^{-1} can be written in the form

$$T^{-1}\begin{pmatrix}x\\y\end{pmatrix} = M^{-1} \cdot \left(\begin{pmatrix}x\\y\end{pmatrix} - \begin{pmatrix}e\\f\end{pmatrix}\right).$$

Now let's turn our attention to the frames F and G and the coordinate systems that they define. Let $F = \{v_x, v_y\}$ and $G = \{w_x, w_y\}$ and let O_F and O_G be the vertices of the frames F and G, respectively. They serve as the origins of the C_F and C_G coordinate systems. Note that $T(O_F) = O_G$ and and that, in the C_F coordinate system, O_F has coordinates $(0,0)$. Therefore, in the C_F coordinate system, the coordinates of O_G are

$$T\begin{pmatrix}0\\0\end{pmatrix} = M\begin{pmatrix}0\\0\end{pmatrix} + \begin{pmatrix}e\\f\end{pmatrix} = \begin{pmatrix}e\\f\end{pmatrix}.$$

Finally, we can talk about the coordinates of a general point P. When we say that P has coordinates (x,y) in the C_G coordinate system, what that means is that the vector from O_G to P can be written as the linear combination $x\vec{w}_x + y\vec{w}_y$ (where $\vec{w}_x$ and $\vec{w}_y$ are the vectors represented by the segments w_x and w_y directed to point away from O_G). In terms of the C_F coordinate system, then, the vector from O_F to P can be written as

$$x\vec{w}_x + y\vec{w}_y + \begin{pmatrix} e \\ f \end{pmatrix}.$$

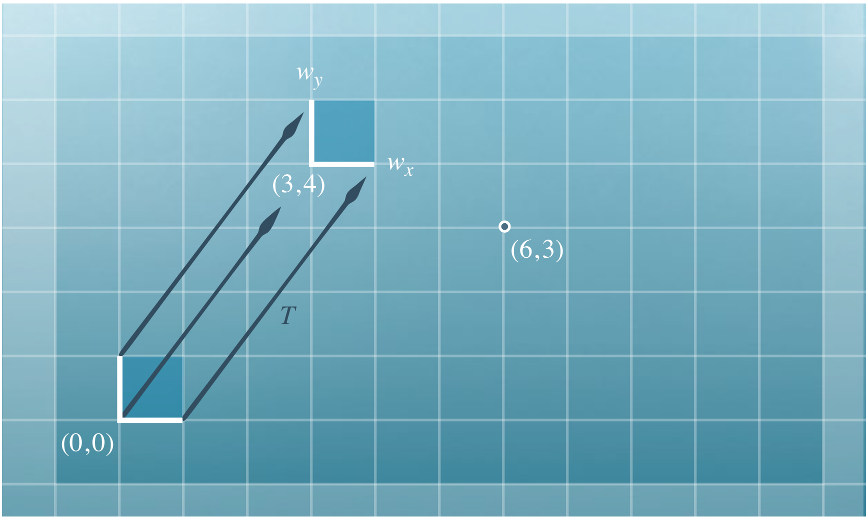
From that, we can now compute $T^{-1}(P)$. Along the way, we will use the fact that the matrix multiplication acts linearly, as discussed right before the start of this proof.

$$\begin{aligned} T^{-1}(P) &= T^{-1}\left(x\vec{w}_x + y\vec{w}_y + \begin{pmatrix} e \\ f \end{pmatrix}\right) \\ &= M^{-1}\left(\left(x\vec{w}_x + y\vec{w}_y + \begin{pmatrix} e \\ f \end{pmatrix}\right) - \begin{pmatrix} e \\ f \end{pmatrix}\right) \\ &= M^{-1}(x\vec{w}_x + y\vec{w}_y) \\ &= M^{-1}(x\vec{w}_x) + M^{-1}(y\vec{w}_y) \\ &= x \cdot M^{-1}(\vec{w}_x) + y \cdot M^{-1}(\vec{w}_y) \end{aligned}$$

Now T maps the segments v_x and v_y to w_x and w_y respectively. It therefore maps the vectors $\vec{v}_x$ and $\vec{v}_y$ to $\vec{w}_x$ and $\vec{w}_y$. In fact, though, the situation with the vectors is a little simpler. The map T has two components: a matrix component M and a translation component B. The translation component has no effect on the vectors– translating a representative of a vector just gives another representative of the same vector– as far as vectors are concerned, all the effect of T is contained in the matrix M. Therefore $M(\vec{v}_x) = \vec{w}_x$ and $M(\vec{v}_y) = \vec{w}_y$, and so $M^{-1}(\vec{w}_x) = \vec{v}_x$ and $M^{-1}(\vec{w}_y) = \vec{v}_y$. Plugging those in,

$$T^{-1}(P) = x\vec{v}_x + y\vec{v}_y,$$

and so the coordinates for $T^{-1}(P)$ in the C_F coordinate system are (x,y), the same as the coordinates for P in the C_G system. □



The real value of this theorem is in situations where calculations are difficult to work out in one coordinate system, but easy in another. In order for you to get a more concrete sense of this result, though, let me look at a few examples where coordinates of a point can be easily determined in both systems.

Example 1. Let G be the frame $\{w_x, w_y\}$ where in C_F coordinates,

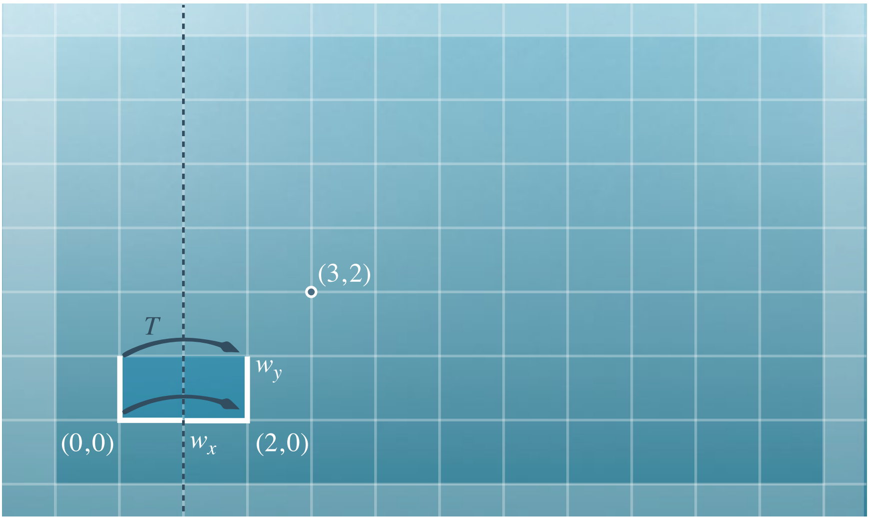
- w_x has endpoint $(3,4)$ and $(4,4)$, and
- w_y has endpoint $(3,4)$ and $(3,5)$.

Consider a point P with C_F coordinates $(6,3)$. It is clear that its C_G coordinates should be $(3,-1)$. Let's see if the previous theorem confirms that. The isometry T that maps F to G is a translation by $\langle 3,4 \rangle$. Its inverse is the translation in the opposite direction:

$$T^{-1}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} 3 \\ 4 \end{pmatrix},$$

and so, as anticipated,

$$T^{-1}\begin{pmatrix} 6 \\ 3 \end{pmatrix} = \begin{pmatrix} 6 \\ 3 \end{pmatrix} - \begin{pmatrix} 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 3 \\ -1 \end{pmatrix}.$$



Example 2. Let G be the frame $\{w_x, w_y\}$ where in C_F coordinates,

- w_x has endpoint $(2,0)$ and $(1,0)$, and
- w_y has endpoint $(2,0)$ and $(2,1)$.

Consider a point P with C_F coordinates $(3,2)$. Again, we can see that the C_G coordinates should be $(-1,2)$. This time, the isometry that maps F to G is a reflection that is given by the equation

$$T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 2 \\ 0 \end{pmatrix} = \begin{pmatrix} -x \\ y \end{pmatrix} + \begin{pmatrix} 2 \\ 0 \end{pmatrix} = \begin{pmatrix} 2-x \\ y \end{pmatrix}.$$

Since it is a reflection, it is its own inverse and we can calculate

$$T^{-1}\begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 2-3 \\ 2 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}.$$

In the last few lessons, we worked out the matrix equations for some, but not all, isometries– in particular, we only gave equation for rotations about the origin and reflections across lines through the origin. With the right change of basis, we can now move the origin around, and so get equations for any rotation or reflection. Let's consider an example.

Example. Suppose we want to find the matrix equation of a counterclockwise rotation by $\pi/2$ around the point $(3,1)$. Begin with the coordinates (x,y) of an arbitrary point P. Now, the only formula we have for a rotation is one for rotation about the origin. In order to use that formula, we are going to have to switch to a coordinate system with $(3,1)$ as its origin. We can do it with a translation. There are three steps to the process:

1. *Find the coordinates of P in the new coordinate system.*
The translation $T : (x,y) \mapsto (x+3, y+1)$ takes the current coordinate frame to one with the origin at $(3,1)$. To find the coordinates of P in the new system, we just need to calculate $T^{-1}(P)$.

2. *Calculate the rotation of this point.*
The matrix for this rotation is

$$\begin{pmatrix} \cos\pi/2 & -\sin\pi/2 \\ \sin\pi/2 & \cos\pi/2 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

3. *Write the the result in the original coordinate system.*
Going the other direction, we now need to apply T to the result.

Combining those three steps gives the equation of the rotation:

$$\begin{aligned} R\begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\left[\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} 3 \\ 1 \end{pmatrix}\right] + \begin{pmatrix} 3 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} x-3 \\ y-1 \end{pmatrix} + \begin{pmatrix} 3 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 1-y \\ x-3 \end{pmatrix} + \begin{pmatrix} 3 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 4-y \\ x-2 \end{pmatrix}. \end{aligned}$$

This single example illustrates the general procedure. Let τ be an isometry. Suppose F and G are frames and that S is the matrix equation of the isometry that maps F to G (written in terms of the F-coordinate system). Suppose that τ can be expressed as a matrix equation T in the G-coordinate system. Then τ can be expressed as the matrix equation $S \circ T \circ S^{-1}$ in the F-coordinate system.

Exercises

1. Verify that vector addition is commutative and associative.

2. Prove that there is a unique isometry from any (orthonormal) frame to any other (orthonormal) frame.

3. Prove the theorem in the lesson called "The Linearity of Matrix Operations". You may assume that M is a 2×2 matrix and that v_1 and v_2 are vectors in the plane.

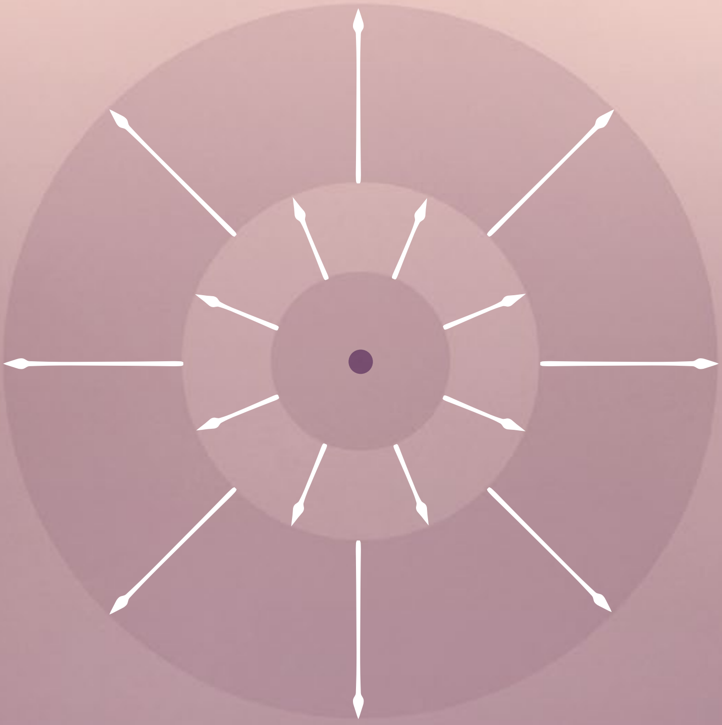
4. What is the image of the point $(3,0)$ under the counterclockwise rotation by an angle $\pi/6$ about the point $(1,1)$?

5. What is the matrix equation for a glide reflection whose glide line is $y = 2x + 1$ and whose glide distance is 5 (and the glide vector points from the origin into the first quadrant)?

6. Use a change of coordinates to find the general form for the counterclockwise rotation by an angle θ about a point (h,k).

7. Use a change of coordinates to find the general form for the reflection across the line $y = mx + b$.

8. (a) Show that the composition of two translations is either a translation or the identity.
   (b) Show that the composition of a translation and a rotation is a rotation.
   (c) Show that the composition of two rotations is: (1) a translation or the identity if the sum of the rotation angles is a multiple of 2π; (2) a rotation otherwise.



30 DILATION

Similarity mappings

Throughout our study of Euclidean geometry, we have dealt with two fundamentally important equivalence relations for triangles– congruence and similarity. The isometries of the last few lessons are closely tied to the congruence relation: if T is any triangle and τ is any isometry, then $\tau(T)$ is congruent to T. In this lesson, we will look at mappings that are tied to the similarity relation.

DEF: SIMILARITY MAPPING

Def. A bijective mapping σ of the Euclidean plane is called a *similarity mapping* if for every triangle T, T and its image $\sigma(T)$ are similar.

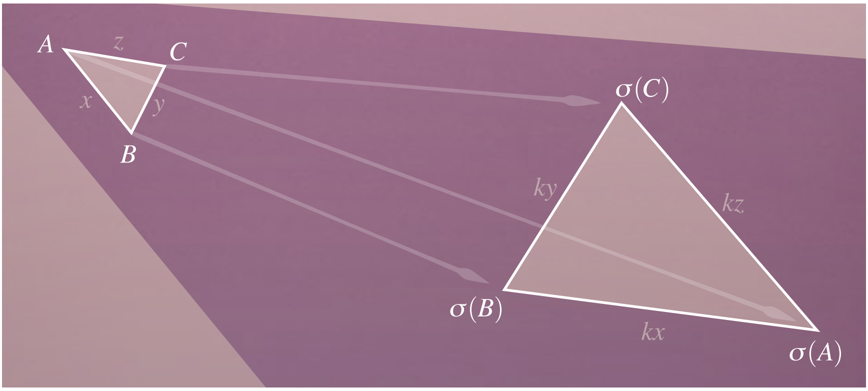
The first and most important thing to do is to understand the effect that a similarity mapping will have on distance.

THM: SIMILARITY MAPPINGS AND DISTANCE

A bijection σ is a similarity mapping if and only if it scales all distances by a constant. That is, σ is a similarity mapping if and only if there is a positive real number k so that $|\sigma(AB)| = k|AB|$ for all segments AB.

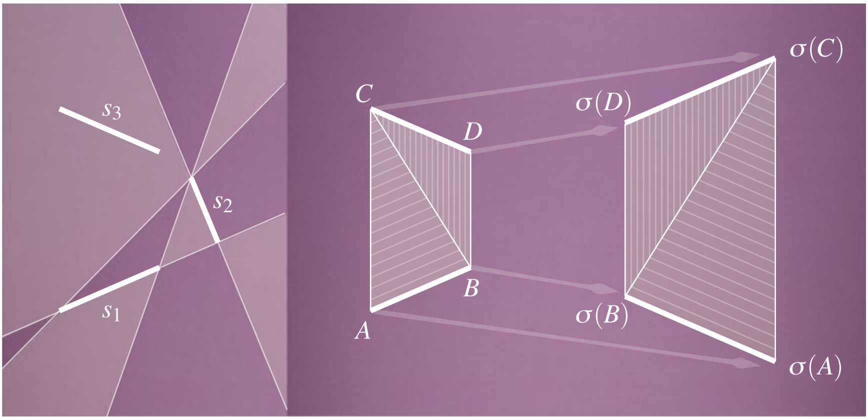
Proof. $\Longrightarrow$ First suppose that σ scales all distances by a constant k. Then given any triangle $\triangle ABC$,

$$|\sigma(AB)| = k|AB| \quad |\sigma(AC)| = k|AC| \quad |\sigma(BC)| = k|BC|.$$



By the S·S·S similarity theorem, $\triangle ABC$ and $\sigma(\triangle ABC)$ are similar, and so σ meets the requirements of a similarity mapping.

$\Longleftarrow$ Now suppose that σ is a similarity mapping. We need to show that σ scales all distances by a constant– suppose instead that there are two segments s_1 and s_2 that are not scaled by the same amount. From that, we will try to get to a contradiction. This proof uses some triangles, and in order to guarantee that the triangles will be properly formed, I need s_1 and s_2 to be in "general position", so that no three endpoints of s_1 and s_2 are collinear. Of course it is possible that s_1 and s_2 are not in general position– what to do in that case? Choose another segment, s_3, and get it right this time: choose one whose two endpoints are *not* on any of the lines formed by a pair of endpoints from s_1 and s_2. This new segment may be scaled by the same amount as s_1, or it may be scaled by the same amount as s_2, or it may be scaled by an entirely different amount. In any case, s_3 can't be scaled by the same amount as both s_1 and s_2 since they themselves differ. So now we have a setup with two segments in general position with different scaling constants. Label them AB and CD.



Fix a bad arrangement by replacing s_2 with s_3. Then look at similar triangles.

Consider $\triangle ABC$. Since σ is a similarity mapping, $\sigma(\triangle ABC)$ is similar to $\triangle ABC$. There is, then, a constant k so that

$$|\sigma(AB)| = k|AB| \quad \& \quad |\sigma(BC)| = k|BC|.$$

Second, $\sigma(\triangle BCD)$ is similar to $\triangle BCD$. We already know that $|\sigma(BC)| = k|BC|$, and so $|\sigma(CD)| = k|CD|$. But that then means that AB and CD are scaled by the *same* amount, a contradiction. □

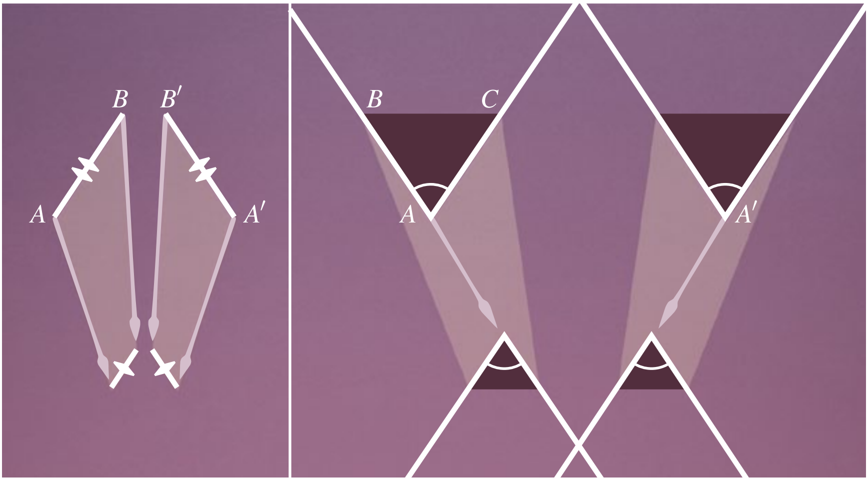
Let's investigate some of the properties of a similarity mapping σ.

LEM: SEGMENT CONGRUENCE
If $AB \simeq A'B'$, then $\sigma(AB) \simeq \sigma(A'B')$.

Proof. This follows immediately from the previous theorem since segments are congruent when they are the same length, and since

$$|\sigma(AB)| = k|AB| = k|A'B'| = |\sigma(A'B')|.$$

□



Congruence of segments. And of angles.

LEM: ANGLE CONGRUENCE
1. For any angle $\angle A$, $\sigma(\angle A) \simeq \angle A$.
2. If $\angle A \simeq \angle A'$, then $\sigma(\angle A) \simeq \sigma(\angle A')$.

Proof. 1. Mark points B and C on the two rays forming A to make a triangle $\triangle ABC$. Since σ is a similarity mapping, $\sigma(\triangle ABC)$ is similar to $\triangle ABC$. The corresponding angles in similar triangles are congruent, so $\sigma(\angle A) \simeq \angle A$.

2. If $\angle A \simeq \angle B$, then using the first part,

$$\sigma(\angle A) \simeq \angle A \simeq \angle A' \simeq \sigma(\angle A').$$

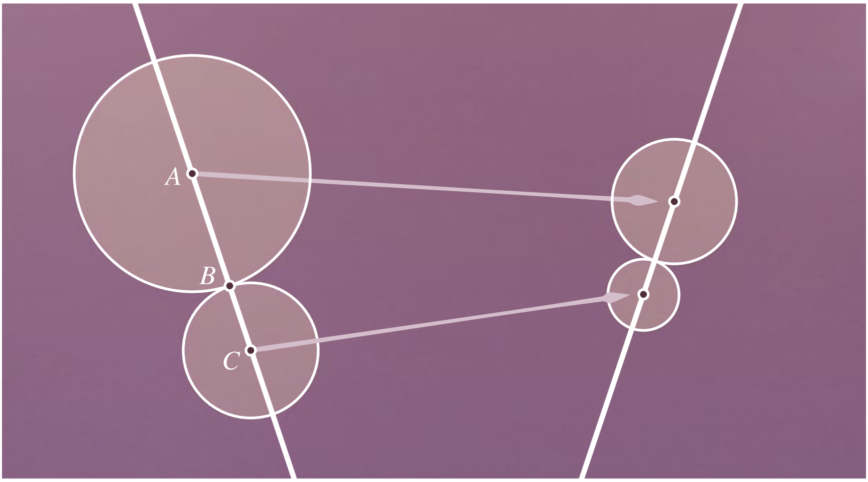
□

Note that this property together with the distance scaling property means that a similarity mapping will map any *polygon* to a similar polygon, not just triangles.

LEM: INCIDENCE AND ORDER
If $A * B * C$, then $\sigma(A) * \sigma(B) * \sigma(C)$.

Proof. Since $A * B * C$, $|AC| = |AB| + |BC|$. Multiply through by the scaling constant k to get

$$k|AC| = k|AB| + k|BC|$$
$$|\sigma(AC)| = |\sigma(AB)| + |\sigma(BC)|.$$



The image of B has to be at the intersection of the circles.

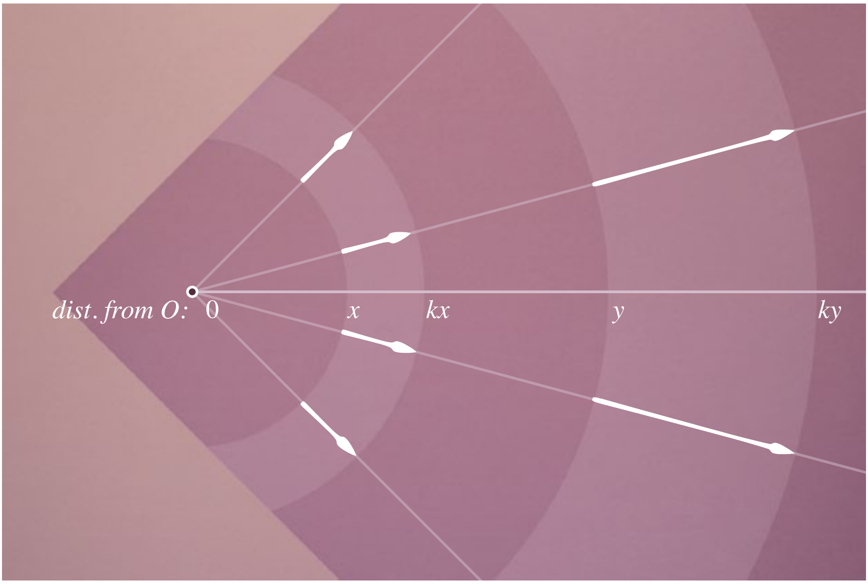
This is the degenerate case of the triangle inequality. The only way it can be true is if $\sigma(A)$, $\sigma(B)$, and $\sigma(C)$ are all collinear, and $\sigma(B)$ is between $\sigma(A)$ and $\sigma(C)$. □

More generally, the images of any number of collinear points are collinear, and their order is retained. Essentially, while similarity mappings distort distances, they do so in a relatively tame way, and the key synthetic relations of incidence, order, and congruence, are preserved.

Dilations

We have looked at some properties of similarities without ever really asking whether there are in fact mappings (other than isometries) that meet this condition. There are, of course– we use them daily whenever we use a map, or a blueprint, or a scale model.

DEF: DILATION
Let O be a point and k be a positive real number. The *dilation* by a factor of k centered at O is the map d of the Euclidean plane so that
1. $d(O) = O$, and
2. for any other point P, $d(P)$ is the point on $OP\rightarrow$ that is a distance $k|OP|$ from O.



Dilations are also called scalings, dilatations, and occasionally homotheties. First of all, it is clear that a dilation is a bijection (that it is both one-to-one and onto). In fact, it is easy to describe its inverse: if d is the dilation by k centered at O, its inverse is another dilation centered at O, this time by a factor of $1/k$. When $k = 1$, d is the identity map. Otherwise, a dilation will not be an isometry– it will alter distance.

THM: DILATIONS AND DISTANCE

A dilation is a similarity mapping.

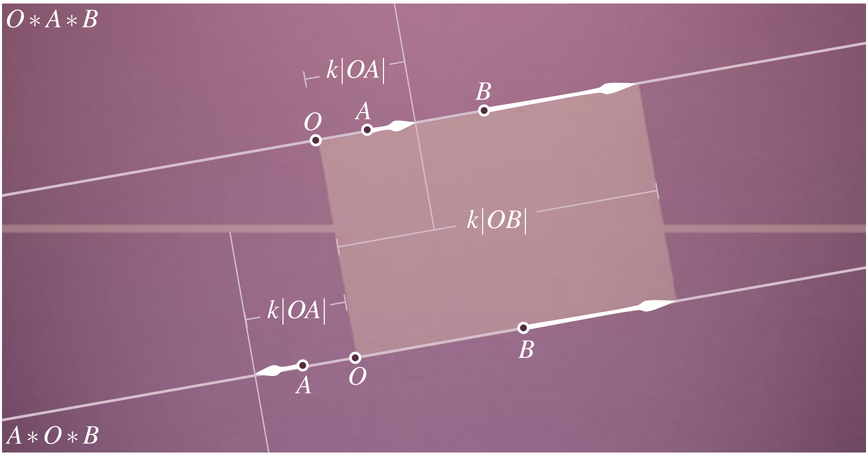
Proof. Let d be a dilation centered at O with a scaling factor of k. By definition, any segment with one endpoint on O will be scaled by k. To show that d is a similarity mapping, we need to show that any other segment AB is scaled by that same amount. There are a handful of cases to consider.

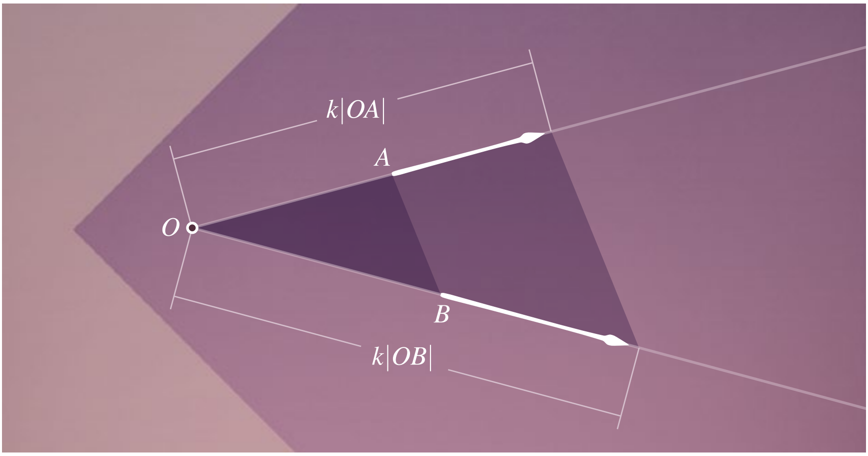
1. Suppose that A and B are on the same ray from O, and for the sake of convenience, let's suppose that A is between O and B. Then $d(A)$ and $d(B)$ are still on that same ray from O, at respective distances of $k|OA|$ and $k|OB|$, and so $d(A)$ is still between O and $d(B)$. Therefore

$$\begin{aligned}|d(AB)| &= |d(OB)| - |d(OA)| \\ &= k|OB| - k|OA| \\ &= k(|OB| - |OA|) \\ &= k|AB|.\end{aligned}$$

2. Suppose that A and B are on opposite rays from O. Then $d(A)$ and $d(B)$ are also on those same opposite rays, and so

$$\begin{aligned}|d(AB)| &= |d(OA)| + |d(OB)| \\ &= k|OA| + k|OB| \\ &= k(|OA| + |OB|) \\ &= k|AB|.\end{aligned}$$





3. Surely the most common case, though, is when A and B are neither on the same ray, nor on opposite rays from O. Compare then the triangles $\triangle AOB$ and $d(\triangle AOB)$. Since $d(O) = O$, and $d(A)$ and $d(B)$ are on the same rays from O as A and B, $\angle AOB = d(\angle AOB)$. In addition, $|d(OA)| = k|OA|$ and $|d(OB)| = k|OB|$. By the S·A·S similarity theorem, then, $\triangle AOB$ and $d(\triangle AOB)$ are similar. Comparing the third sides of those triangles, $|d(AB)| = k|AB|$. □

As with isometries, the effect of a dilation can be described with a matrix equation.

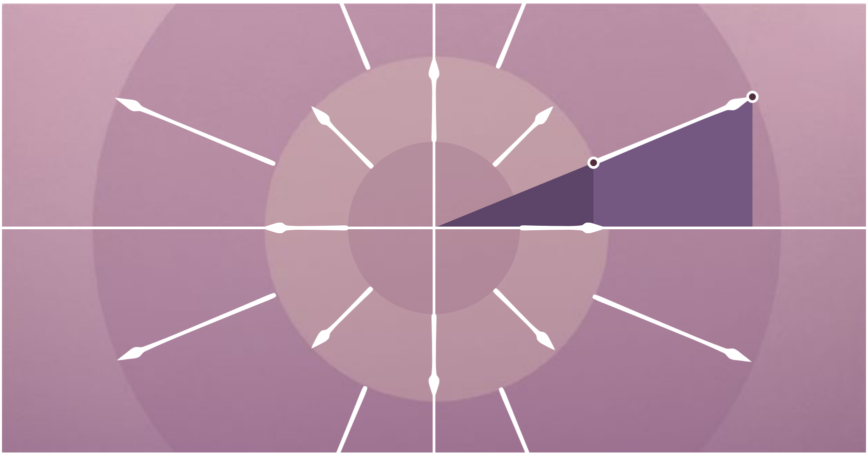
EQN: SCALING ABOUT THE ORIGIN

The matrix equation for a dilation d by a factor of k centered at the point $(0,0)$ is

$$d\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} kx \\ ky \end{pmatrix}.$$

Proof. We need to show that the mapping d that is given by the equation has the same effect on points as a dilation by k does. There are three things to show:

1. that d fixes the origin O;
2. that for any other point P, $d(P)$ is on $OP\rightarrow$; and
3. that the distance from O to $d(P)$ is $k|OP|$.



1. $d\begin{pmatrix}0\\0\end{pmatrix} = \begin{pmatrix}k\cdot 0\\k\cdot 0\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix}.$

2. The slope of the line through the origin and (kx,ky) is $(ky)/(kx) = y/x$, the same as the slope of the line through the origin and (x,y). Therefore (kx,ky) and (x,y) are on the same *line* through the origin. Furthermore, since we specified that the scaling constant k of a dilation is a positive number, kx and ky will have the same signs as x and y, respectively. Therefore (kx,ky) and (x,y) will be the in same quadrant, and so they are on the same *ray* from the origin.

3. The distance from $(0,0)$ to (kx,ky) is

$$\sqrt{(kx-0)^2+(ky-0)^2} = \sqrt{k^2(x^2+y^2)} = k\sqrt{(x-0)^2+(y-0)^2}.$$

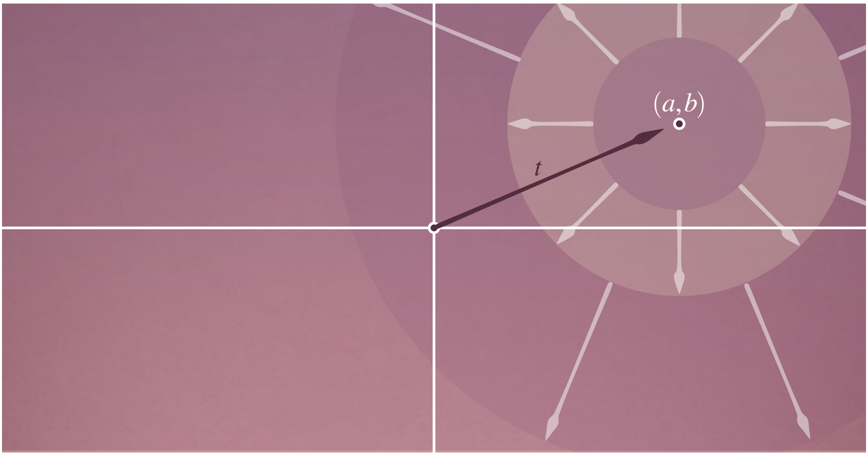
It is k times the distance from the origin to (x,y). □

As we did earlier with isometries, we can now use a change of coordinates to describe dilations about any point.

EQN: DILATION ABOUT AN ARBITRARY POINT

The matrix equation for a dilation d by a factor of k centered at the point (a,b) is

$$d\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}kx+(1-k)a\\ky+(1-k)b\end{pmatrix}.$$



Proof. Let P be an arbitrary point with coordinates (x,y). The strategy of this proof is simple– follow the procedure that we developed in the lesson on changing coordinates:

1. convert (x,y) to a coordinate system whose origin is at (a,b);
2. perform the scaling by a factor of k; and then
3. convert the result back to the original coordinate system.

1. The translation $t(x,y) = (x+a, y+b)$ shifts the standard coordinate frame centered at $(0,0)$ to one that is centered at (a,b). To compute the coordinates of P in the new coordinate system, then, apply t^{-1}:

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x-a \\ y-b \end{pmatrix}$$

2. Now scale by k, using the special formula from the previous theorem.

$$\mapsto \begin{pmatrix} k(x-a) \\ k(y-b) \end{pmatrix}$$

3. Convert back to the original coordinate system by applying t:

$$\mapsto \begin{pmatrix} k(x-a)+a \\ k(y-b)+b) \end{pmatrix} = \begin{pmatrix} kx+(1-k)a \\ ky+(1-k)b \end{pmatrix}.$$

□

Preserving incidence, order, and congruence

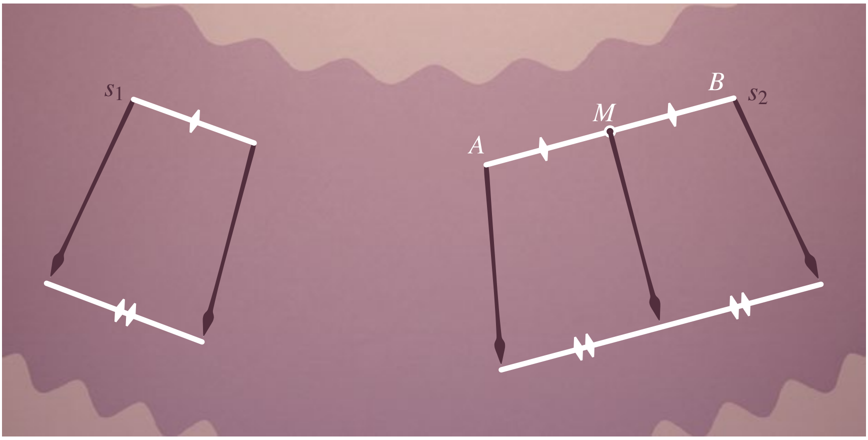
Dilations and isometries are similarity mappings. It is natural to wonder what other types of similarity mappings there might be, but I actually want to investigate what is in theory a slightly more general question. Every similarity mapping preserves the relations of incidence, order and congruence. We have seen two such types of mappings– dilations and isometries. What other types of bijections will preserve these structures? It all hinges on the congruence relation. In the next three lemmas, f is a bijection that preserves incidence, order, and congruence.

LEM: HALVING SEGMENTS
Let s_1 and s_2 be segments. If $|s_1| = \frac{1}{2}|s_2|$, and f scales s_2 by k, then f scales s_1 by k as well.

Proof. Label the two endpoints of s_2 as A and B, and its midpoint as M. Then all three segments s_1, AM, and BM are congruent and so their images must be as well. Then

$$\begin{aligned}|f(s_1)| &= (|f(s_1)| + |f(s_1)|)/2\\ &= (|f(AM)| + |f(BM)|)/2\\ &= |f(AB)|/2\\ &= k|AB|/2\\ &= k \cdot 2|s_1|/2\\ &= k|s_1|.\end{aligned}$$

□



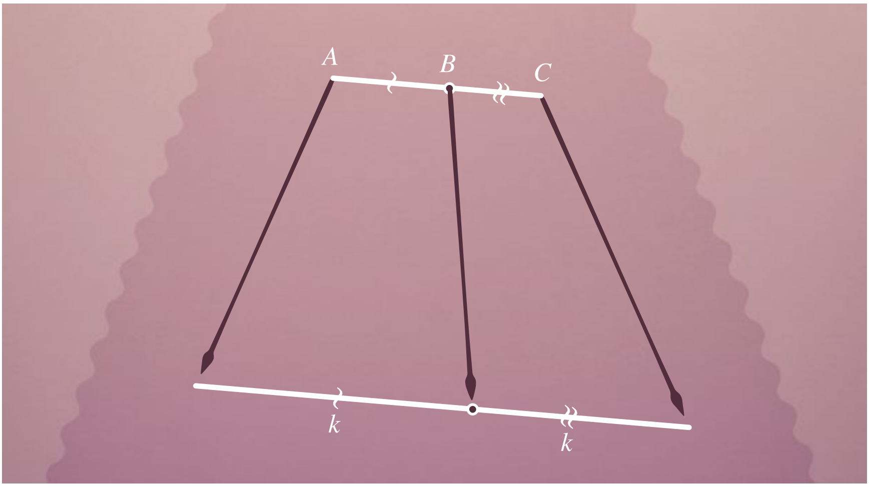
LEM: CHAINING SEGMENTS TOGETHER

If $A * B * C$ and if f scales both AB and BC by k, then f scales AC by that same constant.

Proof. Since f preserves the order of points, $f(A) * f(B) * f(C)$, and so

$$\begin{aligned}|f(AC)| &= |f(AB)| + |f(BC)| \\ &= k|AB| + k|BC| \\ &= k(|AB| + |BC|) \\ &= k(|AC|).\end{aligned}$$

□



LEM: DYADIC LENGTHS

If $|s_1| = (m/2^n) \cdot |s_2|$ where m and n are positive integers, and if f scales s_2 by k, then f scales s_1 by k as well.

Proof. The first lemma tells us that a segment of length $(1/2) \cdot |s_2|$ will be scaled by k. Applied again, it tells us that a segment of length $(1/4) \cdot |s_2|$ will be scaled by k. And so on, so that for all positive integers n, a segment of length $(1/2^n) \cdot |s_2|$ will be scaled by a factor of k. Then we can line up m segments of length $(1/2^n) \cdot |s_2|$, to get a segment of length $(m/2^n) \cdot |s_2|$. By repeatedly applying the second lemma, we can see that it too must be scaled by k. □

THM: THAT IS ALL, PART I

Any bijection of the Euclidean plane that preserves incidence, order, and congruence is a similarity mapping.

Proof. Let f be a bijection that preserves incidence, order, and congruence. Since f maps congruent segments to congruent segments, all segments of a given length will be scaled by the same amount. Let k be the scaling constant for a segment of length one. By subdividing and chaining together (as described above), k is the scaling constant for all segments of length $m/2^n$. We need to show that k is the scaling constant for segments of all other lengths as well. Suppose that segment OA has a length of x and that $|f(OA)| = k'|OA|$. To get an idea of k', we can use dyadic approximations to pin OA between segments that are scaled by k. For each n, there is an m_n so that

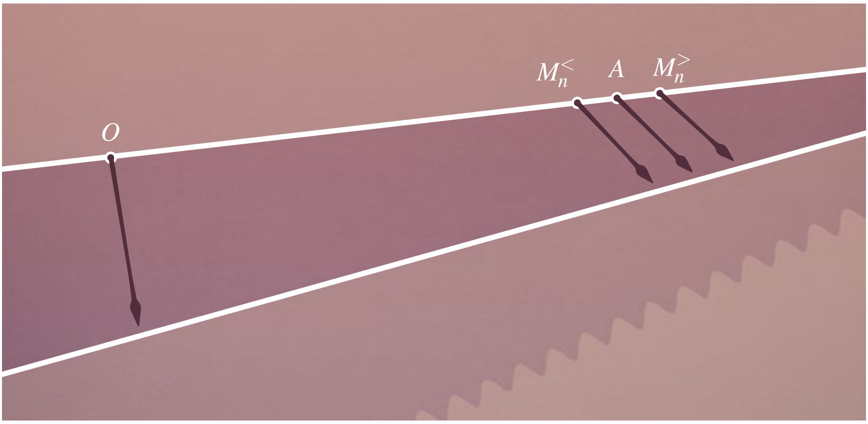
$$\frac{m_n}{2^n} \leq x \leq \frac{m_n+1}{2^n}.$$

Along the ray $OA \rightarrow$, mark off points $M_n^<$ and $M_n^>$ bracketing A so that $|OM_n^<| = m_n/2^n$ and $|OM_n^>| = (m_n+1)/2^n$. Reading off the points in order, then $O * M_n^< * A * M_n^>$. The distance between $M_n^<$ and $M_n^>$ is $1/2^n$, so as n increases, the bracketing of A gets tighter and tighter. Since f preserves incidence and order, when we apply it to these points, we get a bracketing of $f(A)$ that can give us an idea of the scaling of OA:

$$f(O) * f(M_n^<) * f(A) * f(M_n^>)$$

$$|f(OM_n^<)| \leq |f(OA)| \leq |f(OM_n^>)|$$

$$k \cdot m_n/2^n \leq k' \cdot |OA| \leq k \cdot (m_n+1)/2^n.$$



To find k', divide through by $|OA|$

$$k \cdot \frac{m_n/2^n}{|OA|} \le k' \le k \cdot \frac{(m_n+1)/2^n}{|OA|}.$$

This set of inequalities is true for all values of n. Notice that as n increases, both of the terms $(m_n/2^n)/|OA|$ and $((m_n+1)/2^n)/|OA|$ approach one. The only way that this inequality can be satisfied for all n, then, is for k' to be equal to k. Therefore f scales all distances by the same constant k– this means that f is a similarity mapping. □

THM: THAT IS ALL, PART II

Any bijection that preserves incidence, order, and congruence can be written as a composition of an isometry and a dilation.

Proof. Let f be such a bijection. As we have just seen, that means f is a similarity mapping, which in turn means that f scales all distances by some constant k. Let d be the dilation centered at the origin by a factor of k. Its inverse, d^{-1} is a dilation by a factor of $1/k$, so for any segment s,

$$|d^{-1} \circ f(s)| = (1/k) \cdot |f(s)| = (1/k) \cdot k \cdot |s| = |s|.$$

Therefore $d^{-1} \circ f$ is an isometry. Writing τ for this isometry, $d^{-1} \circ f = \tau$. Hit both sides of this equation with the dilation d to get

$$d \circ d^{-1} \circ f = d \circ \tau \implies f = d \circ \tau,$$

and we have just written f as a composition of an isometry and a dilation. □

Exercises

1. What is the image of the point $(2,3)$ under the scaling by a factor of 4 centered at the point $(1,5)$?

2. Show that if d_1 and d_2 are transformations with the same scaling factor, then there is an isometry τ so that $d_2 = \tau \circ d_1$.

3. Show that if d_1 is a scaling by a factor of k_1 and d_2 is a scaling by a factor of k_2, then $d_1 \circ d_2$ is a scaling by a factor of $k_1 \cdot k_2$.

4. Write an equation for the similarity mapping that is formed by

   1. first dilating by a factor of $1/2$ about the point $(1,1)$, and then
   2. reflecting across the x-axis.

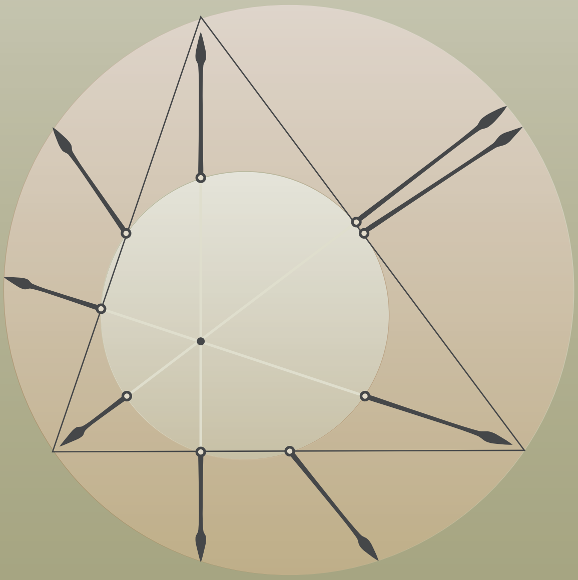
   Does this transformation have any fixed points?

5. Prove that if $\triangle ABC \sim \triangle A'B'C'$, then there is a similarity mapping σ so that $\sigma(A) = A'$, $\sigma(B) = B'$, and $\sigma(C) = C'$.

6. Consider the similar triangles $\triangle ABC$ and $\triangle A'B'C'$ with vertices at the following coordinates:

$$\begin{array}{ccc} A = (0,0) & B = (1,0) & C = (0,1) \\ A' = (2,0) & B' = (0,2) & C' = (0,-2) \end{array}$$

   Find the equation of the similarity mapping that maps $\triangle ABC$ to $\triangle A'B'C'$.



31 APPLICATIONS OF TRANSFORMATIONS

We have spent the last several lessons building up a theory of Euclidean transformations. To do that, we drew upon some of the Euclidean theory that we had previously developed. Now in this lesson we will turn the tables and use the theory of transformations to prove three results of classical Euclidean geometry.

Varignon's Theorem

The first result, Varignon's Theorem, discovers a parallelogram that hides inside of any quadrilateral. The proof of this theorem uses half-turns. Recall from Lesson 26 (on rotations) that

> DEF: HALF-TURN
> A half-turn is a rotation with a rotation angle of π.

Note that a half-turn is its own inverse. Because of that, this is the one instance where we don't have to specify whether the rotation is clockwise or counterclockwise– they are the same. In the exercises at the end of Lesson 29 (change of coordinates), I asked you to investigate what happens when you compose two rotations. In particular, you were supposed to verify that if the two angles of rotation add up to a multiple of 2π, then their composition is either the identity or a translation (it is a fairly straightforward, albeit messy, calculation using the matrix equations for a rotation). Because of that, when we compose any two half-turns, their rotation angles add up to $\pi + \pi = 2\pi$, and the result must be either a translation or the identity.

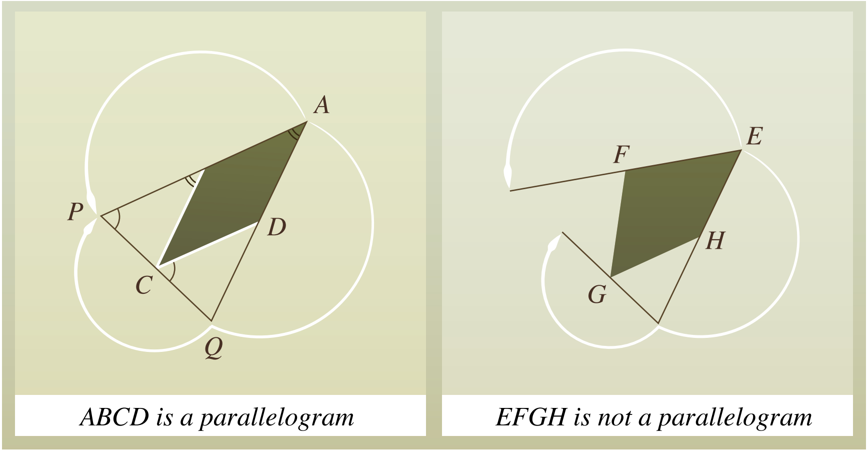
> LEM: COMPOSING FOUR HALF-TURNS
> Let r_A, r_B, r_C, and r_D be half-turns around four distinct points A, B, C, and D. If the composition $r_A \circ r_B \circ r_C \circ r_D$ is the identity map, then the quadrilateral $ABCD$ is a parallelogram.

Proof. Let's break that four-part composition into two pieces: $r_A \circ r_B$ is one and $r_C \circ r_D$ is the other. If we assume that their composition is the identity, then they must be inverses of each other. That is

$$r_C \circ r_D = (r_A \circ r_B)^{-1} = r_B^{-1} \circ r_A^{-1}.$$

Each of r_B and r_A is its own inverse, though, since they are half-turns. Thus $r_C \circ r_D = r_B \circ r_A$. In that case, we can apply both of these maps to the point A, chasing it in two directions around the quadrilateral, and we should end up in the same place. Label that ending point P, and along the way label one more point, $Q = r_D(A)$. That is,

$$r_C \circ r_D(A) = r_C(Q) = P \quad \& \quad r_B \circ r_A(A) = r_B(A) = P.$$



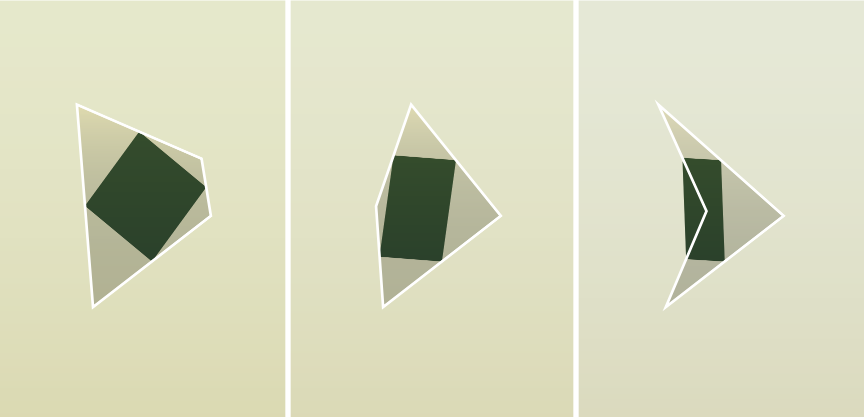
ABCD is a parallelogram

EFGH is not a parallelogram

The points A, P, and Q form a triangle around the original quadrilateral. This triangle is particularly well-balanced with respect to $ABCD$. You see, because r_D is an isometry, $|AD| = |DQ|$; and because r_C is an isometry, $|QC| = |CP|$; and because r_B is an isometry, $|PB| = |AB|$. Thus,

$$|AQ| = 2|DQ|, \quad |QP| = 2|CQ| = 2|CP|, \quad |PA| = 2|PB|.$$

By S·A·S similarity we have created two sets of similar triangles: $\triangle AQP$ is similar to $\triangle DQC$, and $\triangle QPA$ is similar to $\triangle CPB$. Matching up angles in them, $\angle DCQ \simeq \angle P$ and $\angle A \simeq \angle PBC$. Finally, the Alternate Interior Angle Theorem tells us that $CD \parallel AB$ and $AD \parallel BC$ and so $ABCD$ is, by definition, a parallelogram. □

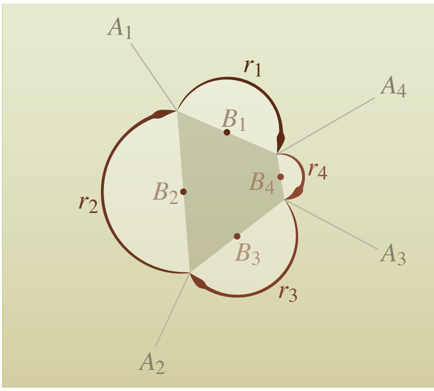


THM: VARIGNON'S THEOREM

Let $A_1A_2A_3A_4$ be any quadrilateral and label the midpoints of the four sides B_1, B_2, B_3 and B_4, so that B_i is the midpoint of A_iA_{i-1} (subscripts are taken "mod 4"). Then $B_1B_2B_3B_4$ is a parallelogram.

Proof. The strategy should be pretty obvious– use the last lemma! That means we need to look at the composition $r_1 \circ r_2 \circ r_3 \circ r_4$ of half-turns around the four midpoints B_1, B_2, B_3, and B_4. We need to show it is the identity. For starters, let's take the four half-turns in pairs again, as $r_1 \circ r_2$ and $r_3 \circ r_4$. Each of these is a translation, and so their composition is either a translation or the identity. Now the easiest way to show that a map is the identity rather than a translation is to find a fixed point– translations don't have any. In the case of $r_1 \circ r_2 \circ r_3 \circ r_4$ there is one fixed point that is easy to find:

$$\begin{aligned} &r_1 \circ r_2 \circ r_3 \circ r_4(A_4) \\ &= r_1 \circ r_2 \circ r_3(A_3) \\ &= r_1 \circ r_2(A_2) \\ &= r_1(A_1) \\ &= A_4. \end{aligned}$$



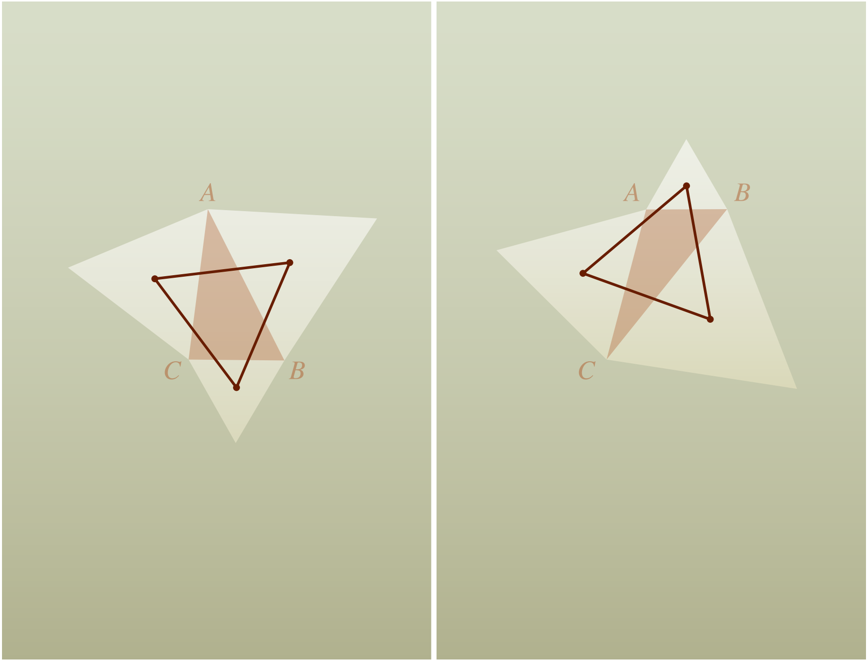
Since $r_1 \circ r_2 \circ r_3 \circ r_4$ has a fixed point, it cannot be a translation, and so it must be the identity. According to the previous lemma, $B_1B_2B_3B_4$ must be a parallelogram. □

Napoleon's Theorem

Like Varignon's Theorem, Napoleon's Theorem reveals an unexpected symmetry. And yes, it is named after *that* Napoleon, although there is some skepticism about whether he in fact discovered it. I guess once you have conquered half of Europe, no one is going to raise a fuss if you claim a theorem or two as well.

THM: NAPOLEON'S THEOREM

Given any triangle $\triangle ABC$, construct three equilateral triangles exterior to it– one on each of the sides AB, BC, and CA. The centers of these three equilateral triangles are the vertices of another triangle. This triangle is also equilateral.



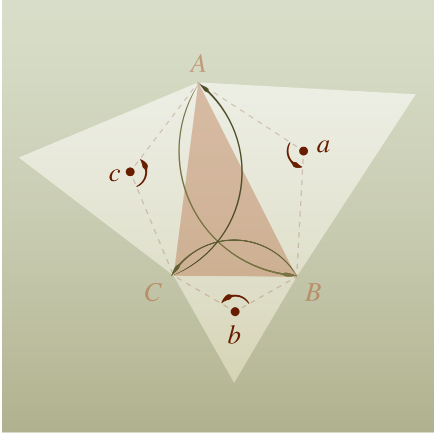
Napoleon's Theorem: two examples

Proof. This proof begins as Varignon's did, with a composition of rotations whose rotation angles add up to 2π. The fixed point is easy to find, meaning that the composition is the identity. It may not be immediately clear how to use that fact in a meaningful way, and so it is admittedly a bit of a scramble to the finish. Anyway, this time around the fundamental symmetry of the situation comes from the three equilateral triangles, and the rotations that capture that symmetry are 1/3-turns around the centers of the equilateral triangles. To make sure that our labeling is consistent, let's do a quick check: I want the path that goes from A to B to C to A to make a *clockwise* loop around the triangle. If it instead makes a counterclockwise loop, you can just swap two of the labels to fix it. Now label the centers of those equilateral triangles as a, b, and c, where

a is the center of the triangle built off of side AB,

b is the center of the triangle built off of side BC, and

c is the center of the triangle built off of side CA.



Label the corresponding $2\pi/3$ counterclockwise rotations around these points as r_a, r_b, and r_c. When we compose these three rotations, their rotation angles add up to $2\pi/3 + 2\pi/3 + 2\pi/3 = 2\pi$, so their composition $r_c \circ r_b \circ r_a$ must be either a translation or the identity. Now take a look inside one of the equilateral triangles, say the one centered at a, and notice that in it $|aA| = |aB|$, and $(\angle AaB) = 2\pi/3$. That means that r_a sends A to B. Likewise, r_b sends B to C and r_c sends C to A. In combination,

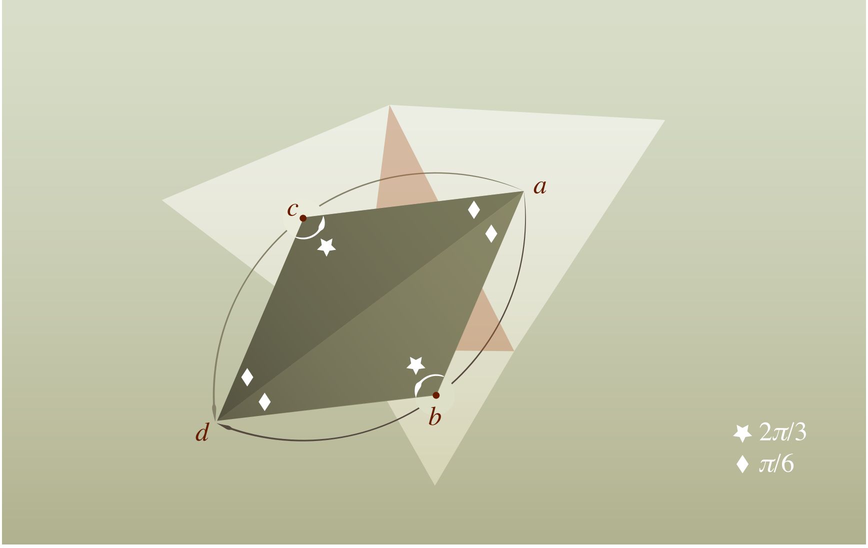
$$r_c \circ r_b \circ r_a(A) = r_c \circ r_b(B) = r_c(C) = A,$$

and so $r_c \circ r_b \circ r_a$ has a fixed point. Well, it can't be a translation then, so it must be the identity.

Now for the scrambling part. Let's see what happens when we plug the point a into this composition (that is really just the identity):

$$r_c \circ r_b \circ r_a(a) = a \implies r_c \circ r_b(a) = a \implies r_b(a) = r_c^{-1}(a)$$

This gives us one last point to label: $d = r_b(a)$. There are two triangles to look at.



The first is $\triangle abd$. Since r_b maps the segment ba to the segment bd, ba and bd are congruent. Thus $\triangle abd$ is an isosceles triangle. Furthermore, at vertex b, we know the angle measure is $2\pi/3$. The other two angles in this triangle must add up to $\pi - 2\pi/3 = \pi/3$. According to the Isosceles Triangle Theorem, they are congruent, though, so they each measure $\pi/6$.

The second triangle is $\triangle acd$. The map r_c^{-1} is also a rotation by $2\pi/3$– it is just a *clockwise* rotation by that amount. It maps the segment ca to the segment cd, and so they must be congruent. Therefore, $\triangle acd$ is also isosceles, its angle at vertex c has a measure of $2\pi/3$, and that means its other two angles also must each measure $\pi/6$.

Finally, when we put the two pieces together, we get

$$(\angle bac) = (\angle bad) + (\angle cad) = \pi/6 + \pi/6 = \pi/3.$$

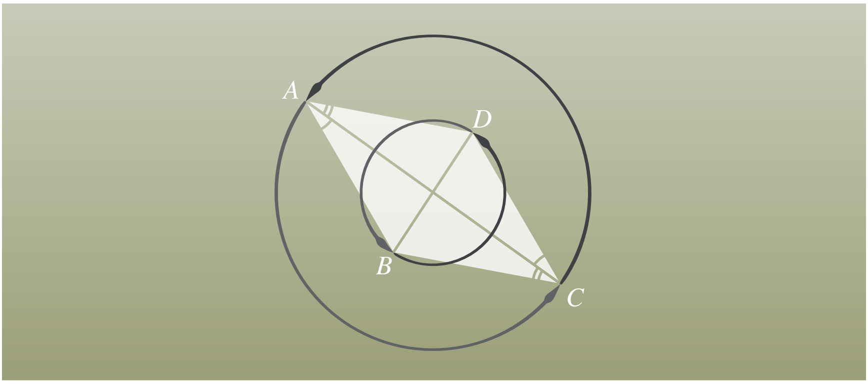
This angle at a is no more special than the angles at vertices b and c, though. A similar argument (in which the the compositions of r_a, r_b, and r_c are taken in different orders) will show that the other two angles of $\triangle abc$ also measure $\pi/3$. Therefore $\triangle abc$ is equiangular and so it is equilateral. □

The Nine Point Circle

For the last part of this lesson, let's look back at the Nine Point Circle Theorem. We proved this theorem way back in Lesson 20 without using transformation methods– the key then was to find a diameter of the nine-point circle. This time, the key is to find a transformation that maps the nine-point circle to the circumcircle. In the Lesson 20 proof, we also needed to know that the diagonals of a parallelogram bisect one another. In this proof, we will need the converse of that.

LEM: BISECTING DIAGONALS
If segments AC and BD bisect each other, then the quadrilateral $ABCD$ is a parallelogram.

Proof. Let h be the half-turn around the point of intersection of AC and BD. Then h interchanges A and C, and it interchanges B and D. Therefore $h(\angle BAC) = \angle DCA$. That means that $\angle BAC$ must be congruent to $\angle DCA$, and according to the Alternate Interior Angle Theorem, then, AB is parallel to CD. Similarly, $h(\angle CAD) = \angle ACB$, meaning $\angle CAD$ is congruent to $\angle ACB$, so AD is parallel to BC. Quadrilateral $ABCD$ has two pairs of parallel sides– it must be a parallelogram. □



THM: THE NINE POINT CIRCLE THEOREM, REVISITED

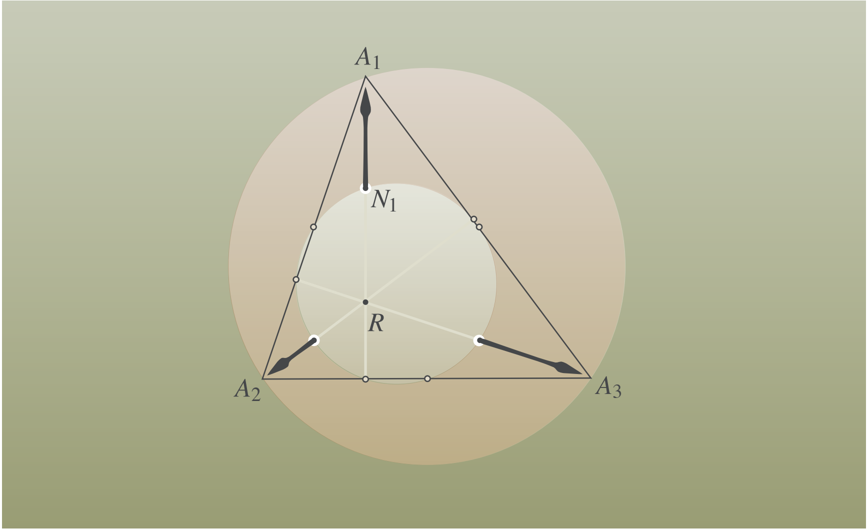
For any triangle, the following nine points all lie on the same circle: (1) the feet of the three altitudes, (2) the midpoints of the three sides, and (3) the midpoints of the three segments connecting the orthocenter to each vertex. This circle is called the nine-point circle associated with that triangle.

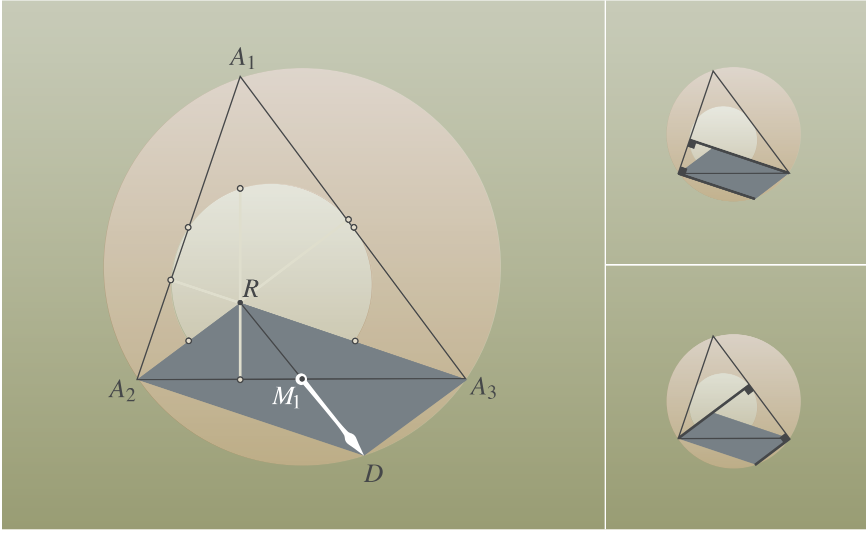
Proof. Given a triangle $\triangle A_1A_2A_3$ with orthocenter R, label

L_i, the foot of the altitude which passes through A_i,
M_i, the midpoint of the side that is opposite A_i, and
N_i, the midpoint of the segment A_iR.

Let d be the dilation by a factor of two centered at the orthocenter. We will show that $d(L_i)$, $d(M_i)$, and $d(N_i)$ are all on the circumcircle $\mathcal{C}$. [Note that this proof does not handle a few degenerate cases: when $M_i = R$, the quadrilateral described in (2) cannot be formed, and when $L_i = M_i$, the right angle described in (3) cannot be formed. Those case are easily resolved though, so I have omitted them to keep the proof as streamlined as possible.]

The points N_i. Since N_i is halfway from R to A_i, d maps each of the points N_i to the corresponding vertex A_i. All three of the vertices are, of course, on $\mathcal{C}$.



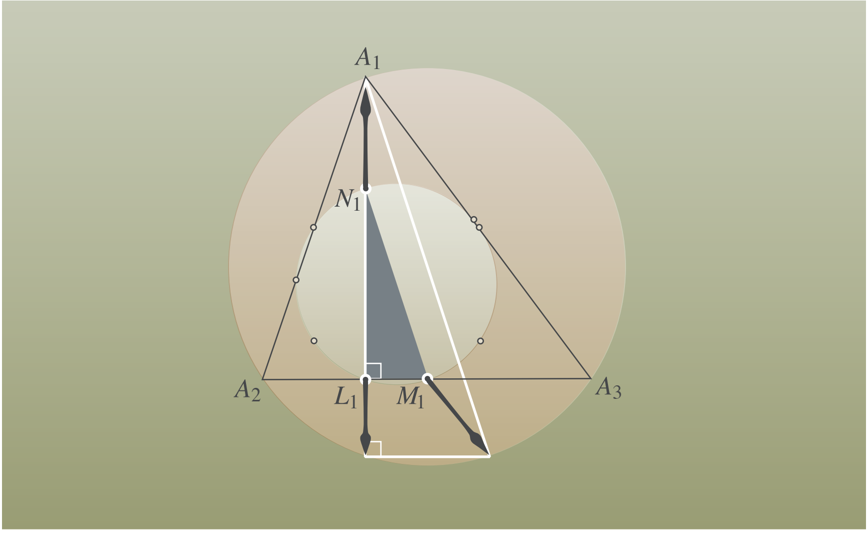


The points M_i. This is the difficult one. Take for example M_1, the midpoint of A_2A_3. The dilation d maps M_1 to a point D that is twice as far away from R as M_1, and so M_1 is the midpoint of RD. Thus M_1 is the intersection of two bisecting diagonals, A_2A_3 and RD. As we just proved, this means that the quadrilateral RA_2DA_3 is a parallelogram. Therefore

1. DA_3 is parallel to RA_2, the altitude perpendicular to A_1A_3. Hence DA_3 is perpendicular to the side A_1A_3.

2. DA_2 is parallel to RA_3, the altitude perpendicular to A_1A_2. Hence DA_2 is perpendicular to the side A_1A_2.

In other words, both $\angle A_1A_2D$ and $\angle A_1A_3D$ are right angles. According to Thales' Theorem, both A_2 and A_3 have to be on the circle with diagonal A_1D. Well, there is only one circle through the three points A_1, A_2, and A_3– it is the circumcircle $\mathcal{C}$. Therefore $D = d(M_1)$ must be on $\mathcal{C}$. It is just a matter of shuffling around the indices to show that d maps M_2 and M_3 to points of $\mathcal{C}$ as well. Furthermore, each of the segments $A_id(M_i)$ is a diameter of $\mathcal{C}$. Note that this is in keeping with the Lesson 20 proof– in that proof, we showed directly that N_iM_i is a diameter of the nine point circle. Here we see that its scaled image $d(N_iM_i) = A_id(M_i)$ is a diameter of the circumcircle.



The points L_i. The intersection of each altitude with its corresponding side forms a right angle $\angle N_iL_iM_i$. Now apply the dilation: the result, $d(\angle N_iL_iM_i)$, will still be a right angle. As we just saw, $d(N_iM_i)$ is a diameter of $\mathcal{C}$. By Thales' Theorem, $d(L_i)$ must be on $\mathcal{C}$ as well.

In conclusion, the dilation d maps the nine points L_i, M_i, and N_i to nine points of $\mathcal{C}$. In reverse, d^{-1} will map nine points of $\mathcal{C}$ to L_i, M_i, and N_i. Since d^{-1} is a Euclidean transformation, it will map the points of one circle, such as $\mathcal{C}$, to the points of another circle. Therefore L_i, M_i and N_i must all be on the same circle. □

These transformations provide a fundamentally different perspective on the problems of geometry. I hope that these few examples give you a little sense of that. Going forward, transformations will be a critical weapon in our arsenal.

References

This proof of Varignon's Theorem (certainly not the most common) is from Wallace and West's *Roads to Geometry* [3]. This proof of Napoleon's Theorem is from the geometry web site cuttheknot.org [1], although there they reference I. M. Yaglom's Geometry Transformations I. This proof of the Nine Point Circle Theorem is from Pedoe's book *Course of Geometry for Colleges and Universities* [2], which is now available under the title *Geometry, a Comprehensive Course* from Dover Publications.

[1] Alexander Bogomolny. Napoleon's theorem by transformation. distributed on World Wide Web. http://www.cut-the-knot.org/ Curriculum/Geometry/NapoleonByTransformation.shtml.

[2] Daniel Pedoe. *A Course of Geometry for Colleges and Universities*. Cambridge University Press, London, 1st edition, 1970.

[3] Edward C. Wallace and Stephen F. West. *Roads to Geometry*. Pearson Education, Inc., Upper Saddle River, New Jersey, 3rd edition, 2004.

Exercises

1. Prove that the composition of two half-turns around distinct points separated by a distance x is a translation by a distance $2x$.

2. Let r be a rotation around point P. Prove that every line through P is invariant (that is, $r(\ell) = \ell$) if and only if r is a half-turn.

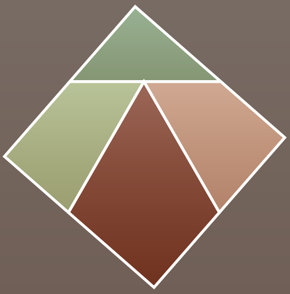
3. Given a triangle $\triangle ABC$, let

   r_A be the half-turn around the midpoint of BC,
   r_B be the half-turn around the midpoint of AC, and
   r_C be the half-turn around the midpoint of AB.

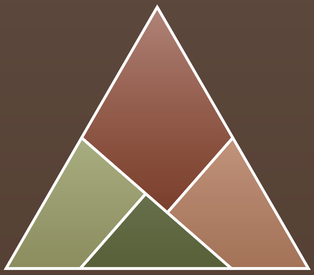
   Then $r_A \circ r_B \circ r_C$ is a half-turn as well. What is its center of rotation?

4. Show that if $ABCD$ is a parallelogram, then the composition $r_A \circ r_B \circ r_C \circ r_D$ of half-turns around A, B, C, D is the identity (the converse of what we proved in the lesson).

5. Consider a triangle $\triangle ABC$ whose vertices are located at the following coordinates: $A = (0,0)$, $B = (2,0)$, and $C = (1,3)$. Find the diameter of the circumcircle of $\triangle ABC$, and from that, the radius of the nine point circle.



32 AREA I



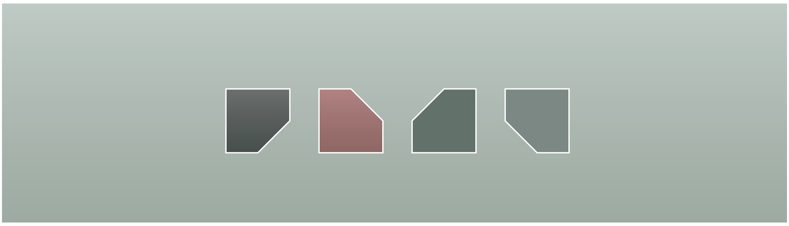
The area function

It took us long enough, but we have finally gotten around to talking about *area*. Fundamentally, when we talk about the area of a polygon, we are talking about a number, a positive real number. So you can think of area as a function from the set of all polygons to the set of positive real numbers

$$A : \{\text{polygons}\} \longrightarrow (0,\infty).$$

That's not all though– if this area function is going to live up to our expectations, it needs to meet a few other requirements as well.

1. If two polygons are congruent, their areas should be the same. This statement can also be interpreted in terms of isometries. Remember that if P is any polygon and τ is any isometry, then $\tau(P)$ and P are congruent. Therefore area should be an invariant of any isometry.



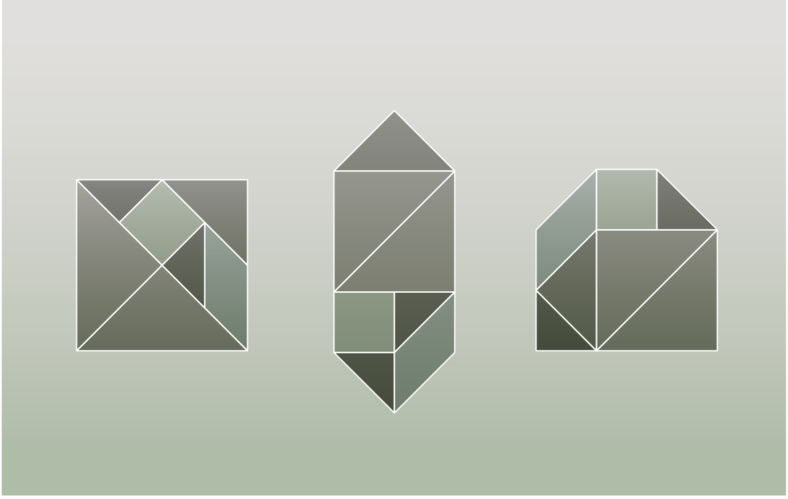
Congruent polygons have the same area.

2. If a polygon can be broken down into smaller pieces, then the area of the polygon should be the sum of the areas of the pieces. More precisely, let $\text{int}\,(P)$ denote the set of points in the interior of a polygon P, and let $\overline{P}$ denote that set of interior points together with the points on the edges of P. A set of polygons $\{P_i\}$ is a *decomposition* of P if

   $\cup\overline{P_i} = \overline{P}$ (the pieces cover P), and

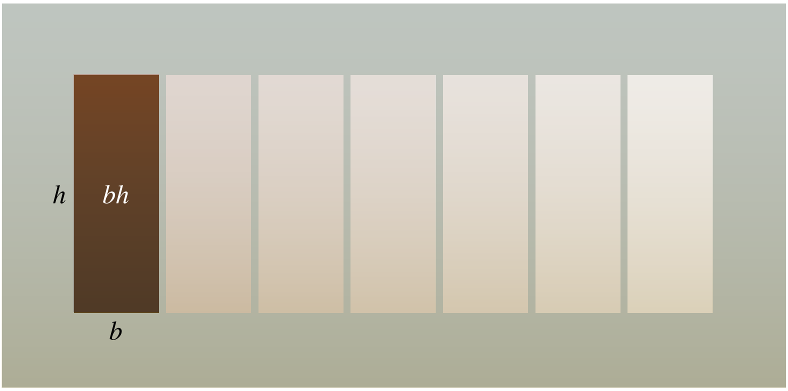
   $\text{int}\,(P_i) \cap \text{int}\,(P_j) = \emptyset$ if $i \neq j$ (the pieces don't overlap).

   In this context, if $\{P_i\}$ is a decomposition of polygon P, then $A(P)$ should equal $\sum A(P_i)$.

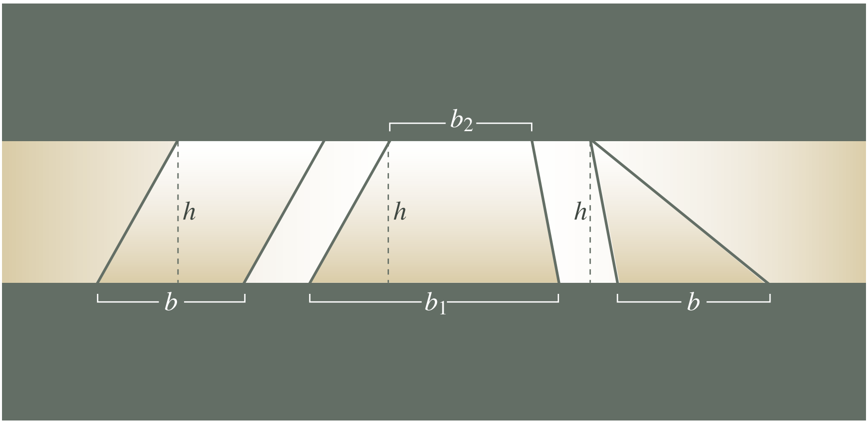


Three convex shapes. Since they can be decomposed into the same set of congruent pieces (the tangram tiles), they must have the same areas.

3. Finally, we need something to get us started, and it is this: the area of a rectangle with a base b and a height h is $A = bh$.



The congruence and decomposition conditions allow us to cut apart and rearrange polygons, starting with rectangles, to find the areas of other, more exotic shapes. We will start that process in the next few results. Because these early results are just a few steps removed from the formula for the area of a rectangle, these formulas also involve bases and heights, so let me first clarify what is meant by "base" and "height" in each of these shapes.



Parallellogram: Any side of a parallelogram can serve as its base. The height is a segment that is perpendicular to the base; one of its endpoints is on the line containing the base and the other is on the line through the opposite side. Usually you will want to use a vertex of the parallelogram as one of the endpoints for the height.

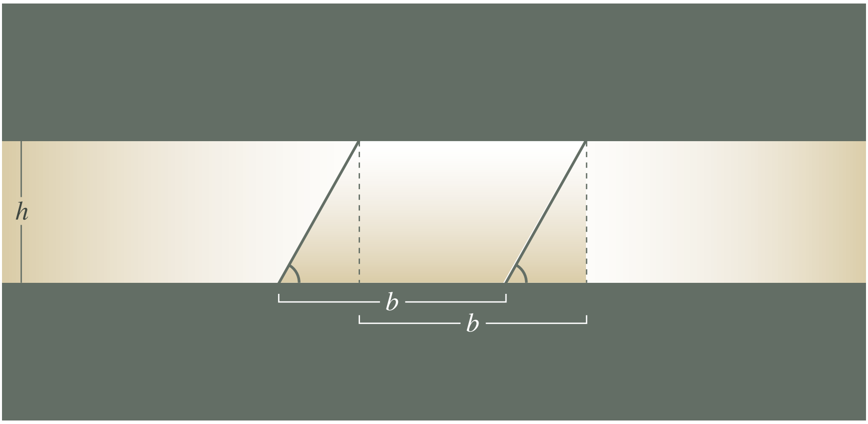
Trapezoid: The two parallel sides are both considered bases (the area formula uses them both). The height is as in the parallelogram– a segment perpendicular to, and connecting, the lines through the two parallel sides.

Triangle: Any of the sides of a triangle can serve as its base. The height is the segment from the opposite vertex to the line containing the base, perpendicular to that base (the height runs along the altitude, but I originally defined an altitude to be a line, not a segment).

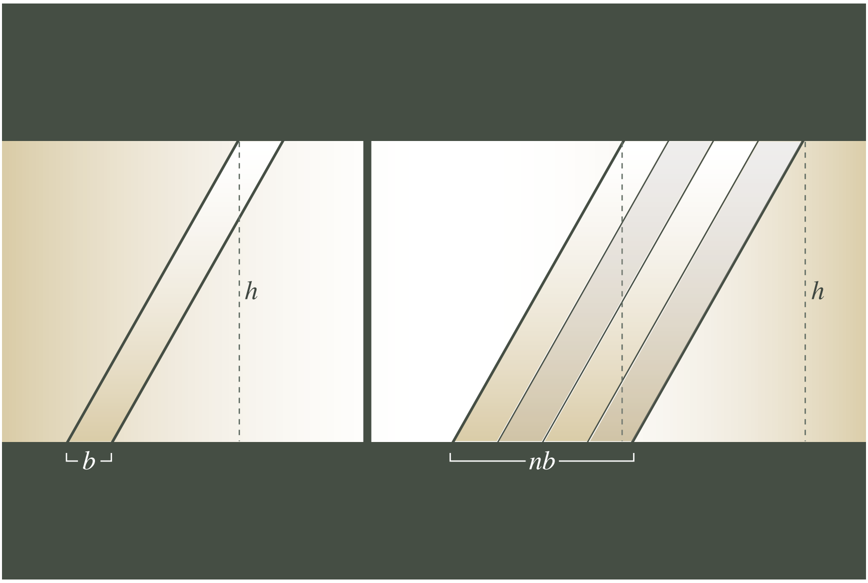
Let's start cutting and gluing to find some area formulas.

THM: AREA OF A PARALLELOGRAM
A parallelogram with base b and height h has area $A = bh$.

Proof. With a well-mannered parallelogram, you just need to cut off the triangular end and shift it to the other side. Since adjacent angles in a parallelogram are supplementary, the two pieces will fit perfectly to form a rectangle with base b and height h. Since cutting and rearranging pieces doesn't change the total area, the area of the parallelogram is the same as the area of the rectangle.



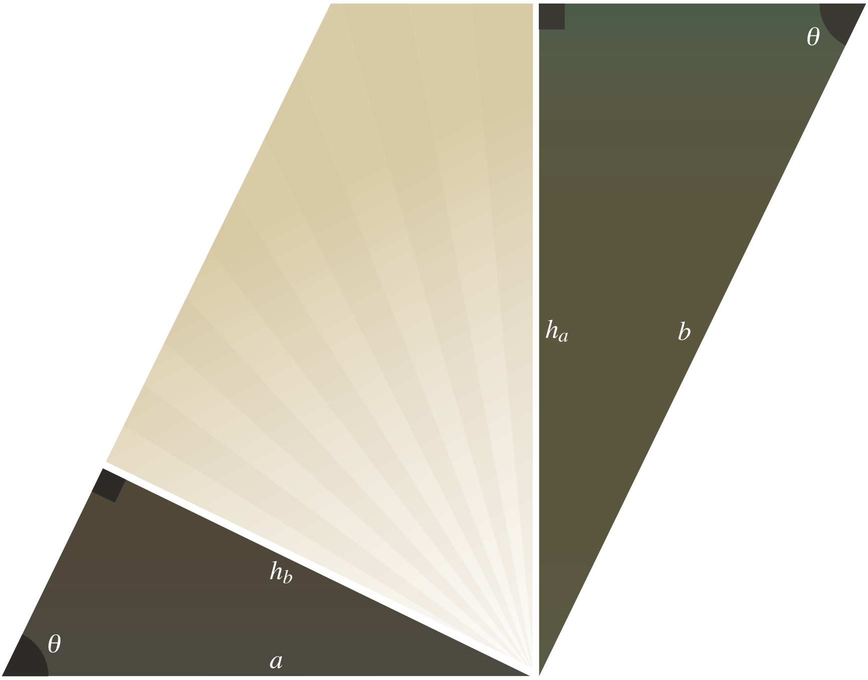
If the parallelogram is particularly narrow, this simple approach may not work– the height line along which you need to cut may slip outside the parallelogram. In this case, you can lay out congruent copies of the parallelogram next to each other to form a wider parallelogram. Do this enough times (let's say n times) and eventually the result will be wide enough to fall in the well-behaved scenario described above. It is a parallelogram with a base of nb and a height of h, so its area is $A = nbh$. It is made up of n congruent pieces, each of which must then have an area $A = (nbh)/n = bh$. □



This formula for the area of a parallelogram raises an important issue: there are two choices for what will be the base of the parallelogram (actually, any of the four sides could be the bases, so there are really are four choices, but since the opposite sides of a parallelogram are congruent, there are two *different* choices). In order for the area of a parallelogram to be well-defined, it must not depend upon which of those choices we make.

THM: THE ILLUSION OF CHOICE I
The area of a parallelogram does not depend upon the choice of base.

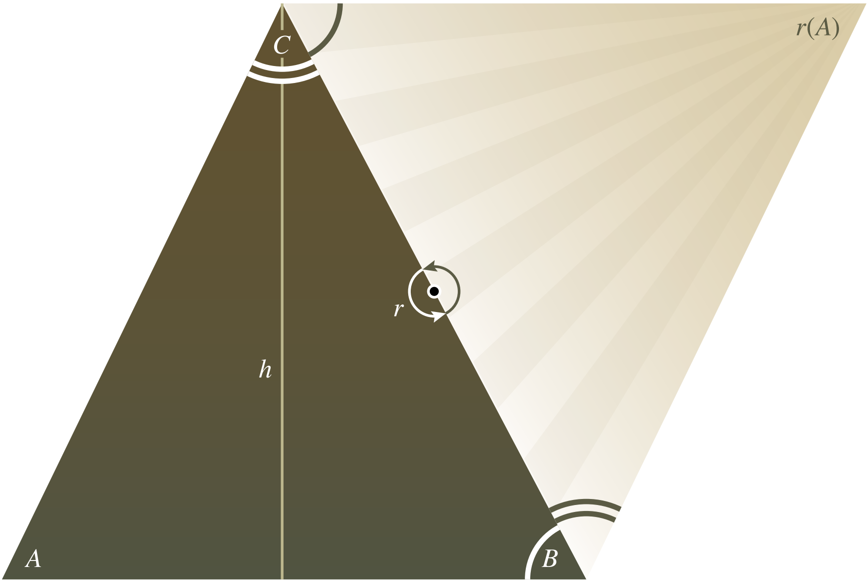
Proof. Consider a parallelogram with sides of length a and b. Let h_a be the height corresponding to the base of length a, and let h_b be the height corresponding to the base of length b.



Then we can write the area of the parallelogram as either $A = ah_a$ or $A = bh_b$. Note, though, that if θ is the angle between the sides of the parallelogram (take the acute angle for convenience), then $h_a = b\sin\theta$ and $h_b = a\sin\theta$, so either way, $A = ab\sin\theta$. □

THM: AREA OF A TRIANGLE

A triangle with base b and height h has area $A = \frac{1}{2}bh$.



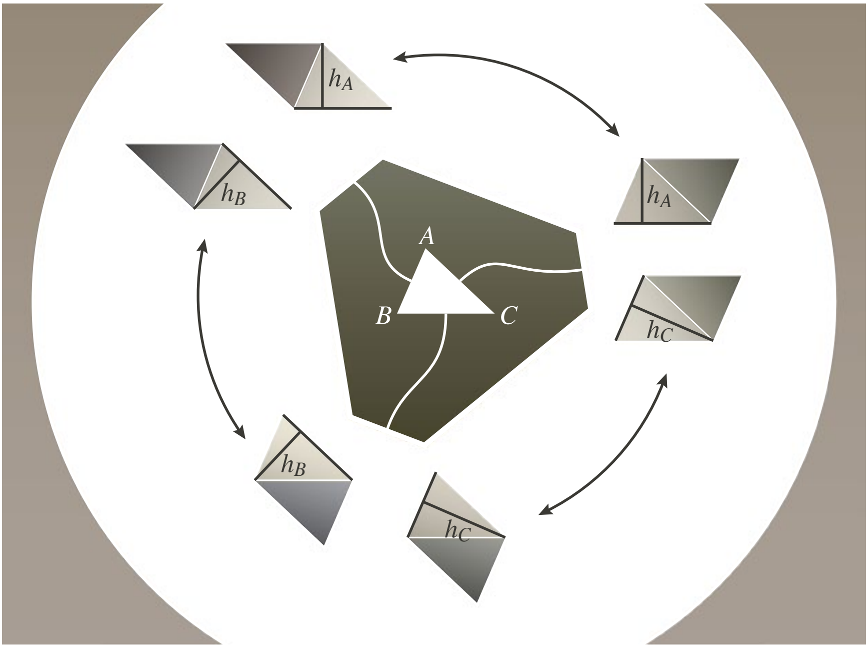
Proof. Begin with a triangle $\triangle ABC$. Identify the base b of this triangle as the segment AB, and the corresponding height h. Consider a half-turn r through the midpoint of BC. The resulting triangle $r(\triangle ABC)$ is congruent to the original and r swaps the points B and C– that means the alternate interior angles at B and C are congruent, so the sides AB and $Cr(A)$ are parallel, as are the sides AC and $Br(A)$. We have created a parallelogram! It has a base b and a height h, so its area is bh. The area of each of the two triangles forming it, then, must be half of that– they will have an area of $bh/2$. □

As with the parallelogram, this raises the issue: there is an apparent choice of base– does that choice effect the result?

THM: THE ILLUSION OF CHOICE II

The area of a triangle does not depend upon the choice of base.

Proof. Start with a triangle $\triangle ABC$. There are three choices of base here, and each can potentially lead to a different, non-congruent, parallelogram.



Label the corresponding heights:

h_A: the height associated with BC,
h_B: the height associated with AC, and
h_C: the height associated with AB.

But look more closely. The two parallelograms formed by turning across AB and AC both have base BC and height h_A, so they have the same area. And the two parallelograms formed by turning across AB and BC both have base AC and height h_B, so they too have the same area. So yes, the parallelograms may not be congruent, but they do have the same area. □

AREA OF A TRAPEZOID

A trapezoid with bases b_1 and b_2 and height h has area

$$A = \frac{b_1 + b_2}{2} \cdot h.$$

I will leave it to you to prove this one.

Laws of Sines and Cosines

Standard trigonometry provides functions that describe the relationships between the sides and angles of a right triangle. Out of the box, though, those relationships are limited to *right* triangles. The Law of Sines builds from those elementary relationships to describe some of the connections between the angles and the sides of an *arbitrary* triangle. We could have derived the Law of Sines way back when we first looked at the trigonometric functions, but we didn't. So now let's do it by thinking in terms of area.

THM: THE LAW OF SINES

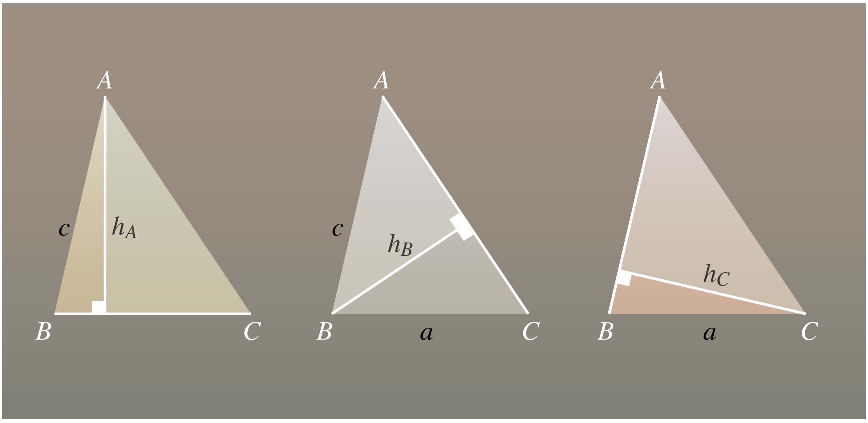
In a triangle $\triangle ABC$, let a denote the length of the side opposite $\angle A$, b denote the length of the side opposite $\angle B$, and c denote the length of the side opposite $\angle C$. Then

$$\frac{\sin A}{a} = \frac{\sin B}{b} = \frac{\sin C}{c}.$$

Proof. We know that each of the three sides of the triangle can serve as the base in the calculation of its area, and that no matter which side is chosen, the result is the same. Doing that calculation with each of the sides:

$$\tfrac{1}{2}ah_A = \tfrac{1}{2}bh_B = \tfrac{1}{2}ch_C$$

where h_A, h_B, h_C are the heights corresponding to the bases a, b, and c respectively.



Work with the first equality and note that we can write $h_A = c\sin B$ and $h_B = c\sin A$. Therefore

$$\begin{aligned}\tfrac{1}{2}ac\sin B &= \tfrac{1}{2}bc\sin A\\ a\sin B &= b\sin A\\ \frac{\sin B}{b} &= \frac{\sin A}{a}.\end{aligned}$$

That gets the first half of the Law of Sines, and working with the second equality is similarly productive: with $h_B = a\sin C$ and $h_C = a\sin B$,

$$\begin{aligned}\tfrac{1}{2}ba\sin C &= \tfrac{1}{2}ca\sin B\\ b\sin C &= c\sin B\\ \frac{\sin C}{c} &= \frac{\sin B}{b}.\end{aligned}$$

□

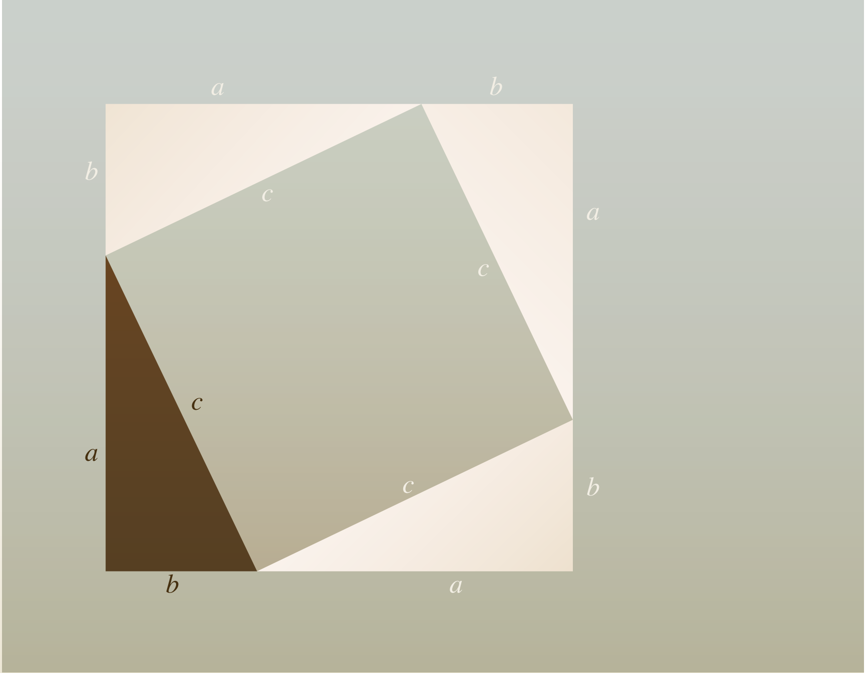
I am pretty sure that the first proof I ever saw in my life was a proof of the Pythagorean Theorem that I stumbled across while flipping through my parent's copy of Bronowski's *The Ascent of Man*. It was a proof based upon calculating the areas of triangles and squares. Of course, we have already seen one proof of the Pythagorean Theorem, but (1) the Pythagorean Theorem is fairly important; (2) this proof is personally significant to me; and (3) it suggests a way to use area to prove the Law of Cosines.

THM: THE PYTHAGOREAN THEOREM

In a right triangle with legs of length a and b, and hypotenuse of length c,

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

Proof. Position four congruent copies of the triangle around a square with sides of length c as shown. Now look at how the angles come together at each corner of the square– the two acute angles of the right triangle, and then the right angle of the square. Taken together, these three angles add up to π– that means the edges of the triangles join up in a straight line. The pieces fit perfectly to form a square with sides of length $a+b$. We can calculate the area of the big square in two ways.



1. Directly in terms of its sides:

$$(a+b)^2 = a^2 + 2ab + b^2.$$

2. By adding the areas of the center square and surrounding triangles:

$$c^2 + 4 \cdot \tfrac{1}{2}ab = c^2 + 2ab$$

Set the two equal, subtract $2ab$ from both sides, to get $c^2 = a^2 + b^2$, the Pythagorean Theorem. □

The Pythagorean Theorem only applies to right triangles. There is, however, an extension of the Pythagorean Theorem called the Law of Cosines that can be used in any triangle.

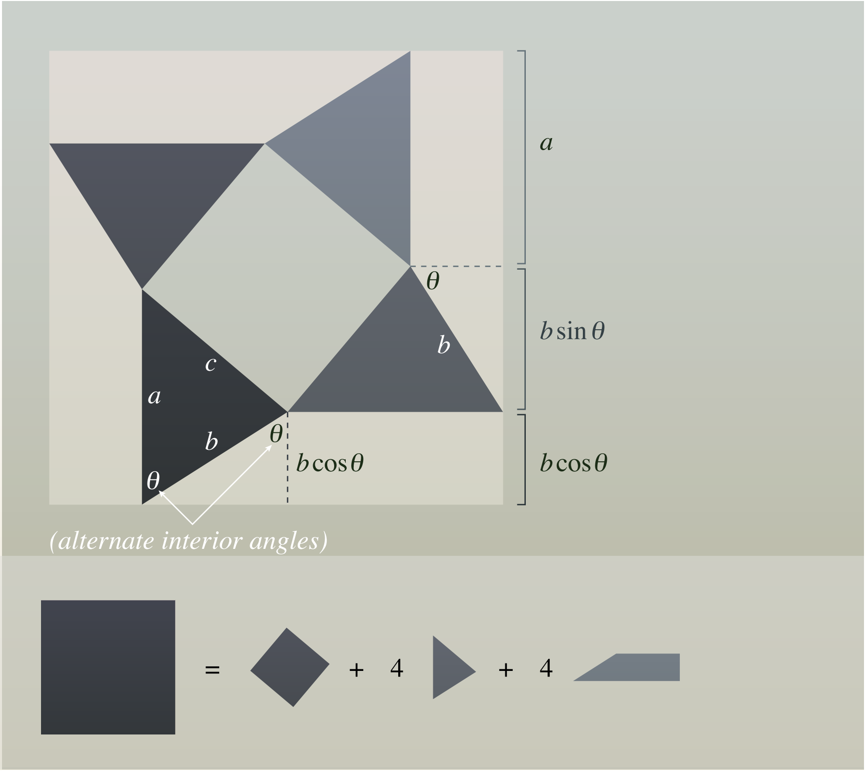
THM: THE LAW OF COSINES

Given a triangle with sides of length a, b, and c, and angle θ opposite side c,

$$c^2 = a^2 + b^2 - 2ab\cos\theta.$$

Proof. As in the last proof, what we want to do is to build four congruent copies of the triangle around a square with sides c. If $\angle\theta$ is a right angle, then the $2ab\cos\theta$ term in the equation is zero, and this really is just the Pythagorean Theorem. In terms of the proof, it is when $\angle\theta$ is right that the sides of the two neighboring triangles line up with each other to form a square. If $\angle\theta$ is not a right angle, this does not happen, and so we will have to work a little harder. That special Pythagorean arrangement neatly splits the more general problem into two cases– one when $\angle\theta$ is acute and one when $\angle\theta$ is obtuse. I will take the acute case, and leave you the obtuse case.

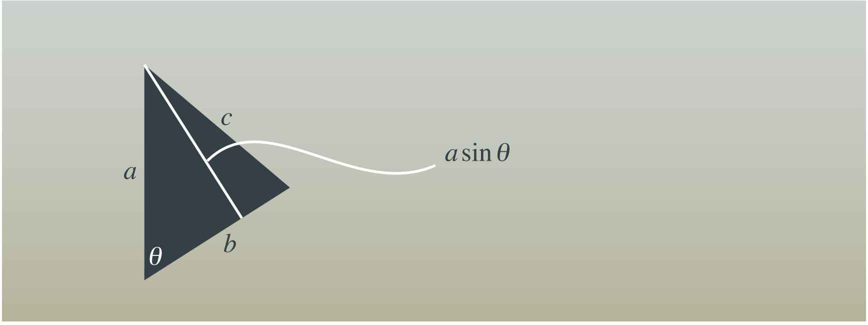
The four congruent copies of the triangle form a pinwheel shape around the square. We can build a square that frames that pinwheel by drawing lines through each pinwheel tip parallel to the "a" sides of the triangle. Since adjacent triangles in the pinwheel are turned at right angles to each other, these new lines will also intersect at right angles. So we have a big square which is divided into four trapezoids, four triangles and a smaller square. Now let's calculate the dimensions of these shapes.



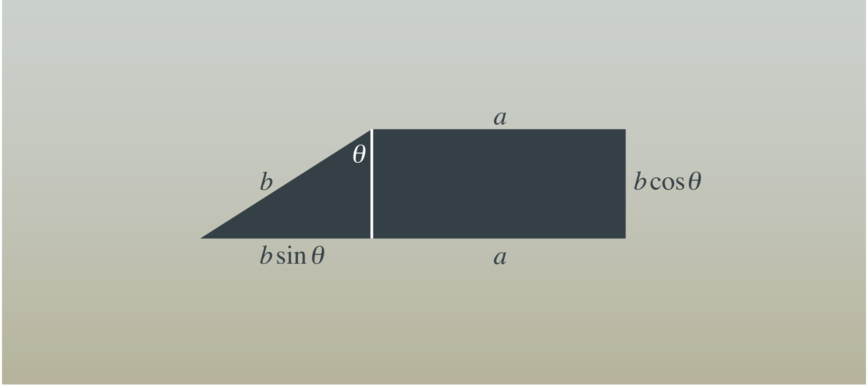
Area of the big square:

$$\begin{aligned}(a+b(\sin\theta+\cos\theta))^2 \\ &= a^2+2ab(\sin\theta+\cos\theta)+b^2(\sin\theta+\cos\theta)^2 \\ &= a^2+2ab\sin\theta+2ab\cos\theta+b^2+2b^2\sin\theta\cos\theta.\end{aligned}$$

Area of the small square: c^2.



Area of one of the four triangles: $\frac{1}{2}ab\sin\theta$.



Area of one of the four trapezoids:

$$\begin{aligned}\tfrac{1}{2}(a+(a+b\sin\theta))\cdot b\cos\theta \\ &= \tfrac{1}{2}(2ab\cos\theta+b^2\sin\theta\cos\theta) \\ &= ab\cos\theta+\frac{1}{2}b^2\sin\theta\cos\theta.\end{aligned}$$

Since the area of the whole is the sum of the areas of the parts

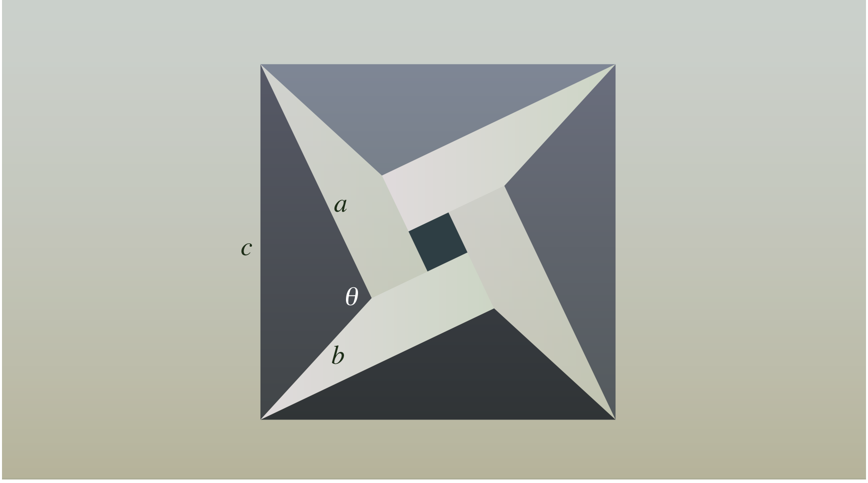
$$\begin{aligned} a^2 &+ 2ab\sin\theta + 2ab\cos\theta + b^2 + 2b^2\sin\theta\cos\theta \\ &= c^2 + 4\left(\tfrac{1}{2}ab\sin\theta\right) + 4\left(ab\cos\theta + \tfrac{1}{2}b^2\sin\theta\cos\theta\right). \end{aligned}$$

Simplify and cancel out common terms to get the Law of Cosines,

$$a^2 + b^2 - 2ab\cos\theta = c^2.$$

□

Hint: if you are interested in proving the obtuse case, then I would suggest you build the triangles *inside* the square with sides c, as shown in the following illustration, rather than out around it.



Heron's formula

To close out this lesson, I want to use the Law of Cosines to derive another formula for the area of a triangle called *Heron's Formula*. The S·S·S Triangle Congruence Theorem says that a triangle is uniquely determined by the lengths of its three sides. That means there should be a formula to calculate the area of a triangle using the just the lengths of its sides. The formula $A = bh/2$ does not do that, since it also requires a height. But Heron's Formula does.

DEF: SEMIPERIMETER

The semiperimeter s of a triangle is half its perimeter. If a triangle has sides of length a, b, and c, then its semiperimeter is

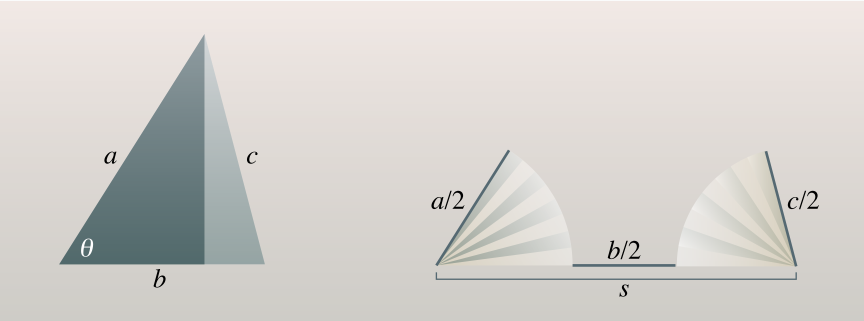
$$s = \tfrac{1}{2}(a+b+c).$$

THM: HERON'S FORMULA

The area of a triangle with sides of length a, b, and c, and semiperimeter s is

$$A = \sqrt{s(s-a)(s-b)(s-c)}.$$

Proof. This theorem is not difficult from a theoretical point of view. It is a nuisance, however, on the calculation side. Label the sides of the triangle so that side a is the base and the angle θ between a and b is acute (at least two angles in any triangle have to be acute, so this is no problem).



Then the area of the triangle is

$$A = \frac{1}{2}ab\sin\theta.$$

We want to get that θ out of the picture. The Law of Sines might seem like the obvious choice, but it always relate an {angle & side} to another {angle & side}, so it doesn't help eliminate angles entirely. The Law of Cosines does give a way to relate an angle to the three sides– that's what we need to use– so we have to write the area in terms of cosine, not sine. Use the Pythagorean Identity:

$$\sin^2\theta + \cos^2\theta = 1 \implies \sin^2\theta = 1-\cos^2\theta.$$

Normally at this point, taking the square root of both sides would yield two solutions. In this case, since I required that θ be an acute angle, $\sin\theta$ will be a positive number, and we can go with the positive root

$$\sin\theta = \sqrt{1-\cos^2\theta},$$

so the area of the triangle is

$$A = \tfrac{1}{2}ab\sqrt{1-\cos^2\theta}.$$

Now use the Law of Cosines

$$c^2 = a^2+b^2-2ab\cos\theta \implies \cos\theta = \frac{c^2-a^2-b^2}{2ab},$$

and substitute into the area formula to get a big algebra problem:

$$\begin{aligned}
A &= \tfrac{1}{2}ab\sqrt{1-\left[\frac{c^2-a^2-b^2}{2ab}\right]^2} \\
&= \tfrac{1}{2}ab\sqrt{\frac{4a^2b^2-(c^2-a^2-b^2)^2}{4a^2b^2}} \\
&= \tfrac{1}{2}ab\cdot\frac{1}{2ab}\sqrt{4a^2b^2-(c^4-2a^2c^2-2b^2c^2+a^4+2a^2b^2+b^4)} \\
&= \tfrac{1}{4}\sqrt{-(a^4-2a^2b^2+b^4)+2(b^2c^2+c^2a^2)-c^4} \\
&= \tfrac{1}{4}\sqrt{-(a^2-b^2)^2+2c^2(a^2+b^2)-c^4} \\
&= \tfrac{1}{4}\sqrt{-(a^2-b^2)^2+(a^2+b^2)^2-(a^2+b^2)^2+2c^2(a^2+b^2)-c^4} \\
&= \tfrac{1}{4}\sqrt{(-a^4+2a^2b^2-b^4+a^4+2a^2b^2+b^4)-((a^2+b^2)^2-c^2)^2}
\end{aligned}$$

$$= \tfrac{1}{4}\sqrt{4a^2b^2 - ((a^2+b^2)-c^2)^2}$$

$$= \tfrac{1}{4}\sqrt{(2ab-(a^2+b^2-c^2))(2ab+(a^2+b^2-c^2))}$$

$$= \tfrac{1}{4}\sqrt{((-a^2+2ab-b^2)+c^2)((a^2+2ab+b^2)-c^2)}$$

$$= \tfrac{1}{4}\sqrt{(c^2-(a-b)^2)((a+b)^2-c^2)}$$

$$= \tfrac{1}{4}\sqrt{(c+(a-b))(c-(a-b))((a+b)+c)((a+b)-c)}$$

$$= \sqrt{\frac{(a-b+c)(-a+b+c)(a+b+c)(a+b-c)}{16}}$$

$$= \sqrt{\frac{a-b+c}{2}\cdot\frac{-a+b+c}{2}\cdot\frac{a+b+c}{2}\cdot\frac{a+b-c}{2}}$$

$$= \sqrt{\left[\frac{a+b+c}{2}-b\right]\cdot\left[\frac{a+b+c}{2}-a\right]\cdot\left[\frac{a+b+c}{2}\right]\cdot\left[\frac{a+b+c}{2}-c\right]}$$

$$= \sqrt{(s-b)(s-a)s(s-c)}.$$

□

In this lesson, we started from area of a rectangle and worked our way down to area of a triangle. In the next lesson, we will build up from the area of a triangle to the area of polygons in general.

References

Time has apparently clouded my memory. Bronowski does discuss an area-based proof of the Pythagorean Theorem in his book [1] (pages 158-162), but it is not the one I have given here. The idea behind the proof of Heron's Formula is simple enough, but without Coxeter's *Introduction to Geometry*[2] (pages 12-13) I may have given up somewhere in the calculation.

[1] J. Bronowski. *The Ascent of Man*. Little, Brown and Company, Boston/Toronto, 1973.

[2] H.S.M. Coxeter. *Projective Geometry*. Blaisdell Publishing Co., New York, 1st edition, 1964.

Exercises

1. We took as definition that the area of a rectangle is given by the formula $A = bh$. That can be derived from a much more minimal condition– that the area of a 1×1 square is one. Derive the general formula for the area of a rectangle from this.

2. In a parallelogram with sides of length a and b, and acute interior angle θ, describe the number of strips n required to cut and form a rectangle (as described in the proof of the parallelogram area formula) in terms of a, b, and θ.

3. Prove the area formula for the trapezoid.

4. The Penrose tiles are a pair of rhombuses that in conjunction tile the Euclidean plane in an aperiodic manner (they are named after Roger Penrose, who discovered their aperiodicity). One of the tiles has interior angles measuring $2\pi/5$ and $3\pi/5$. The other has interior angles measuring $\pi/5$ and $4\pi/5$. Find the areas of the two Penrose tiles assuming the lengths of each side of a tile is 1.

5. Consider a triangle $\triangle ABC$ whose vertices are located at the following coordinates: $A = (0,0)$, $B = (3,1)$, and $C = (4,2)$. Calculate the area of $\triangle ABC$, first using the formula $A = bh/2$, then using Heron's Formula.

6. Find a formula for the area of an equilateral triangle in terms of the length s of one of its sides.

7. Prove the obtuse case of the Law of Cosines.

8. Consider a regular n-sided polygon inscribed in a circle with radius r. By dividing the polygon into triangles, find a formula for its area in terms of n and r. [We will take a different approach to this problem in the next lesson].

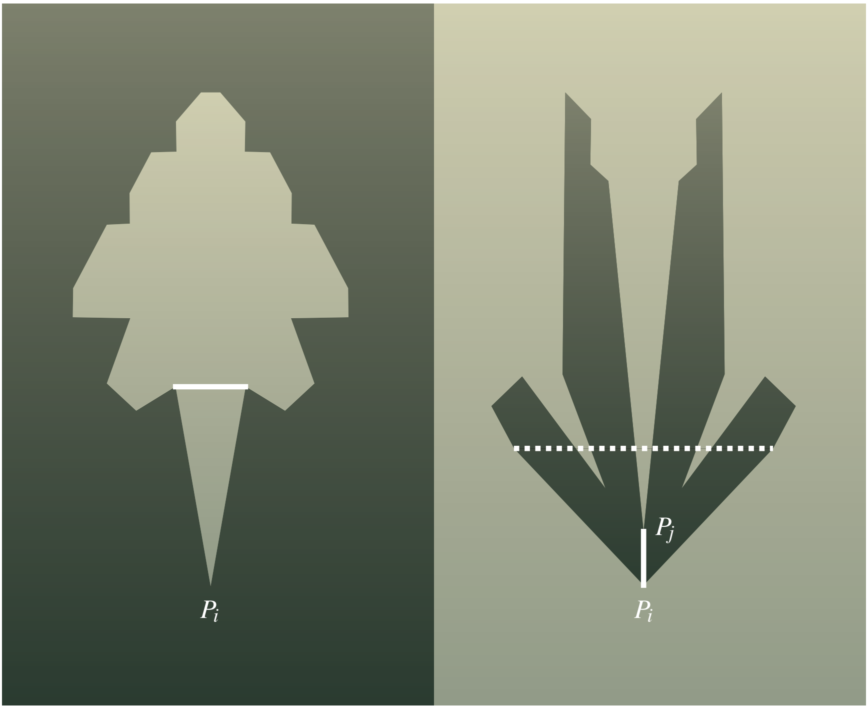
33 AREA II

Areas of polygons

The goal of this section is to establish a formula for the area of a general simple polygon. Ultimately, we will use a proof by induction to prove this formula. We need two things to make this proof work. (1) We need a way to decompose a polygon into smaller pieces– this is handled by the next result, which states that any simple polygon has a diagonal that cuts it into two smaller pieces. (2) We need a working formula for the "base" case– the area of a triangle. We found a few formulas for the area of a triangle in the last lesson, but none of those are really appropriate for this problem, so we will derive another one, this time in terms of the coordinates of its vertices. Those two steps are the hard work of this section– once they are done, it is easy to slot those pieces into the induction proof.

THM: EXISTENCE OF A DIAGONAL
Every simple polygon $\mathcal{P}$ has a diagonal that lies entirely in its interior.

Proof. If $\mathcal{P}$ is convex, then any diagonal will work. If $\mathcal{P}$ is not convex, the situation becomes a little more complicated– some of the diagonals will not be contained entirely in $\mathcal{P}$. We need to show, then, that even the most contorted polygon has at least one diagonal that lies entirely inside it. To do that, let's consider the coordinates of $\mathcal{P}$– we are looking for the "lowest" point on the polygon– the vertex with the smallest y-coordinate. Call this point P_i. Now consider the segment that connects P_i's two neighbors, P_{i-1} and P_{i+1}. If $P_{i-1}P_{i+1}$ lies entirely inside of $\mathcal{P}$, then we have found our diagonal, easy enough.



$P_{i-1}P_{i+1}$ is a diagonal.

$P_{i-1}P_{i+1}$ is not, but P_iP_j is.

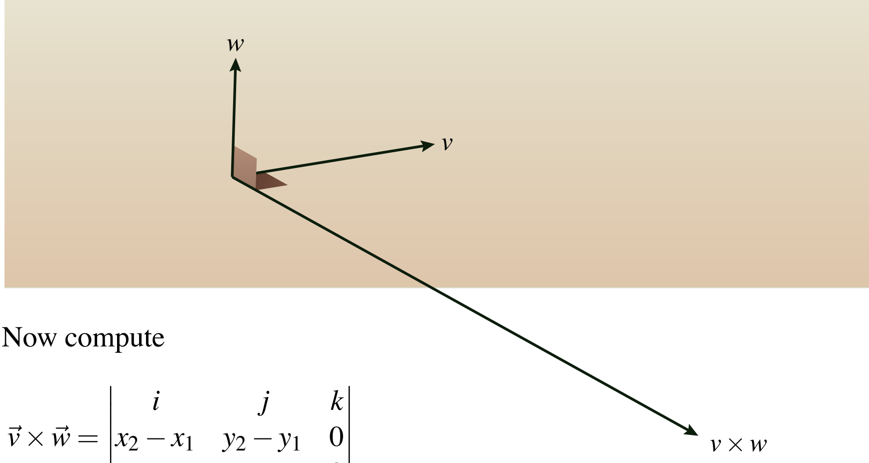
What if it doesn't? In that case, it is because at least some of the remaining vertices of $\mathcal{P}$ lie inside the triangle $\triangle P_{i-1}P_iP_{i+1}$. From this subset of vertices, let P_j be the lowest one– the one with the smallest y-coordinate. I claim that the segment P_iP_j lies entirely inside $\mathcal{P}$, so that it can serve as our diagonal. To see why, you need to remember that a point Q is inside a polygon $\mathcal{P}$ if any ray from Q crosses the polygon an odd number of times (counting multiplicities). In this case, if Q is any point on P_iP_j, it is lower than any of the vertices of $\mathcal{P}$ except for P_i, and possibly P_{i-1} or P_{i+1}. Therefore the ray $QP_i \rightarrow$ only intersects the sides $P_{i-1}P_i$ and P_iP_{i+1} once at the shared endpoint P_i, and it does not intersect any of the other sides of $\mathcal{P}$ at all. Since P_j is inside the triangle $\triangle P_{i-1}P_iP_{i+1}$, the ray $QP_i \rightarrow$ splits the polygon at P_i (the adjacent vertices P_{i-1} and P_{i+1} are separated by the line P_iP_j). Therefore, there is one intersection of $QP_i \rightarrow$ with $\mathcal{P}$ and it has multiplicity one– that's an odd number of intersections, so Q is inside $\mathcal{P}$. That is true for all points on the segment P_iP_j, so P_iP_j is a diagonal that lies entirely inside $\mathcal{P}$. □

Now let's go back to the question of the area of a triangle. Let me first try to motivate this new area formula from the perspective of vector calculus. For any two three-dimensional vectors $\vec{v} = \langle v_x, v_y, v_z \rangle$ and $\vec{w} = \langle w_x, w_y, w_z \rangle$, the cross product $\vec{v} \times \vec{w}$ is given by the determinant

$$\vec{v} \times \vec{w} = \begin{vmatrix} i & j & k \\ v_x & v_y & v_z \\ w_x & w_y & w_z \end{vmatrix}.$$

Furthermore, it is a well-known fact from calculus that the length of $\vec{v} \times \vec{w}$ is the area of the parallelogram formed by $\vec{v}$ and $\vec{w}$, so half of that would be the area of the triangle with sides $\vec{v}$ and $\vec{w}$. Let's use that idea to calculate the area of the triangle with vertices at (x_1, y_1), (x_2, y_2), and (x_3, y_3). We can make vectors out of two of the sides and embed them in 3-dimensional space by setting the last coordinate equal to zero:

$$\vec{v} = \langle x_2 - x_1, y_2 - y_1, 0 \rangle \quad \& \quad \vec{w} = \langle x_3 - x_1, y_3 - y_1, 0 \rangle.$$



Now compute

$$\begin{aligned}
\vec{v} \times \vec{w} &= \begin{vmatrix} i & j & k \\ x_2 - x_1 & y_2 - y_1 & 0 \\ x_3 - x_1 & y_3 - y_1 & 0 \end{vmatrix} \\
&= [(x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1)]k \\
&= [(x_2y_3 - x_2y_1 - x_1y_3 + x_1y_1) - (x_3y_2 - x_3y_1 - x_1y_2 + x_1y_1)]k \\
&= [(x_1y_2 - x_2y_1) + (x_2y_3 - x_3y_2) + (x_3y_1 - x_1y_3)]k \\
&= \left(\begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix} + \begin{vmatrix} x_2 & y_2 \\ x_3 & y_3 \end{vmatrix} + \begin{vmatrix} x_3 & y_3 \\ x_1 & y_1 \end{vmatrix} \right) k.
\end{aligned}$$

It is easy to read off the length of $\vec{v}\times\vec{w}$, and half that amount gets you a formula for the area of a triangle. And while all of this may be familiar to you, it does take us out of the plane, and it does draw upon some facts about vectors that we have not yet developed. So let me give a more elementary proof of this formula.

THM: DETERMINANT FORMULA FOR THE AREA OF A TRIANGLE
Label the three vertices of a triangle in counterclockwise order: $P_1 = (x_1, y_1)$, $P_2 = (x_2, y_2)$, and $P_3 = (x_3, y_3)$. The area of $\triangle P_1P_2P_3$ is

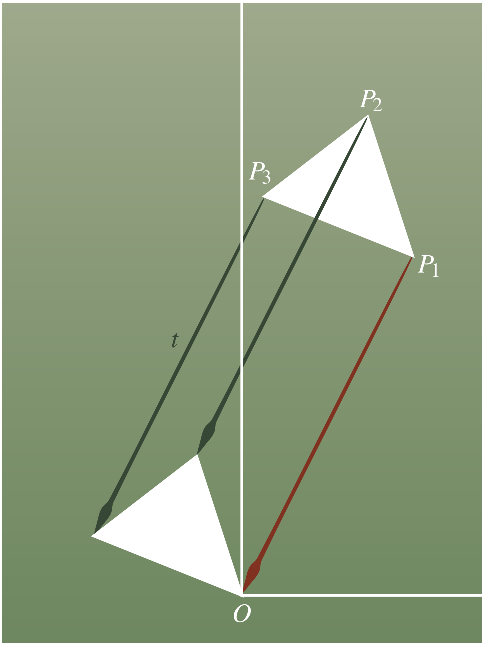
$$A = \frac{1}{2}\left(\begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix} + \begin{vmatrix} x_2 & y_2 \\ x_3 & y_3 \end{vmatrix} + \begin{vmatrix} x_3 & y_3 \\ x_1 & y_1 \end{vmatrix}\right).$$

Proof. Designate the side P_1P_2 to be the base of the triangle, and put $b = |P_1P_2|$. With the right isometry (and remember that isometries do not alter areas of shapes), we can reposition the triangle so that its base lies along the x-axis. Then it is easy to read off the height. The necessary isometry is composed of two pieces.

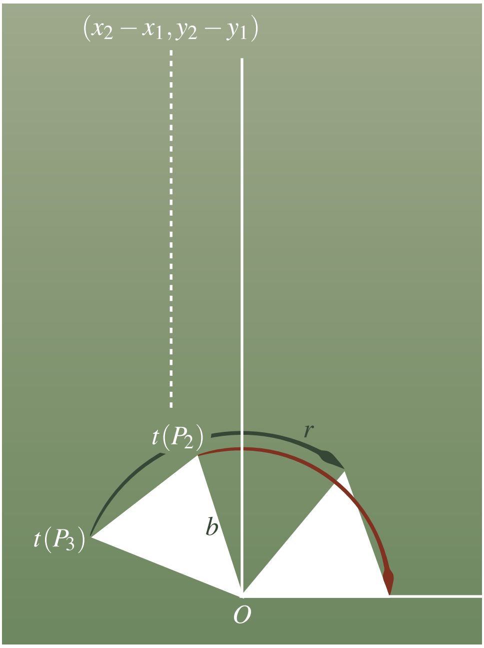
1) The first piece is a translation t to move P_1 to the origin:

$$t\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix}$$

$$t(P_1) = (0,0)$$
$$t(P_2) = (x_2 - x_1, y_2 - y_1)$$
$$t(P_3) = (x_3 - x_1, y_3 - y_1)$$



2) The second piece is a rotation r about the origin to move $t(P_2)$ onto the x-axis. To find the angle for this rotation, look at the angle θ between the x-axis and the line from the origin through $t(P_2)$.



In particular, the sine and cosine values of this angle are

$$\cos\theta = \frac{x_2 - x_1}{b} \quad \& \quad \sin\theta = \frac{y_2 - y_1}{b}.$$

In order to put the base of the triangle along the x-axis, then, we need to rotate by $-\theta$. The matrix equation for that rotation is

$$r\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos(-\theta) & -\sin(-\theta) \\ \sin(-\theta) & \cos(-\theta) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

$$= \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

$$= \frac{1}{b}\begin{pmatrix} x_2 - x_1 & y_2 - y_1 \\ y_1 - y_2 & x_2 - x_1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

The point $t(P_1)$ stays at the origin, while the point $t(P_2)$ rotates around to $(b,0)$. The key to finding the height of the triangle, though, lies with the third point:

$$\begin{aligned}
r\circ t(P_3) &= \frac{1}{b}\begin{pmatrix} x_2-x_1 & y_2-y_1 \\ y_1-y_2 & x_2-x_1 \end{pmatrix}\begin{pmatrix} x_3-x_1 \\ y_3-y_1 \end{pmatrix} \\
&= \frac{1}{b}\begin{pmatrix} [x_2-x_1][x_3-x_1]+[y_2-y_1][y_3-y_1] \\ [x_3-x_1][y_1-y_2]+[x_2-x_1][y_3-y_1] \end{pmatrix}.
\end{aligned}$$

Since $r\circ t$ is a composition of a rotation and a translation, it is orientation-preserving. Since the points P_1, P_2, P_3 are listed in counterclockwise order, their images under $r\circ t$ must also be in counterclockwise order. That means $r\circ t(P_3)$ must lie *above* the x-axis, and so the height of the triangle is just the y-coordinate of $r\circ t(P_3)$:

$$h = \frac{1}{b}\left[(x_3-x_1)(y_1-y_2)+(x_2-x_1)(y_3-y_1)\right].$$

The rest is algebra

$$\begin{aligned}
A &= \frac{1}{2}bh \\
&= \frac{1}{2}b\cdot\frac{1}{b}\left[(x_3-x_1)(y_1-y_2)+(x_2-x_1)(y_3-y_1)\right] \\
&= \frac{1}{2}\left(x_3y_1-x_1y_1-x_3y_2+x_1y_2+x_2y_3-x_2y_1-x_1y_3+x_1y_1\right) \\
&= \frac{1}{2}\left((x_1y_2-x_2y_1)+(x_2y_3-x_3y_2)+(x_3y_1-x_1y_3)\right) \\
&= \frac{1}{2}\left(\begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \end{vmatrix}+\begin{vmatrix} x_2 & y_2 \\ x_3 & y_3 \end{vmatrix}+\begin{vmatrix} x_3 & y_3 \\ x_1 & y_1 \end{vmatrix}\right).
\end{aligned}$$

□

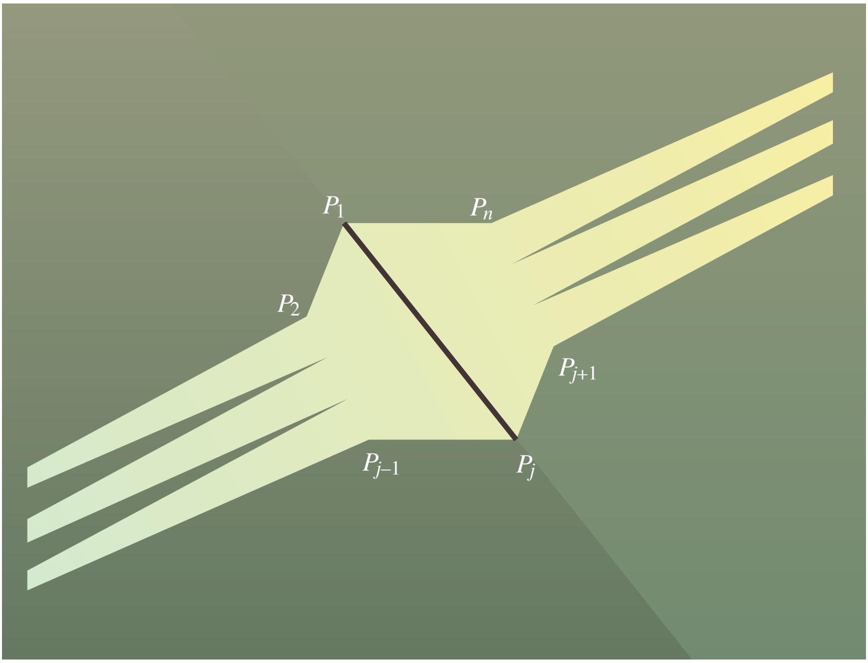
With this new area formula in hand, we can now turn back to the bigger question of polygon area.

THM: AREA OF A POLYGON

Let $P_1 = (x_1, y_1), P_2 = (x_2, y_2), P_3 = (x_3, y_3), \ldots, P_n = (x_n, y_n)$ be the coordinates of the vertices of simple polygon listed in counterclockwise order For notational convenience, put $x_{n+1} = x_1$ and $y_{n+1} = y_1$. Then the area of the polygon is

$$A = \frac{1}{2} \sum_{k=1}^{n} \begin{vmatrix} x_k & y_k \\ x_{k+1} & y_{k+1} \end{vmatrix}.$$

Proof. The proof I will give uses induction on n, the number of sides of the polygon. In the base case, when $n = 3$, the polygons are just triangles, and the area formula given here is really just the coordinate formula for triangular area that we proved above. Now move to the inductive step: suppose that this formula does give the proper area for every polygon with at most n sides, and let $\mathcal{P}$ be an arbitrary polygon with $n + 1$ sides. As we saw at the start of the lesson, $\mathcal{P}$ can be cut in two along a diagonal Δ.

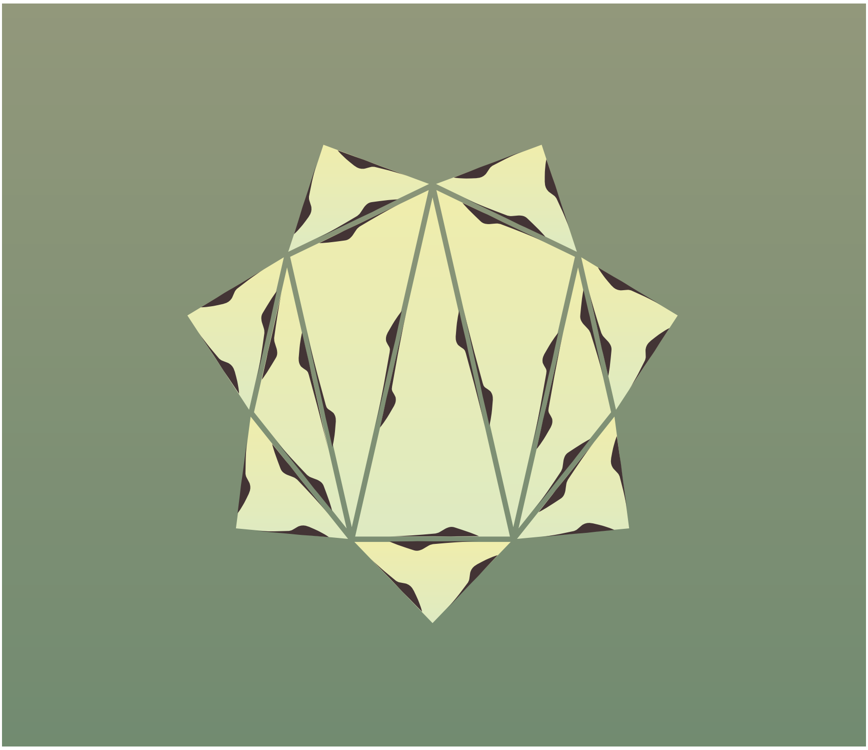


In the interest of keeping indices as simple as possible, let's relabel the points P_i and the corresponding coordinates (x_i, y_i) so that one end of Δ is the vertex $P_1 = (x_1, y_1)$. Continue around $\mathcal{P}$ in the counterclockwise direction, labeling the remaining vertices $P_2 = (x_2, y_2)$, $P_3 = (x_3, y_3)$, ..., $P_n = (x_n, y_n)$ and then loop back around to the start by setting $x_{n+1} = x_1$ and $y_{n+1} = y_1$. At some point along the way, we come to the other end of Δ. Identify that point as $P_j = (x_j, y_j)$. With those labels, Δ cuts $\mathcal{P}$ into two smaller polygons with at most n sides, $P_1P_2...P_j$ and $P_jP_{j+1}...P_nP_1$. The area of $\mathcal{P}$ is the sum of the areas of these two pieces, and by the induction hypothesis we know the area formula works for both of those pieces. Therefore

$$\begin{aligned}
A(\mathcal{P}) &= A(P_1P_2...P_j) + A(P_jP_{j+1}...P_nP_1) \\
&= \frac{1}{2}\sum_{k=1}^{j-1} \begin{vmatrix} x_k & y_k \\ x_{k+1} & y_{k+1} \end{vmatrix} + \frac{1}{2}\begin{vmatrix} x_j & y_j \\ x_1 & y_1 \end{vmatrix} \\
&\quad + \frac{1}{2}\sum_{k=j}^{n} \begin{vmatrix} x_k & y_k \\ x_{k+1} & y_{k+1} \end{vmatrix} + \frac{1}{2}\begin{vmatrix} x_1 & y_1 \\ x_j & y_j \end{vmatrix} \\
&= \frac{1}{2}\sum_{k=1}^{n} \begin{vmatrix} x_k & y_k \\ x_{k+1} & y_{k+1} \end{vmatrix} + \frac{1}{2}(x_jy_1 - x_1y_j) + \frac{1}{2}(x_1y_j - x_jy_1) \\
&= \frac{1}{2}\sum_{k=1}^{n} \begin{vmatrix} x_k & y_k \\ x_{k+1} & y_{k+1} \end{vmatrix}.
\end{aligned}$$

By induction, the formula holds for all polygons. □

What's really going on here is that by repeatedly cutting $\mathcal{P}$ along diagonals, we can eventually break $\mathcal{P}$ down into a bunch of triangles– we can "triangulate" $\mathcal{P}$. The area of each triangle in the triangulation is calculated by three determinants, one for each edge of the triangle. Different triangulations lead to different edges, but (and this is key) each internal edge is actually an edge of *two* triangles, and if the counterclockwise orientation of one triangle points it in the direction from v_i to v_j, then the counterclockwise orientation of the other triangle points it in the direction from v_j to v_i.



Each internal edge is shared by two neighboring triangles, but is oriented oppositely in those triangles. In the overall area calculation, those components cancel one another.

In the end, that internal edge makes a contribution to the overall area computation of

$$\begin{vmatrix} x_i & y_i \\ x_j & y_j \end{vmatrix} + \begin{vmatrix} x_j & y_j \\ x_i & y_i \end{vmatrix} = (x_i y_j - x_j y_i) + (x_j y_i - x_i y_j) = 0.$$

The contributions of all the internal edges cancel out, leaving just the contributions from the edges of the polygon! Note that this is what happens along the internal edge Δ in the proof above. If you have studied multivariable calculus, this internal cancellation may seem familiar. This same thing happens in Green's Theorem, where a double integral across a region is converted to a line integral around the region. In fact, this area formula is a special case of Green's Theorem– this connection is explored more thoroughly in the exercises.

The area of a circle

Time to hit another big landmark– the area of a circle. There are several ways to derive this famous area formula, but of course I want to incorporate the coordinate formula we just derived. First we will use this formula to find the area of a regular polygon. Then we will trap the circle between circumscribed and circumscribing regular polygons, and use their areas as upper and lower bounds for the area of the circle (as we did in the derivation of the circumference formula in Lesson 17).

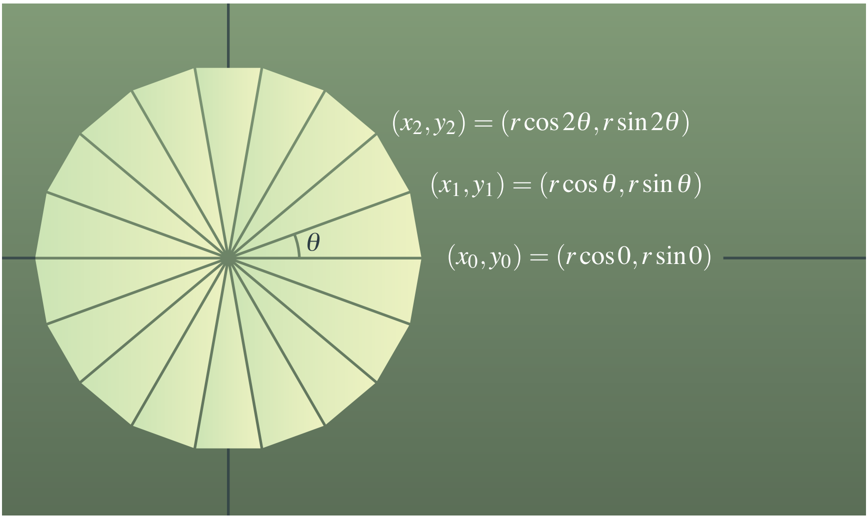
AREA OF A REGULAR POLYGON

Let $\mathcal{P}$ be a regular polygon with n sides and a radius of r (this is the radius of the circumscribing circle). Then the area of $\mathcal{P}$ is

$$A = \frac{1}{2}nr^2 \sin\left(\frac{2\pi}{n}\right).$$

Proof. All regular polygons with the same radius and the same number of sides are congruent, so we will just the one that is easiest, and that is when $\mathcal{P}$ is centered at the origin with its n vertices at the coordinates

$$\left(r\cos\left(\frac{2\pi k}{n}\right), r\sin\left(\frac{2\pi k}{n}\right)\right), \ \ 1 \le k \le n.$$


$(x_2, y_2) = (r\cos 2\theta, r\sin 2\theta)$
$(x_1, y_1) = (r\cos\theta, r\sin\theta)$
θ
$(x_0, y_0) = (r\cos 0, r\sin 0)$

It makes for easier reading if we put $\theta_x = 2\pi x/n$. Then the area of $\mathcal{P}$ is

$$A = \frac{1}{2}\sum_{k=1}^{n} \begin{vmatrix} r\cos\theta_k & r\sin\theta_k \\ r\cos\theta_{k+1} & r\sin\theta_{k+1} \end{vmatrix}$$

$$= \frac{1}{2}\sum_{k=1}^{n} \left(r^2\cos\theta_k \sin\theta_{k+1} - r^2 \sin\theta_k \cos\theta_{k+1}\right).$$

Factor out the r^2 and play around with the signs, using the fact that sine is an odd function and that cosine is an even one, to get this into the right form to use the addition formula for sine:

$$A = \frac{1}{2}\sum_{k=1}^{n} r^2[\cos(-\theta_k)\sin(\theta_{k+1}) + \sin(-\theta_k)\cos(\theta_{k+1})].$$

$$= \frac{1}{2}\sum_{k=1}^{n} r^2\sin(\theta_{k+1} - \theta_k)$$

$$= \frac{1}{2}\sum_{k=1}^{n} r^2\sin(2\pi/n)$$

$$= \frac{1}{2}n\cdot r^2\sin(2\pi/n).$$

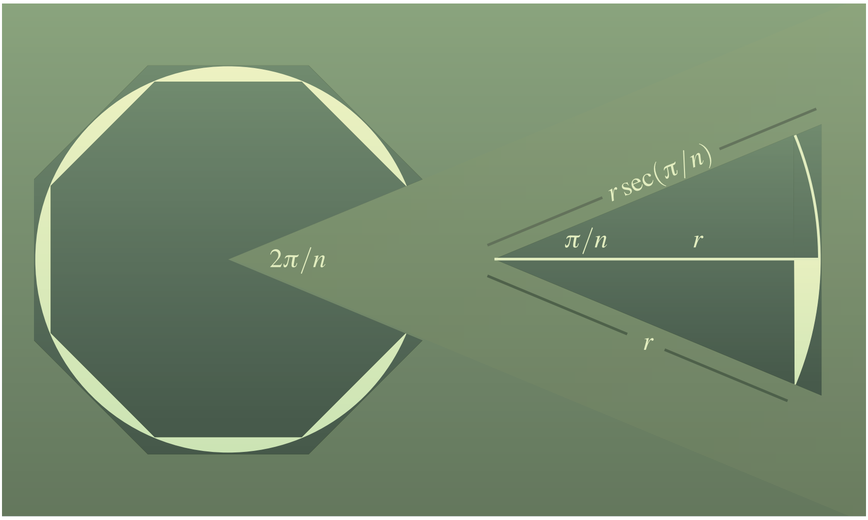
□

By trapping a circle between circumscribed and circumscribing regular polygons, it is possible to pin down its area.

THM: AREA OF A CIRCLE
The area of a circle with radius r is $A = \pi r^2$.

Proof. The radius of the inscribed regular polygon is r. The radius of the circumscribed regular polygon is $r\sec(\pi/n)$ (as illustrated). By plugging those radii into the area formula we just derived, we get upper and lower bounds on the area of the circle itself:

$$\frac{1}{2}nr^2\sin\left(\frac{2\pi}{n}\right) \le A \le \frac{1}{2}n\,(r\sec\pi/n)^2\sin\left(\frac{2\pi}{n}\right).$$



This set of inequalities is true for all n, so let's see what happens to those pieces when we take the limit as n approaches ∞:

(1) π/n approaches 0, so $\sec(\pi/n)$ approaches 1.

(2) for the term $n\sin(2\pi/n)$, make the substitution $m = n/2$. As n approaches ∞, so does m, and so

$$\lim_{n\to\infty} n\sin(2\pi/n) = \lim_{m\to\infty} 2m\sin(\pi/m) = 2\lim_{m\to\infty} m\sin(\pi/m).$$

Back in the lesson on circumference, this limit was the very definition of π (although we were using degrees instead of radians at the time), so this term is approaching 2π. Now let's put it back together:

$$\lim_{n\to\infty} \frac{1}{2}nr^2\sin\left(\frac{2\pi}{n}\right) \le A \le \lim_{n\to\infty}\frac{1}{2}n\left(r\sec\pi/n\right)^2\sin\left(\frac{2\pi}{n}\right).$$

$$\frac{1}{2}2\pi r^2 \le A \le \frac{1}{2}2\pi r^2.$$

Therefore A is trapped between two values that are both closing in upon πr^2. That means A itself must be πr^2. □

Exercises

1. Use the determinant formula for area from this lesson to find the area of the triangle with vertices at $(0,0)$, $(3,1)$, and $(4,2)$.

2. Find the area of a regular five pointed star (Scläfli symbol $\{5/2\}$) inscribed in a circle of radius one.

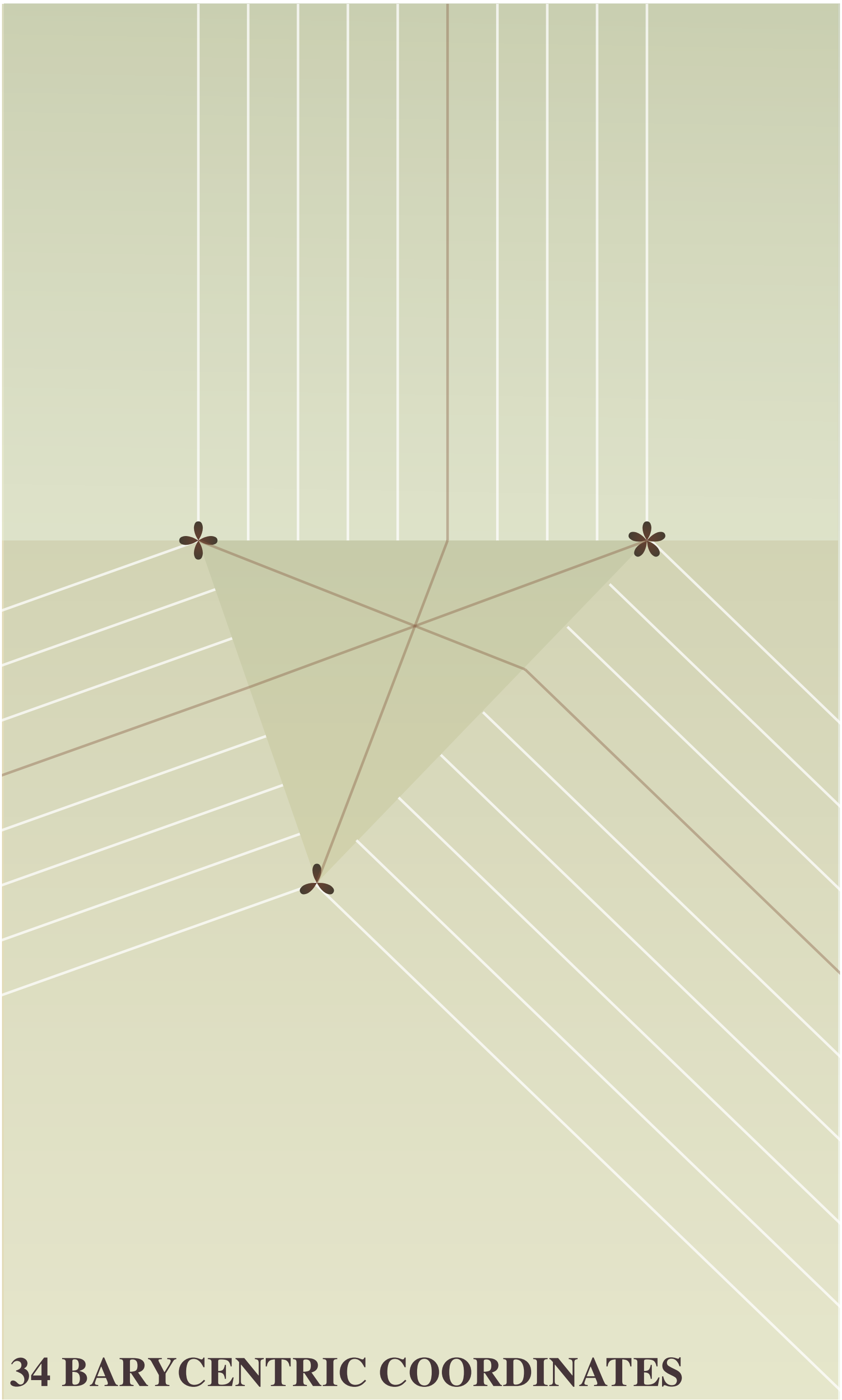
3. Give an inductive proof that any simple polygon $\mathcal{P}$ can be triangulated.

4. An alternate proof of the formula for the area of the circle involves cutting n congruent pie pieces, and then rearranging those pieces into an approximate parallelogram. Work out the details of this approach.

5. (For those who have studied Green's Theorem in calculus) Let $\mathcal{P}$ be a simple n-sided polygon with vertices (taken in the counterclockwise direction) at coordinates (x_i, y_i), for $1 \leq i \leq n$. Use Green's Theorem to show that the area of $\mathcal{P}$ is given by the integral

$$\frac{1}{2}\oint_{\mathcal{P}} -y\,dx + x\,dy.$$

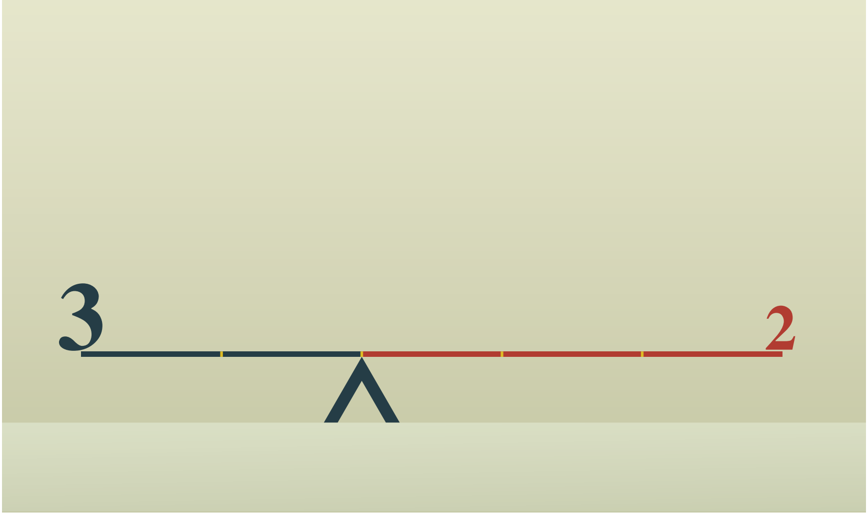
Now parametrize each edge of $\mathcal{P}$, and compute this integral to get the area formula given in the lesson.



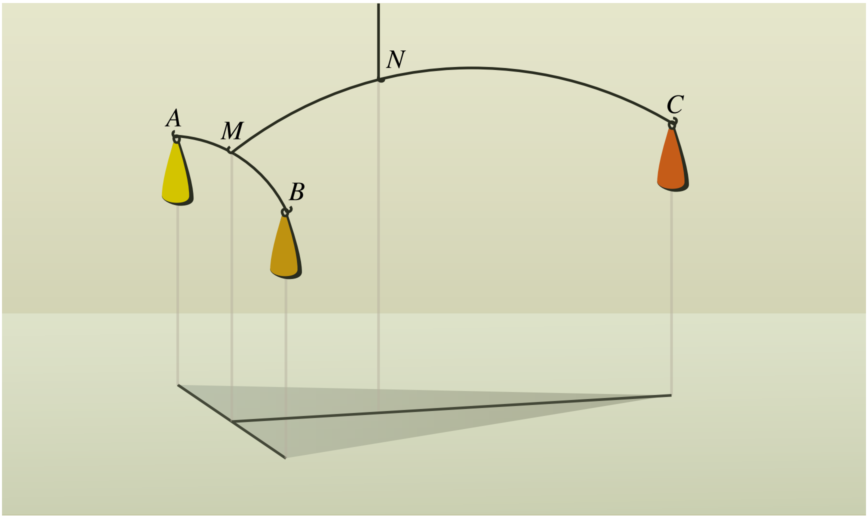
34 BARYCENTRIC COORDINATES

In Lesson 22 we studied the trilinear coordinate system. At the time, I postponed discussion of the closely-related barycentric coordinate system, because we hadn't yet dealt with area. Now that we have looked at area, we can get some closure on this topic. Barycentric coordinates are closely connected to the idea of the center of mass – the balance point of a set of weights. Archimedes has the first word on this topic that is near and dear to heart of every kindergarten kid.

> [The principle of the lever] Place two masses m_1 and m_2 on a seesaw at distances d_1 and d_2 from the fulcrum. The seesaw balances if
> $$m_1 d_1 = m_2 d_2.$$



Archimedes' lever is essentially a one-dimensional construct– the points and the fulcrum are all on one line. For the two-dimensional case, with points lying in the plane, I think of the work of a more contemporary figure– the mobiles of Alexander Calder. Let's think about how he (or some other mobile maker) would build a very simple mobile – one that consists of just three equal weights located at points A, B, and C, and wired together like this:



From a mathematical point of view, the interesting questions are: (1) where should he put the hook M so that A and B balance?, and (2) where should he put the hook N so that everything balances when the mobile is hung from the string? The answer to question (1) is easy: since the two weights are the same, the principle of the lever says that M needs to be at the midpoint of AB. The answer to question (2) is just a bit more involved: since M is now supporting twice the weight of C, the principle of the lever says that the distance from N to C must be twice the distance from N to M. In other words, N must be located two-thirds of the way down CM from C– it is at the centroid of the triangle $\triangle ABC$. Now this was a simple system since all three weights were the same, but imagine if we changed those weights so that they were not all the same. The corresponding balance point of the system (the N) would move as well. This is the key to barycentric coordinates– by putting different weights at A, B, and C, we get different balance points. The barycentric coordinates of a point P are the weights that make P the balance point.

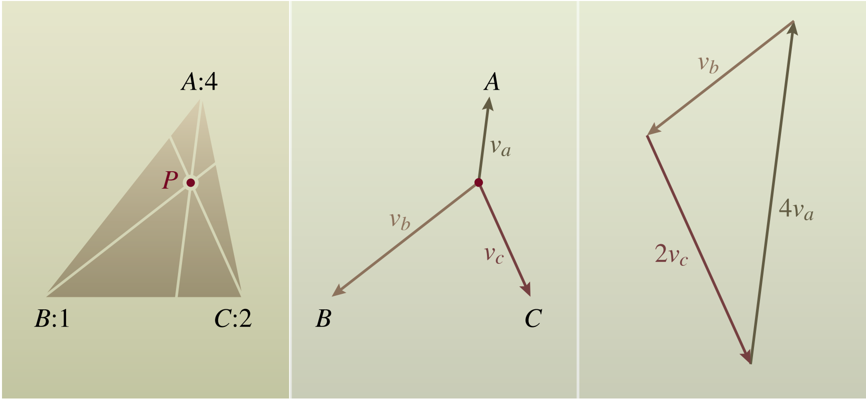
The vector approach

There is a vector approach to this problem as well. Start again with the two person seesaw, with masses m_A and m_B at points A and B, respectively. The balance point occurs at the center of mass M, when the two vectors $m_a \cdot \overrightarrow{MA}$ and $m_b \cdot \overrightarrow{MB}$ cancel out:

$$m_a \cdot \overrightarrow{MA} + m_b \cdot \overrightarrow{MB} = 0.$$

More generally, we can consider when a sum of terms of the form $m_i \cdot \overrightarrow{MP_i}$ add up to zero. The quantities $m_i \cdot \overrightarrow{MP_i}$ are measures of the the tendency of the system to turn in the direction of P_i. The balancing point, the center of mass M, is where all those cancel out:

$$\sum_i m_i \cdot \overrightarrow{MP_i} = 0.$$



Vectors from P to A,B,C. $4v_a + v_b + 2v_c = 0$

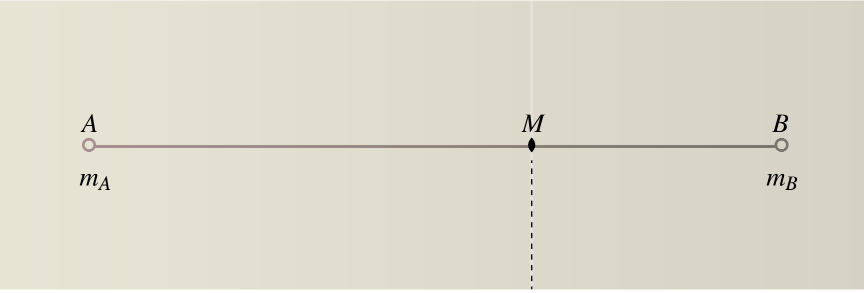
Of course, the idea of mass exists outside of the geometry that we have developed. For our purposes, it is not really necessary to think of the coefficients m_i as masses at all– if you want to avoid physics entirely, you can just think of these as arbitrary scalar coefficients in a vector equation. Whether you think of them as masses or not, it is these coefficients that form the basis for barycentric coordinates. Let's start by investigating some properties of these centers of mass, beginning with a two mass system. Of course, the center of mass of two objects will lie between them, as long as those two masses both have positive mass.

If you are willing to allow for negative mass, then that center of mass may not be between them, but everything else still “just works”. You do then have to talk about in terms of signed distance and signed area. I don’t feel like dealing with that, so I will just work with positive masses and positive distances.

PROPOSITION I

If M is the center of mass of a two mass system, with mass m_A at point A and mass m_B at point B, then

$$|MA| = \frac{m_B}{m_A+m_B}\cdot|AB| \quad \& \quad |MB| = \frac{m_A}{m_A+m_B}\cdot|AB|.$$



Proof. Since M is the center of mass, by definition

$$m_A\cdot\overrightarrow{MA}+m_B\cdot\overrightarrow{MB}=0.$$

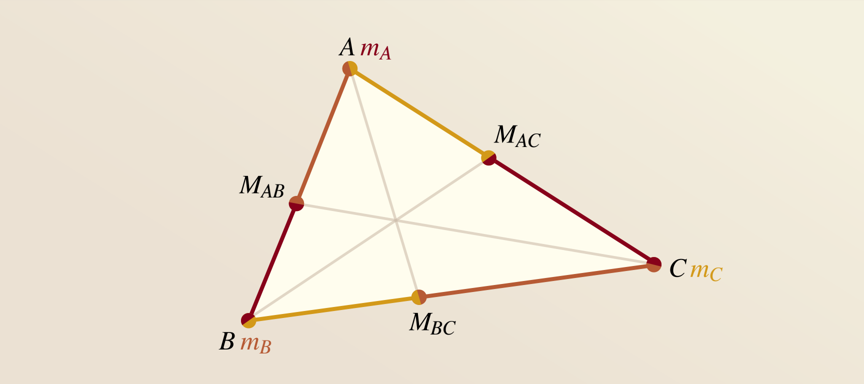
In order for $m_A\cdot\overrightarrow{MA}$ and $m_B\cdot\overrightarrow{MB}$ to cancel, they have to be the same length, so $m_A|MA| = m_B|MB|$, so $|MA|/|MB| = m_B/m_A$. Now let’s look at the ratio of $|MA|$ to $|AB|$:

$$\frac{|MA|}{|AB|} = \frac{|MA|}{|MA|+|MB|} = \frac{1}{1+(|MA|/|MB|)}$$
$$= \frac{1}{1+(m_A/m_B)} = \frac{m_B}{m_A+m_B}.$$

Therefore

$$|MA| = \frac{m_B}{m_A+m_B}\cdot|AB|.$$

The calculation of $|MB|$ is, of course, similar. □



PROPOSITION II

In a triangle $\triangle ABC$ with masses m_A at A, m_B at B, and m_C at C, label: M_{AB}, the center of mass of A and B; M_{AC}, the center of mass of A and C; and M_{BC}, the center of mass of B and C. Then the segments AM_{BC}, BM_{AC}, and CM_{AB} are concurrent.

Proof. This is a straightforward application of Ceva's Theorem, using the measurements from the previous calculation. Recall that Ceva's Theorem guarantees a point of concurrence if a product of ratios around the edges of the triangle equals out to one. In this case, that product is

$$\frac{|AM_{AB}|}{|M_{AB}B|} \cdot \frac{|BM_{BC}|}{|M_{BC}C|} \cdot \frac{|CM_{AC}|}{|M_{AC}A|}.$$

If we focus on just the first ratio in that product and use the previous proposition,

$$\frac{|AM_{AB}|}{|M_{AB}B|} = \frac{m_B/(m_A+m_B)\cdot|AB|}{m_A/(m_A+m_B)\cdot|AB|} = \frac{m_B}{m_A}.$$

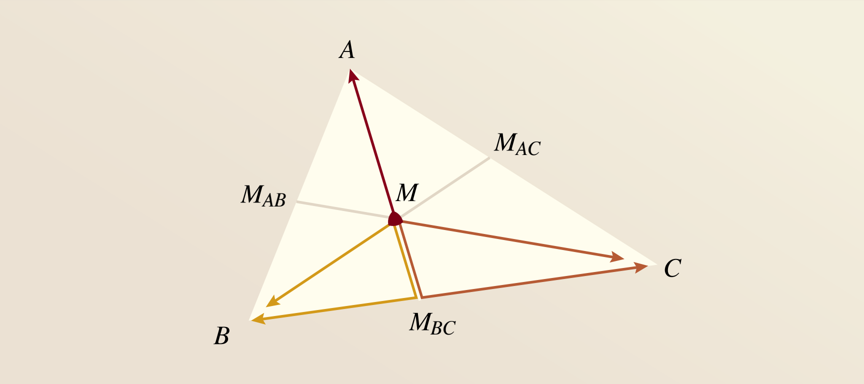
Likewise,

$$\frac{|BM_{BC}|}{|M_{BC}C|} = \frac{m_C}{m_B} \quad \& \quad \frac{|CM_{AC}|}{|M_{AC}A|} = \frac{m_A}{m_C},$$

and so

$$\frac{|AM_{AB}|}{|M_{AB}B|} \cdot \frac{|BM_{BC}|}{|M_{BC}C|} \cdot \frac{|CM_{AC}|}{|M_{AC}A|} = \frac{m_B}{m_A} \cdot \frac{m_C}{m_B} \cdot \frac{m_A}{m_C} = 1.$$

By Ceva's Theorem, the three segments are concurrent. □



PROPOSITION III

The center of mass M of masses m_A at A, m_B at B, and m_C at C, is the point of concurrence of the segments AM_{BC}, BM_{AC}, and CM_{AB}.

Proof. Let's show that M is on the segment AM_{BC}. A similar argument will work to show it is on the other two segments, and therefore that it is at their mutual intersection. Since M is the center of mass of the three mass system, we may write

$$m_A\overrightarrow{MA}+m_B\overrightarrow{MB}+m_C\overrightarrow{MC}=0.$$

Now a little vector arithmetic gets us

$$m_A\overrightarrow{MA}+m_B(\overrightarrow{MM_{BC}}+\overrightarrow{M_{BC}B})+m_C(\overrightarrow{MM_{BC}}+\overrightarrow{M_{BC}C})=0,$$
$$m_A\overrightarrow{MA}+(m_B+m_C)\overrightarrow{MM_{BC}}+(m_B\overrightarrow{M_{BC}B}+m_C\overrightarrow{M_{BC}C})=0.$$

The last piece of that is zero since M_{BC} is the center of mass of the system with masses m_B at B and m_C at C. Therefore

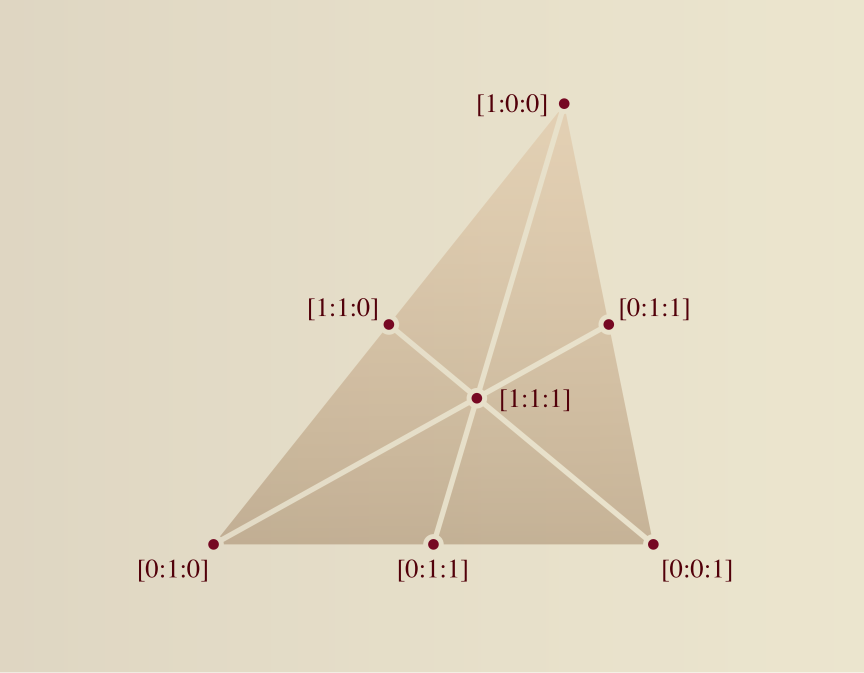
$$m_A\overrightarrow{MA}+(m_B+m_C)\overrightarrow{MM_{BC}}=0.$$

In order for these two vectors to cancel out like this, they must be oppositely directed. That is, A, M, and M_{BC} must be collinear. □

DEF: BARYCENTRIC COORDINATES

Given a triangle $\triangle ABC$ and a point M. A set of barycentric coordinates of M (relative to $\triangle ABC$) is a triple $[m_a : m_b : m_C]$ (with not all of m_A, m_B, and m_C equal to zero) so that

$$m_a\overrightarrow{MA} + m_b\overrightarrow{MB} + m_c\overrightarrow{MC} = 0.$$



The most immediate observation is that barycentric coordinates are defined only up to a constant multiple: if

$$m_a\overrightarrow{MA} + m_b\overrightarrow{MB} + m_c\overrightarrow{MC} = 0$$

then

$$k \cdot m_a\overrightarrow{MA} + k \cdot m_b\overrightarrow{MB} + k \cdot m_c\overrightarrow{MC} = 0$$

as well. Therefore, the barycentric coordinates of a point are not really a triple $[m_a : m_b : m_c]$, but instead an equivalence class of triples where $[m_a : m_b : m_c] = [n_a : n_b : n_c]$ if there is a nonzero constant k so that $m_a = kn_a$, $m_b = kn_b$, and $m_c = kn_c$.

The connection to area and trilinears

Barycentric coordinates can be calculated, quite directly, using either areas of triangles or trilinear coordinates. The key to it is the following theorem that relates the masses m_A, m_B and m_C to the areas of certain triangles. Throughout the rest of this lesson I will use the notation $(\triangle ABC)$ to denote the *area* of $\triangle ABC$ (it appears to be somewhat common to use the absolute value signs to denote area, but I used that notation for perimeter a long time ago).

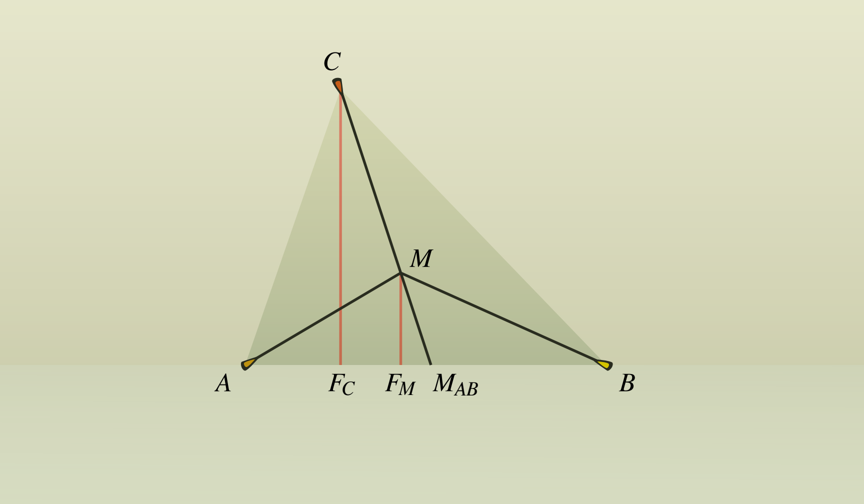
THM: MASS AND AREA

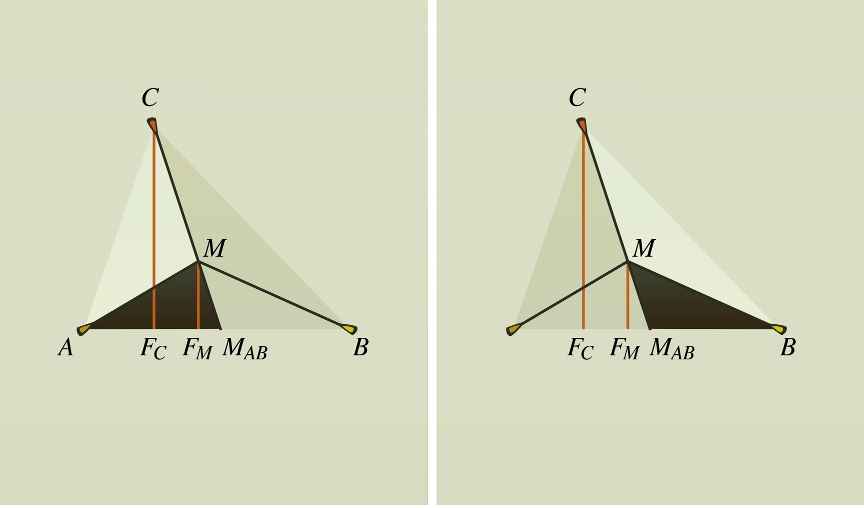
Given a triangle $\triangle ABC$, with masses m_A at A, m_B at B, and m_C at C, and a center of mass M. Then

$$\frac{m_A}{m_B} = \frac{(\triangle BCM)}{(\triangle ACM)}, \quad \frac{m_B}{m_C} = \frac{(\triangle ACM)}{(\triangle ABM)}, \quad \frac{m_C}{m_A} = \frac{(\triangle ABM)}{(\triangle BCM)}.$$

Proof. Let's look at the first of these (the other two are just a shuffling of labels). Label

F_C: the foot of the altitude from A
F_M: the foot of the altitude from M
M_{AB}: the center of mass of AB.





Then

$$\begin{aligned}(\triangle CAM) &= (\triangle CAM_{AB}) - (\triangle MAM_{AB})\\ &= |CF_C|\cdot|AM_{AB}| - |MF_M|\cdot|AM_{AB}|\\ &= |AM_{AB}|(|CF_C| - |MF_M|).\end{aligned}$$

and

$$\begin{aligned}(\triangle CBM) &= (\triangle CBM_{AB}) - (\triangle MBM_{AB})\\ &= |CF_C|\cdot|BM_{AB}| - |MF_M|\cdot|BM_{AB}|\\ &= |BM_{AB}|(|CF_C| - |MF_M|)\end{aligned}$$

so

$$\frac{(\triangle CAM)}{(\triangle CBM)} = \frac{|AM_{AB}|(|CF_C| - |MF_M|)}{|BM_{AB}|(|CF_C| - |MF_M|)} = \frac{|AM_{AB}|}{|BM_{AB}|} = \frac{m_A}{m_B}.$$

Likewise, with the proper interchange of letters,

$$\frac{m_B}{m_C} = \frac{(\triangle ACM)}{(\triangle ABM)} \quad \& \quad \frac{m_C}{m_A} = \frac{(\triangle ABM)}{(\triangle BCM)}.$$

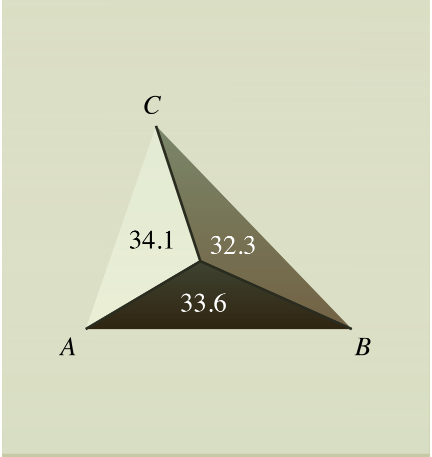
□

As an immediate consequence, we get a way to use triangle areas to calculate barycentric coordinates.

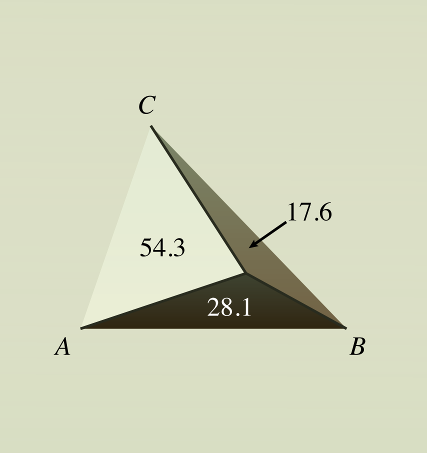
COR: BARYCENTRIC COORDINATES AND AREA

Any point M subdivides a triangles $\triangle ABC$ into three pieces, $\triangle ABM$, $\triangle ACM$, and $\triangle BCM$. The barycentric coordinates of M can be computed from the the areas of those triangles as

$$[(\triangle BCM) : (\triangle ACM) : (\triangle ABM)].$$



[32.3 : 34.1 : 33.6]



[17.6 : 54.3 : 28.1]

Proof. Let $[m_a : m_b : m_c]$ be the barycentric coordinates of M. At least one of the three coordinates must be nonzero. Let's assume it is m_C. Then it is a three-step calculation: (1) divide through by m_C, (2) use the previous theorem to make the connection to area, and (3) multiply through by $(\triangle BCM)$.

$$\begin{aligned}[m_a : m_b : m_c] &= [m_a/m_c : m_b/m_c : 1] \\ &= [(\triangle BCM)/(\triangle ABM) : (\triangle ACM)/(\triangle ABM) : 1] \\ &= [(\triangle BCM) : (\triangle ACM) : (\triangle ABM)]\end{aligned}$$

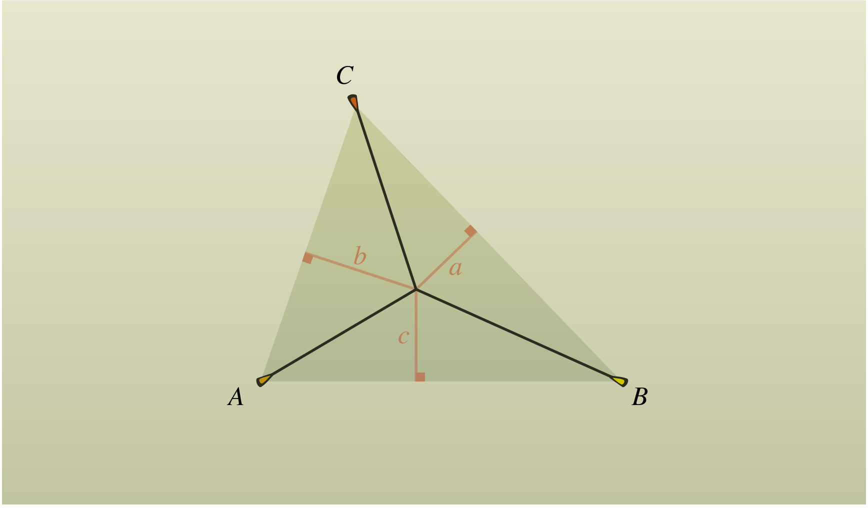
□

So just how closely related are barycentric and trilinear coordinates?

THM: BARYCENTRIC COORDINATES AND TRILINEARS

If the trilinear coordinates of a point M relative to $\triangle ABC$ are $[a : b : c]$, then the barycentric coordinates of M (relative to that same triangle) are

$$[a \cdot |BC| : b \cdot |AC| : c \cdot |AB|].$$



Proof. The barycentric coordinates of M can be computed from the areas of triangles as

$$[(\triangle BCM) : (\triangle ACM) : (\triangle ABM)] = [h_a|BC| : h_b|AC| : h_C|AB|].$$

where h_a, h_b, and h_c are the lengths of the altitudes from M in each of the three triangles. But the trilinear coordinates of M can be normalized to measure exactly these lengths. Therefore $a = h_a$, $b = h_b$, and $c = h_c$. □

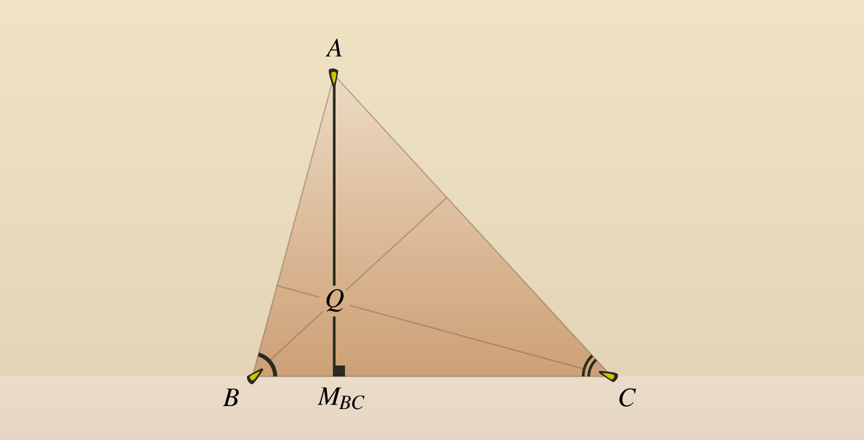
Barycentric coordinates of important triangle centers

Based upon the conversation at the start of the lesson, the barycentric coordinates of the centroid are $[1 : 1 : 1]$. What about some of the other triangle centers we have encountered? Of course, we already have a very easy way to convert from trilinear coordinates to barycentric coordinates, but what would be the fun in that? So let's start with the orthocenter.

THM: BARYCENTRIC COORDINATES OF THE ORTHOCENTER
In $\triangle ABC$, the barycentric coordinates of the orthocenter are

$$[\cot A : \cot B : \cot C].$$

Proof. Let M_{BC} be the foot of the altitude which passes through A and is perpendicular to BC. Look at the two right triangles $\triangle ABM_{BC}$ and $\triangle ACM_{BC}$.



In them,

$$|BM_{BC}| = |AM_{BC}|\cot(\angle B) \quad \& \quad |CM_{BC}| = |AM_{BC}|\cot(\angle C).$$

Therefore

$$\frac{|BM_{BC}|}{|CM_{BC}|} = \frac{\cot B}{\cot C}.$$

Likewise, if M_{AC} is the foot of the altitude which passes through B and is perpendicular to AC, then

$$\frac{|AM_{AC}|}{|CM_{AC}|} = \frac{\cot A}{\cot C}.$$

Now the masses m_A, m_B, and m_C must be in those same ratios. That is,

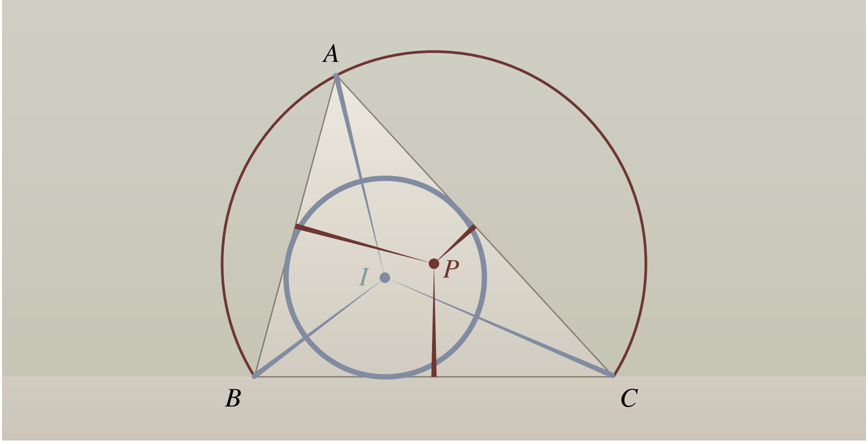
$$\frac{\cot A}{\cot C} = \frac{m_A}{m_C} \quad \& \quad \frac{\cot B}{\cot C} = \frac{m_B}{m_C}.$$

That means that the barycentric coordinates of the orthocenter are

$$\left[\frac{\cot A}{\cot C} : \frac{\cot B}{\cot C} : 1\right] \sim [\cot A : \cot B : \cot C].$$

□

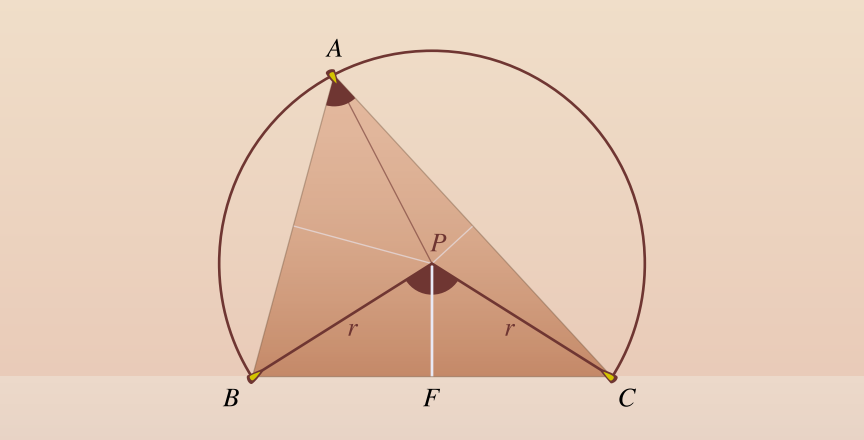
The key to finding the barycentric coordinates of the circumcenter and incenter is the fact that they are the centers of circles– the circumcircle and the incircle.



THM: BARYCENTRIC COORDINATES OF THE CIRCUMCENTER

In $\triangle ABC$, the barycentric coordinates of the circumcenter are

$$[|BC|\cos(\angle A) : |AC|\cos(\angle B) : |AB|\cos(\angle C)].$$



Proof. Let P denote the circumcenter and remember that it is the center of the circumcircle, a circle that passes through each of A, B, and C, so that $|PA| = |PB| = |PC|$. Let r be the radius of this circumcircle. The argument in this proof is essentially a rip-off of the argument in the calculation of the circumcenter's trilinear coordinates, so you may want to review that now. If F is the foot of the perpendicular through P to the line BC, then note that $\angle BPF = \frac{1}{2}\angle BPC = \angle A$ (by the Inscribed Angle Theorem). Therefore

$$|PF| = r\cos(\angle A)$$

and so

$$(\triangle PBC) = \tfrac{1}{2} r\cos(\angle A)|BC|.$$

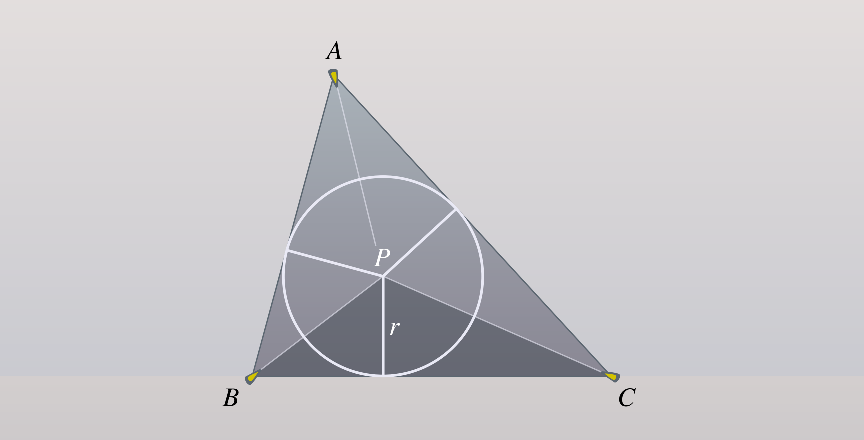
Similarly

$$(\triangle PAC) = \tfrac{1}{2} r\cos(\angle B)|AC| \quad \& \quad (\triangle PAB) = \tfrac{1}{2} r\cos(\angle C)|AB|,$$

and we have seen that the areas of these triangles determine the barycentric coordinates of P:

$$\begin{gathered}\left[\tfrac{1}{2}r|BC|\cos(\angle A) : \tfrac{1}{2}r|AC|\cos(\angle B) : \tfrac{1}{2}r|AB|\cos(\angle C)\right] \\ \sim [|BC|\cos(\angle A) : |AC|\cos(\angle B) : |AB|\cos(\angle C)].\end{gathered}$$

□



THM:BARYCENTRIC COORDINATES OF THE INCENTER

In $\triangle ABC$, the barycentric coordinates of the incenter are

$$[|BC| : |AC| : |AB|].$$

Proof. Let P be the incenter of $\triangle ABC$. Recall that the incenter is equidistant from each of the sides of the triangle– it is the center of the inscribed circle of $\triangle ABC$. Let r be the radius of this incircle. Then

$$(\triangle PBC) = \tfrac{1}{2}r|BC|, \quad (\triangle PAC) = \tfrac{1}{2}r|AC|, \quad (\triangle PAB) = \tfrac{1}{2}r|AB|,$$

so the barycentric coordinates of P are

$$\left[\tfrac{1}{2}r|BC| : \tfrac{1}{2}r|AC| : \tfrac{1}{2}r|AB|\right] \sim [|BC| : |AC| : |AB|].$$

□

References

I referenced Clark Kimberling's web site [3] in an earlier lesson, but it also includes barycentric coordinates for many, many triangle centers. Once again, Coxeter's *Introduction to Geometry*[2] provides a good perspective on this topic. There is also a "letter" from John Conway to Steve Sigur[1] floating around the web that extolls the virtues of barycentric coordinates.

[1] John Conway. Trilinear vs barycentric coordinates. Correspondence, distributed on World Wide Web. Currently available at http://mathforum.org/kb/message.jspa?messageID=1091956.

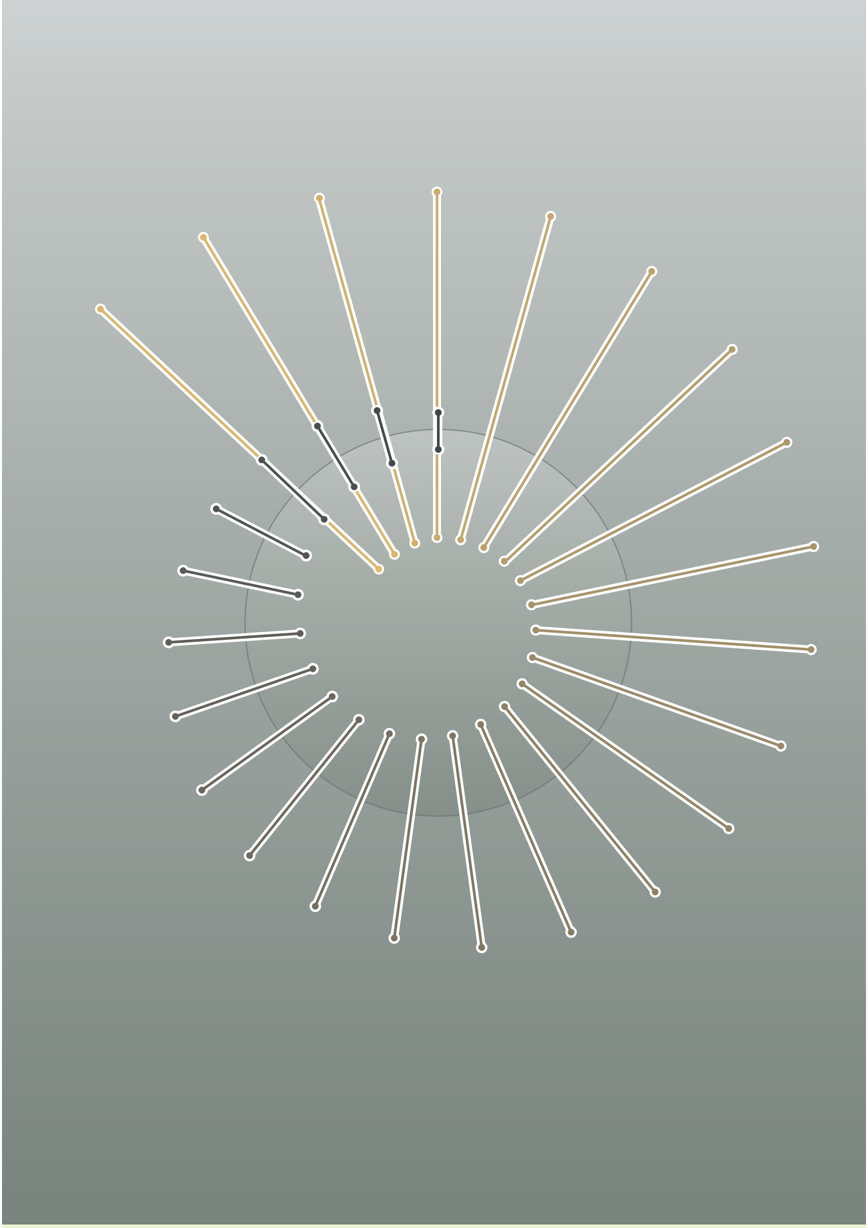
[2] H.S.M. Coxeter. *Projective Geometry*. Blaisdell Publishing Co., New York, 1st edition, 1964.

[3] Clark Kimberling. Encyclopedia of triangle centers - etc. distributed on World Wide Web. http://faculty.evansville.edu/ck6/encyclopedia /ETC.html.

Exercises

1. Consider the triangle $\triangle ABC$ whose vertices are at the coordinates $A = (0,0)$, $B = (2,0)$ and $C = (0,4)$. Find barycentric coordinates for the point $(1,1)$.

2. Show that the barycentric coordinates of the excenters of $\triangle ABC$ are $[-|BC| : |AC| : |AB|]$, $[|BC| : -|AC| : |AB|]$, and $[|BC| : |AC| : -|AB|]$.

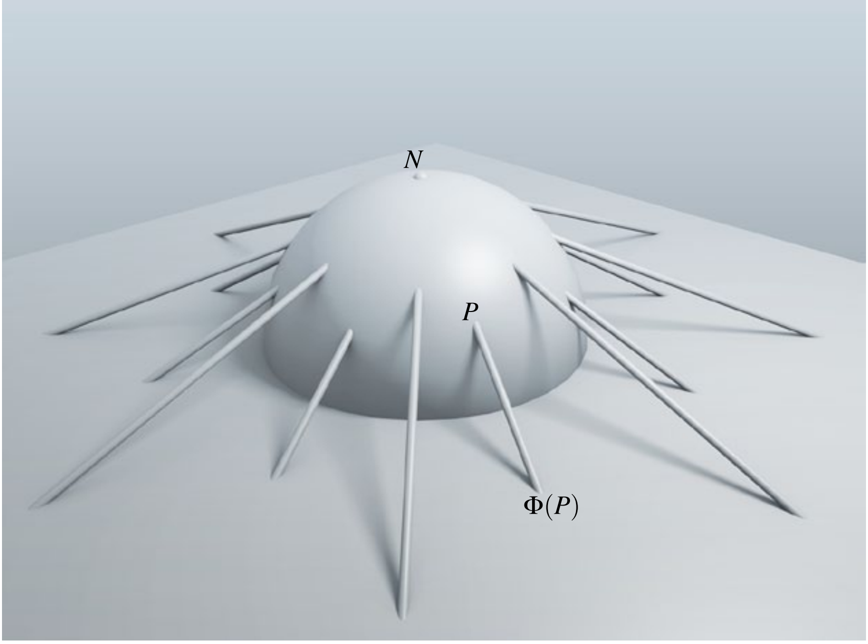


35 INVERSION

In the last few lessons we classified all of the bijective mappings of the Euclidean plane that respect incidence, order, and congruence. Now we are going to have to look for mappings that fall short of that stringent list of conditions, but that still preserve enough remnants of the Euclidean structure to tell us something interesting. An optimist would view the additional freedom as an opportunity, and indeed I think that this is a time to be optimistic. The particular type of mapping that we will investigate in the next couple of lessons is called inversion. Inversions provide interesting insight into some of the classical problems of Euclidean geometry, particularly those that involve circles. Inversions also play an important role in the study of non-Euclidean geometry. I think that the most natural path into the topic of inversion is via stereographic projection. This means that we will have to momentarily step outside of the plane. Don't worry– by the time we get around to formally defining inversions, we will be comfortably back in the plane.

Stereographic Projection

Ever since map-makers realized that the earth is round, they have sought ways to project a spherical surface down to a flat plane. One approach which is nice mathematically (although maybe not so nice cartographically) is called stereographic projection. It works as follows. First put the center of the sphere (say of radius r) at the origin of the plane. Then draw rays out from the north pole through each other point of the sphere. Those rays will each intersect the plane, establishing a bijection between the points of the sphere (except the north pole itself) and the points of the plane. That mapping from the sphere to the plane is called *stereographic projection*.



With a few symbols, we can describe the process more precisely. Label

$\mathbb{E}$: the plane $z = 0$;

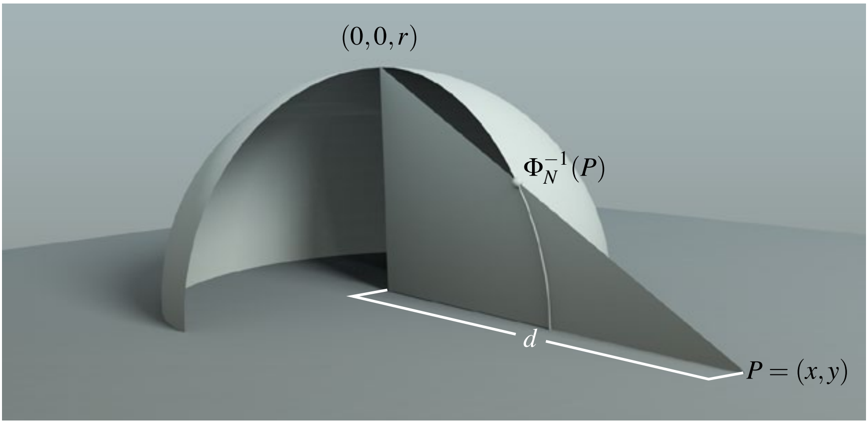
$\mathbb{S}$: the sphere of radius r, centered at the origin;

N: the "north pole"– the point with coordinates $(0,0,r)$;

Φ: the stereographic projection from $\mathbb{S}$ to $\mathbb{E}$;

P: any point of $\mathbb{S}$ other than N.

Then $NP\rightarrow$ will intersect $\mathbb{E}$ exactly once, and $\Phi(P)$ is defined to be this intersection point. Since Φ is a bijection, it has an inverse, Φ^{-1}, that is called *inverse stereographic projection*. For those of you that worry about a possible northern hemisphere bias, we can do the same kind of projection equally well from the south pole. In fact, to define inversion, we will need to work from both poles– first an inverse stereographic from the north pole, and then a stereographic projection from the south pole. It is pretty straightforward to work out analytic equations to describe these mappings, and that is the first task of this lesson.



THM: EQUATIONS FOR STEREOGRAPHIC PROJECTION

The inverse stereographic projection Φ_N^{-1} from the north pole $(0,0,r)$ is given by the equation

$$\Phi_N^{-1}(x,y)=\left(\frac{2xr^2}{d^2+r^2},\frac{2yr^2}{d^2+r^2},\frac{rd^2-r^3}{d^2+r^2}\right)$$

where $d=\sqrt{x^2+y^2}$ is the distance from O to the point (x,y). The stereographic projection Φ_S from the south pole $(O,O,-r)$ is given by the equation

$$\Phi_S(x,y,z)=\left(\frac{rx}{r+z},\frac{ry}{r+z}\right).$$

Proof. I will prove the first of these formulas, and leave the second to you. The point (x,y) in the plane corresponds to the point $(x,y,0)$ in 3-dimensional space. Start with a parametrized equation for the line through $(x,y,0)$ and the north pole, $(0,0,r)$:

$$s(t)=\langle 0,0,r\rangle+t\langle x-0,y-0,0-r\rangle=\langle tx,ty,r-rt\rangle.$$

We need to find out when this line hits the sphere. All the points on the sphere are a distance r from the origin, so this basically boils down to the equation $|s(t)|^2=r^2$:

$$(tx)^2+(ty)^2+(r-rt))^2=r^2$$
$$t^2x^2+t^2y^2+r^2-2r^2t+r^2t^2=r^2.$$

Cancel out the r^2 on both sides, and factor to solve for t:

$$\begin{gathered} t^2(x^2+y^2) - 2r^2t + r^2t^2 = 0 \\ t^2d^2 + r^2t^2 - 2r^2t = 0 \\ t((d^2+r^2)t - 2r^2) = 0. \end{gathered}$$

The first solution, when $t = 0$, is at the north pole – that's not the one we want. The other intersection occurs when

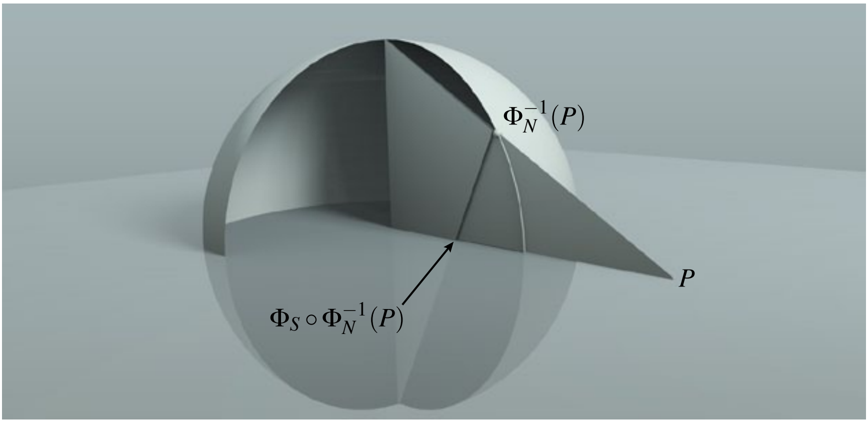
$$t = \frac{2r^2}{d^2+r^2}.$$

Plugging that into $s(t)$ gives the vector that points to $\Phi_N^{-1}(x,y)$:

$$\left\langle \frac{2xr^2}{d^2+r^2}, \frac{2yr^2}{d^2+r^2}, r - \frac{2r^3}{d^2+r^2} \right\rangle.$$

You can use a similar argument for the second part– find the equation of the line through the south pole and the point (x,y,z), and then locate its intersection with the plane $z = 0$. □

This is a book on *plane* geometry, so we should really be looking for maps from the plane to itself. We can get such a map by composing Φ_N^{-1} and Φ_S– the first step in the composition takes the plane to the sphere, but the second step brings it back. Notice that when we do this, there is clearly a problem at the origin O, since $\Phi_N^{-1}(O) = S$, and $\Phi_S(S)$ is undefined. If we just toss out that one bad point, though, what's left is a perfectly good bijection from $\mathbb{E} - O$ to itself.



Let's call that bijection σ. Then

$$\sigma(x,y) = \Phi_S \circ \Phi_N^{-1}(x,y) = \Phi_S\left(\frac{2xr^2}{d^2+r^2}, \frac{2yr^2}{d^2+r^2}, r - \frac{2r^3}{d^2+r^2}\right).$$

The x and y coordinates of this are similar– we can just focus on the first:

$$\frac{r\left(\dfrac{2xr^2}{d^2+r^2}\right)}{r+\left(r-\dfrac{2r^3}{d^2+r^2}\right)}.$$

Multiply through, top and bottom, by d^2+r^2 to get

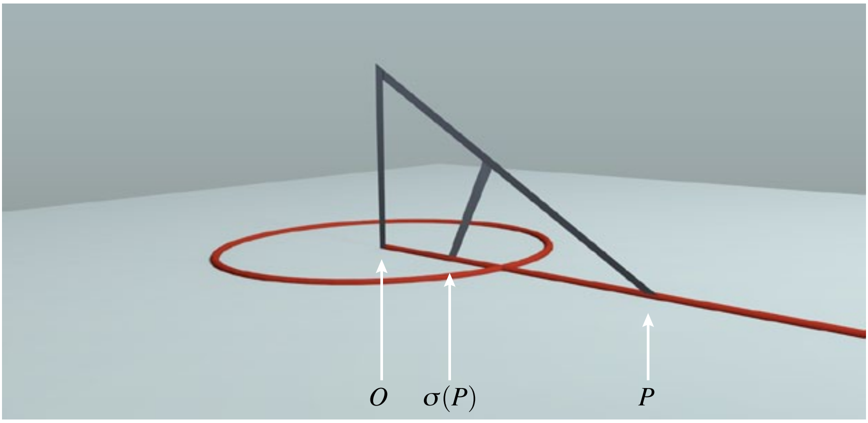
$$\frac{2xr^3}{2r(d^2+r^2)-2r^3} = \frac{2xr^3}{2rd^2+2r^3-2r^3} = \frac{2xr^3}{2rd^2} = \frac{xr^2}{d^2}.$$

The second coordinate works similarly and eventually simplifies down to yr^2/d^2, so

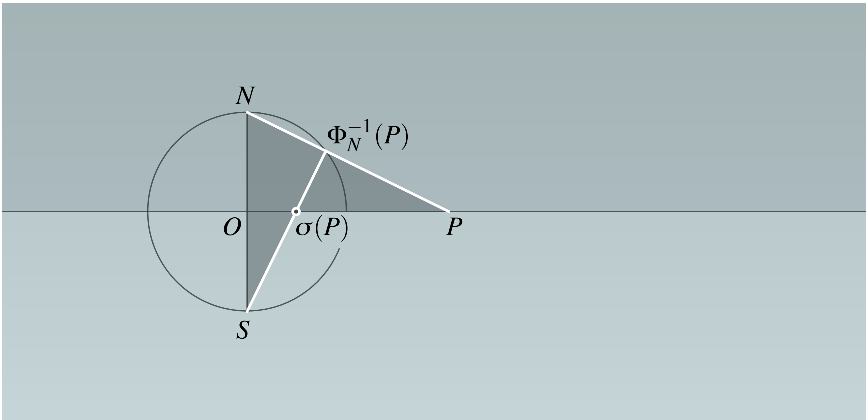
$$\sigma(x,y) = \left(x\cdot\frac{r^2}{d^2}, y\cdot\frac{r^2}{d^2}\right).$$

Note then that $\sigma(x,y)$ is on the same ray from the origin as (x,y), but its distance from the origin has been altered– the distance from the origin is now

$$\sqrt{\left(\frac{xr^2}{d^2}\right)^2+\left(\frac{yr^2}{d^2}\right)^2} = \sqrt{\frac{x^2r^4+y^2r^4}{d^4}} = \sqrt{\frac{d^2r^4}{d^4}} = \frac{r^2}{d}.$$



There is a more geometric view of this that may be more appealing than the previous calculations. Take a cross section of the sphere and plane as illustrated:



By Thales' Theorem, the two lines $\leftarrow NP \rightarrow$ and $\leftarrow S\sigma(P) \rightarrow$ intersect at right angles at $\Phi_N^{-1}(P)$. Then by A·A similarity,

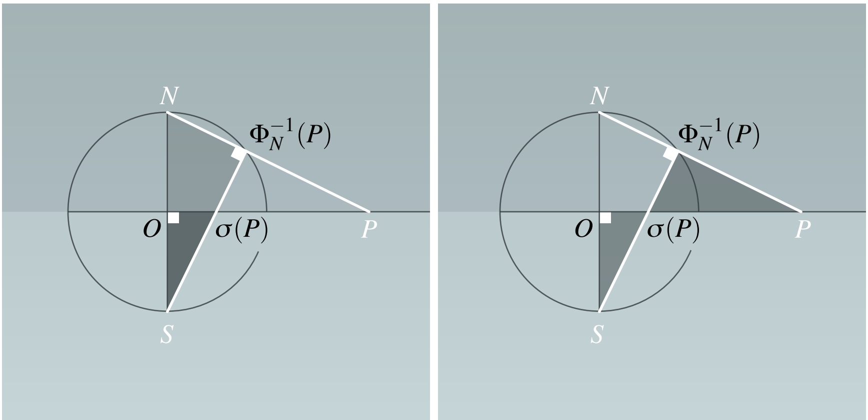
$$\triangle SN\Phi_N^{-1}(P) \sim \triangle S\sigma(P)O$$

(since they both have a right angle and they share the angle at S).

Also by A·A similarity,

$$\triangle S\sigma(P)O \sim \triangle P\sigma(P)\Phi_N^{-1}(P)$$

(using the right angles and the vertical angle pair at $\sigma(P)$).



Therefore

$$\triangle SN\Phi_N^{-1}(P) \sim \triangle P\sigma(P)\Phi_N^{-1}(P).$$

Matching the corresponding ratios of the two legs of these triangles,

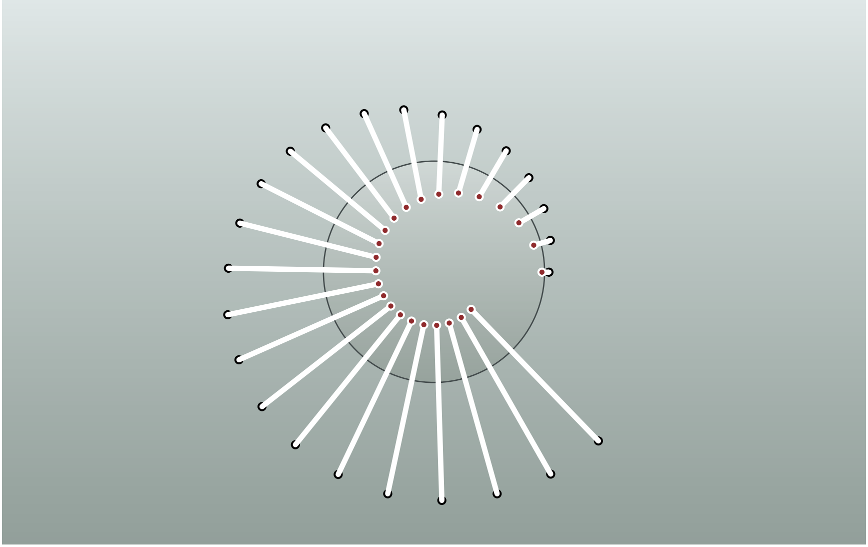
$$\frac{|O\sigma(P)|}{r} = \frac{r}{|OP|} \implies |O\sigma(P)| = \frac{r^2}{|OP|} = r^2/d.$$

Inversion

This map σ that we constructed in the previous section is, in fact, an inversion. Using the above properties, we can now give a proper definition of inversion that does not stray from the plane. The sphere of radius r is replaced by its intersection with the plane, a circle of radius r. Furthermore, there is no longer any real advantage to centering the circle at the origin.

DEF: INVERSION

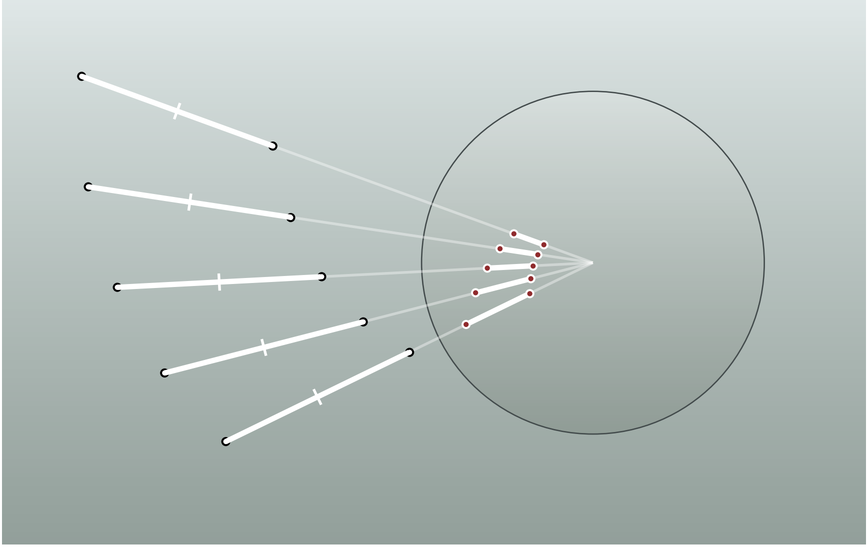
Let $\mathcal{C}$ be a circle with center O and radius r. The inversion σ across $\mathcal{C}$ is the bijection of the points of $\mathbb{E} - O$ defined as follows. For any point $P \in \mathbb{E} - O$, $\sigma(P)$ is the point on the ray $OP\rightarrow$ that is a distance $r^2/|OP|$ from O.



Inversion of a collection of points across a circle.

Note that an inversion turns a circle inside out–

1. If P is inside $\mathcal{C}$, then $|OP|$ is less than r, so $r^2/|OP|$ is greater than r, so $\sigma(P)$ is outside $\mathcal{C}$.
2. If P is outside $\mathcal{C}$, then $|OP|$ is greater than r, so $r^2/|OP|$ is less than r, so $\sigma(P)$ is inside $\mathcal{C}$.
3. If P is on $\mathcal{C}$, then $|OP| = r$, so $r^2/|OP| = r$, so $\sigma(P)$ is again on $\mathcal{C}$. In fact, since $OP\rightarrow$ only intersects $\mathcal{C}$ once, in this case $P = \sigma(P)$.



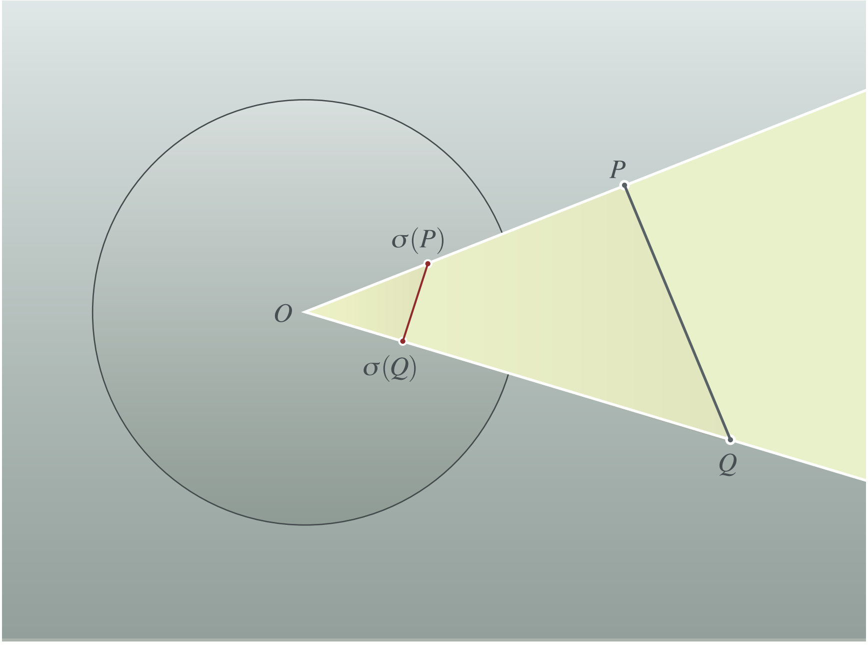
Distances are not all scaled by the same amount.

That last observation is an important one– σ fixes all the points of $\mathcal{C}$. In this regard, an inversion is a little like a reflection. Whereas a reflection fixes a line and swaps the two sides of it, an inversion fixes a circle and swaps the interior and exterior of it. Furthermore, it is easy to see that, like a reflection, an inversion is its own inverse. But it is also important to note how an inversion differs from a reflection; perhaps most importantly, an inversion does *not* scale all distances by a constant– points that are very close to O may be thrown very apart from one another, while points that are very far from O will all be squeezed into a tiny space right around O.

All is not lost, however. The first sign of hope is a result on similarity.

THM: A SIMILARITY CREATED BY INVERSION
Let σ be the inversion across a circle $\mathcal{C}$ with radius r and center O. Then for any two distinct points P and Q in $\mathbb{E} - O$,

$$\triangle POQ \sim \triangle \sigma(Q)O\sigma(P).$$



Proof. First of all, the two triangles in question share an angle at O. Now take a look at the sides:

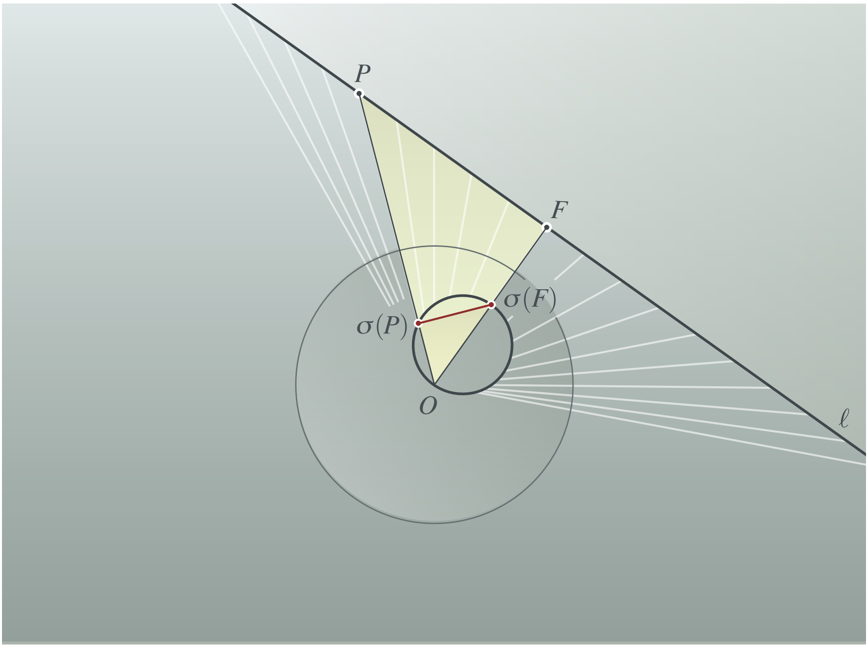
$$|O\sigma(P)| = r^2/|OP| \quad \& \quad |O\sigma(Q)| = r^2/|OQ|,$$

so

$$\frac{|O\sigma(P)|}{|O\sigma(Q)|} = \frac{r^2/|OP|}{r^2/|OQ|} = \frac{|OQ|}{|OP|}.$$

By the S·A·S similarity theorem, then, the two triangles are similar. Note carefully, though, that the sides OP and OQ are "crossed up" by this similarity. □

Let's look at some larger structures. We have seen that all Euclidean transformations map lines to lines, but what happens when we *invert* a line? One situation is easy– any line that passes through O is mapped to itself. [Technically, it isn't quite mapped to itself, because there is a problem at O. Forgive me, but for the rest of the section, it is just more convenient to ignore the problems that arise at O.] For a line that does not pass through O, the situation gets more interesting.

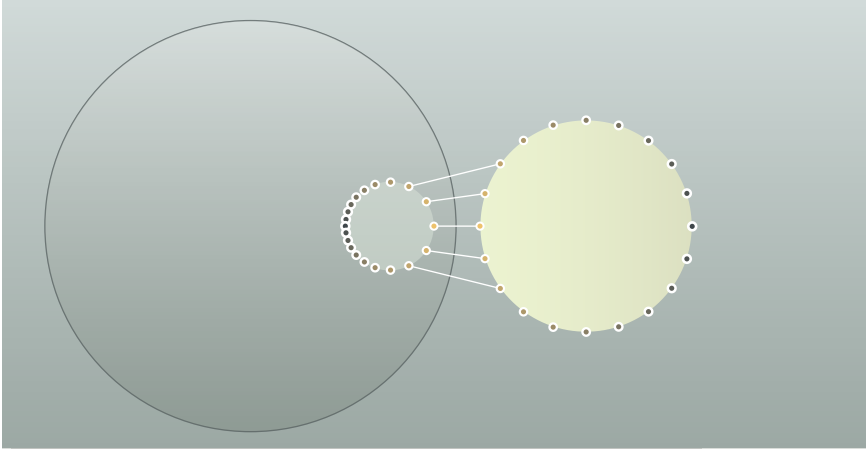


THM: INVERTING A LINE

Let σ be the inversion across a circle $\mathcal{C}$ with radius r and center O. Let ℓ be a line that does not pass through O. Then $\sigma(\ell)$ is a circle that passes through O.

Proof. Let F be the foot of the perpendicular from O to ℓ. I claim that $O\sigma(F)$ is the diameter of the circle $\sigma(\ell)$. To see why, take any other point P on ℓ. Then $\triangle OFP$ is a right triangle with right angle at F. As we have just seen, $\triangle OFP$ is similar to $\triangle O\sigma(P)\sigma(F)$ which means that $\triangle O\sigma(P)\sigma(F)$ is a right triangle whose right angle is at $\sigma(P)$. By Thales' Theorem, $\sigma(P)$ must be on the circle with diameter $O\sigma(P)$. □

It is easy to play that argument in reverse: any circle which passes through O inverts to a line (which does not pass through O). But that obviously leads to another question– what about circles that don't pass through O?



THM: INVERTING A CIRCLE

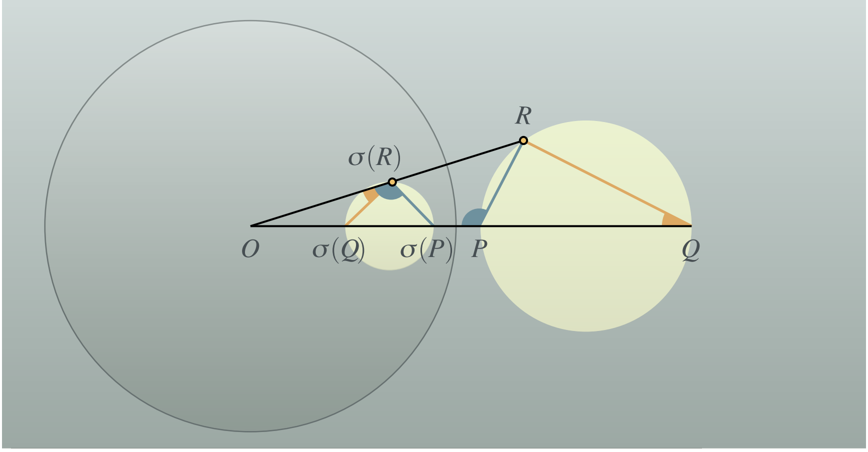
Let Let σ be the inversion across a circle $\mathcal{C}$ with radius r and center O. Let c be a circle that does not pass through O. Then $\sigma(c)$ is again a circle (that does not pass through O).

Proof. This proof again uses Thales' Theorem…it is just a little more complicated. The ray from O through the center of c will intersect c twice. Label those two points P and Q. Then PQ is a diameter of c and I claim that $\sigma(P)\sigma(Q)$ is a diameter of $\sigma(c)$. Now let R be another point on c. Then

$$\triangle OPR \sim \triangle O\sigma(R)\sigma(P) \implies \angle OPR \simeq \angle O\sigma(R)\sigma(P)$$
$$\triangle OQR \sim \triangle O\sigma(R)\sigma(Q) \implies \angle OQR \simeq \angle O\sigma(R)\sigma(Q).$$

A little angle arithmetic:

$$\begin{aligned}(\angle\sigma(P)\sigma(R)\sigma(Q)) &= (\angle O\sigma(R)\sigma(P)) - (\angle O\sigma(R)\sigma(Q))\\ &= (\angle OPR) - (\angle OQR).\end{aligned}$$

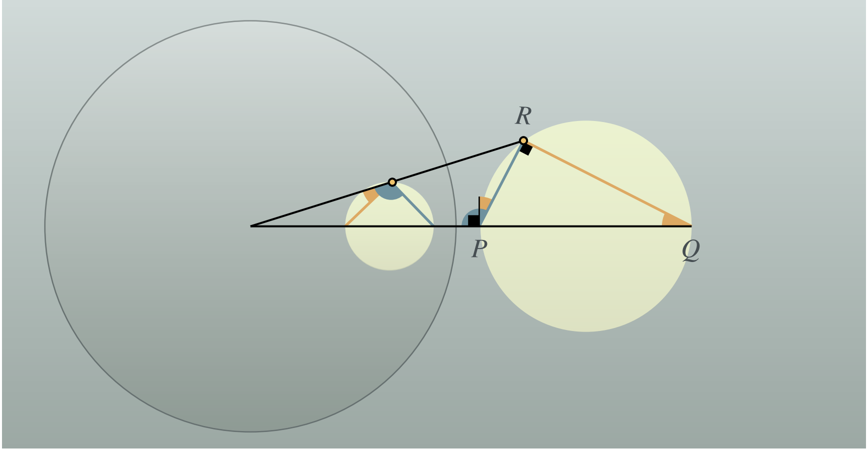


Note though that $\angle OPR$ is an exterior angle of $\triangle PQR$, so

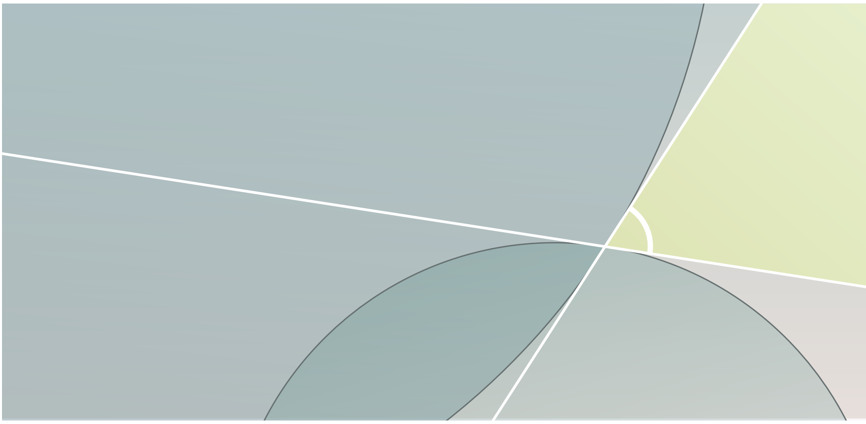
$$(\angle OPR) = (\angle PQR) + (\angle PRQ).$$

Substituting that in,

$$(\angle\sigma(P)\sigma(R)\sigma(Q)) = ((\angle PQR) + (\angle PRQ)) - (\angle OQR) = (\angle PRQ).$$



Since R is on the circle with diameter PQ, $\angle PRQ$ is a right angle. Therefore $\angle\sigma(P)\sigma(R)\sigma(Q)$ is a right angle as well, and so $\sigma(R)$ lies on the circle with diameter $\sigma(P)\sigma(Q)$. A word of warning: while $\sigma(c)$ is a circle, σ does not map the center of c to the center of $\sigma(c)$. □



Since an inversion σ doesn't map lines to lines, it doesn't really make much sense to ask whether $\sigma(\angle ABC) \simeq (\angle ABC)$. Instead, let's take a page from the book of calculus. In calculus, the angle between intersecting curves is measured by zooming into the infinitesimal level, at which point the angle between the curves becomes the angle between their tangent lines. A mapping that preserves those angles between curves is called a *conformal map*. Inversion does preserve angles in this sense.

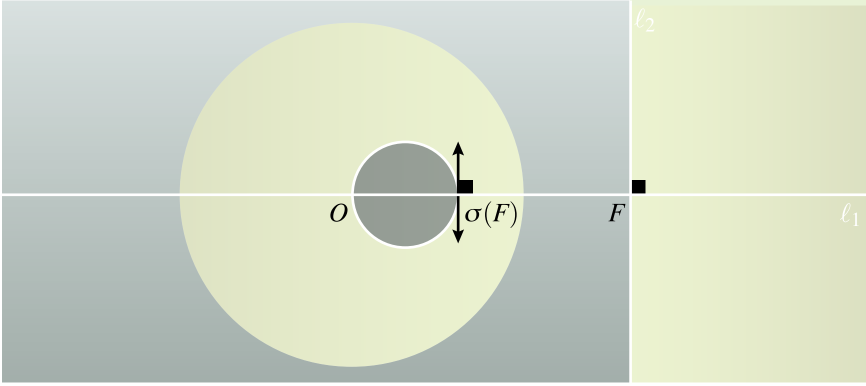
THM: INVERSION IS CONFORMAL

Let σ be the inversion across the circle $\mathcal{C}$ with center O and radius r. Let ℓ_1 and ℓ_2 be curves that intersect at some point P other than O. The curves may be both lines, both circles, or one of each. Let P be the intersection of ℓ_1 and ℓ_2. Then the angle between ℓ_1 and ℓ_2 at P is the same as the angle between $\sigma(\ell_1)$ and $\sigma(\ell_2)$ at $\sigma(P)$.

Proof. There are a lot of cases here, particularly since the scenarios where one or both of the curves pass through O require their own attention. I will do the part where ℓ_1 and ℓ_2 are lines, but leave the rest as an exercise. Note first that ℓ_1 and ℓ_2 cannot both pass through O, for if they did, then their inversion P would occur at O– it doesn't make sense to talk of the image of that point, which is why that scenario was specifically prohibited in the statement of the theorem.

Suppose that ℓ_1 passes through O, but that ℓ_2 does not.

Then σ will map ℓ_1 to itself and will map ℓ_2 to a circle which passes through O. In the course of the proof of that second fact, we found out that if F is the foot of the perpendicular to ℓ_2 from O, then $O\sigma(F)$ will be a diameter of $\sigma(\ell_2)$. On the chance that ℓ_1 and ℓ_2 intersect exactly at F, then ℓ_1 and ℓ_2 will intersect at right angles, and in that case, the diameter of $\sigma(\ell_2)$ will lie along the line ℓ_1. Thus the tangent line to the circle $\sigma(\ell_2)$ at $\sigma(F)$ will again intersect $\sigma(\ell_1)$ at a right angle.

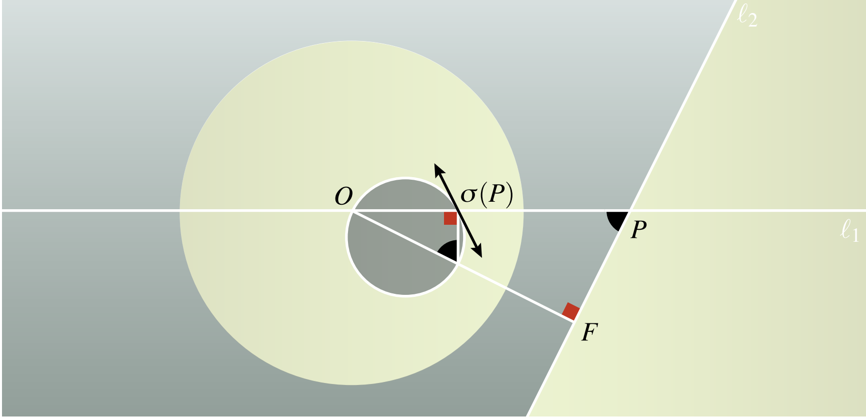


More generically, when ℓ_1 and ℓ_2 intersect at a point P other than F, then their angle of intersection is $\angle OPF$, and

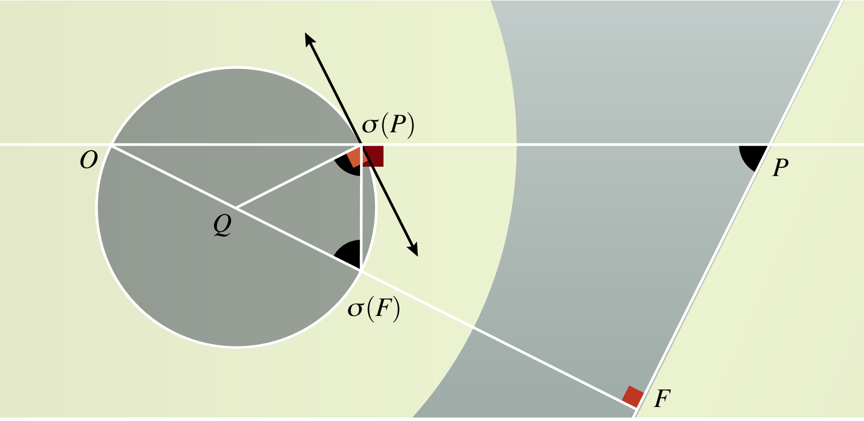
$$\triangle OPF \sim \triangle O\sigma(F)\sigma(P),$$

so $\angle OPF \simeq \angle O\sigma(F)\sigma(P)$. Let Q be the center of the circle $\sigma(\ell_2)$. Both $Q\sigma(F)$ and $Q\sigma(P)$ are radii of that circle, so $\triangle Q\sigma(F)\sigma(P)$ is isosceles, and by the Isosceles Triangle Theorem,

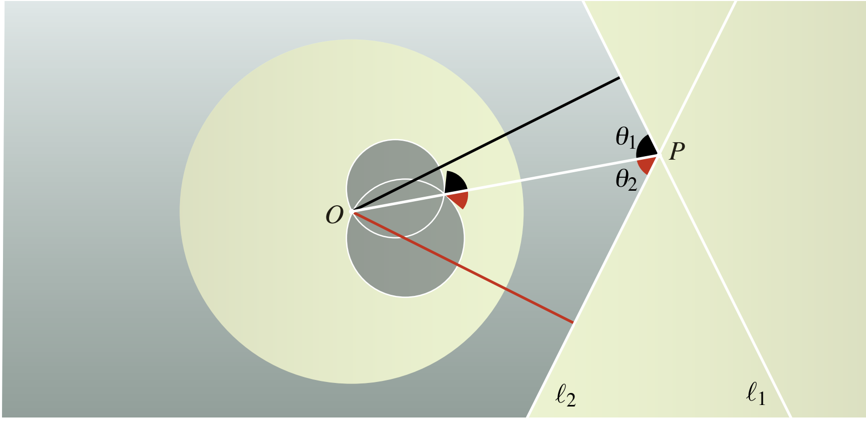
$$\angle Q\sigma(F)\sigma(P) \simeq \angle Q\sigma(P)\sigma(F).$$



Now focus on what is happening right around $\sigma(P)$. Both $\angle Q\sigma(P)\sigma(F)$ and the angle between $\sigma(\ell_1)$ and $\sigma(\ell_2)$ are complementary to the same angle. That means they must be congruent.



Suppose that neither ℓ_1 nor ℓ_2 pass through O.
In this case, the line $\leftarrow OP \rightarrow$ splits the angle formed by ℓ_1 and ℓ_2 into two pieces. Let θ_1 be the angle between ℓ_1 and OP, and let θ_2 be the angle between ℓ_2 and $\leftarrow OP \rightarrow$. Now $\leftarrow OP \rightarrow$ will also split the angle between $\sigma(\ell_1)$ and $\sigma(\ell_2)$. From our previous work, the angle between $\sigma(\ell_1)$ and $\leftarrow OP \rightarrow$ is the same as the angle between ℓ_1 and $\leftarrow OP \rightarrow$, and the angle between $\sigma(\ell_2)$ and $\leftarrow OP \rightarrow$ is the same as the angle between ℓ_2 and $\leftarrow OP \rightarrow$. Adding the pieces together, the angle between $\sigma(\ell_1)$ and $\sigma(\ell_2)$ is the same as the angle between ℓ_1 and ℓ_2. □

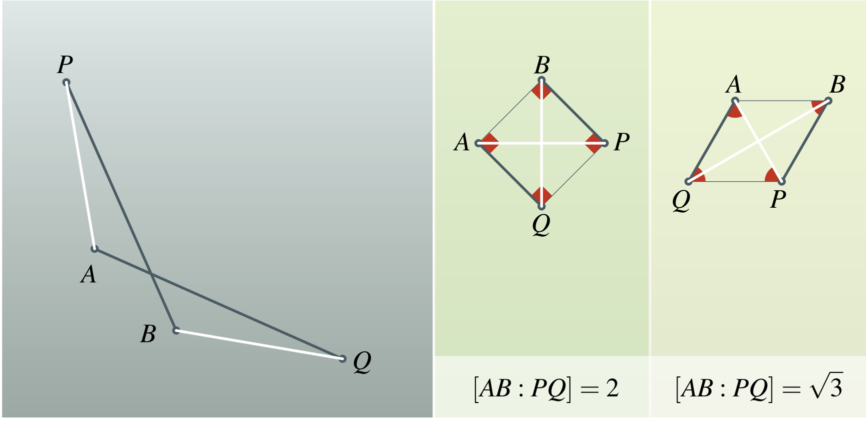


That's some good news about angle measure. Unfortunately, we already know that the news isn't so good when it comes to measuring distance. Does inversion have *any* kind of distance invariant? As a matter of fact, yes– to find it you have to play around with the similarity property of inversion. The invariant is something called the *cross ratio*.

DEF: CROSS RATIO

Let A, B, P and Q be four distinct points. The cross ratio of A, B, P, and Q, written $[AB, PQ]$ is the product of ratios

$$[AB, PQ] = \frac{|AP|}{|AQ|} \cdot \frac{|BQ|}{|BP|}.$$



$[AB : PQ] = 2$

$[AB : PQ] = \sqrt{3}$

THM: INVERTING THE CROSS RATIO

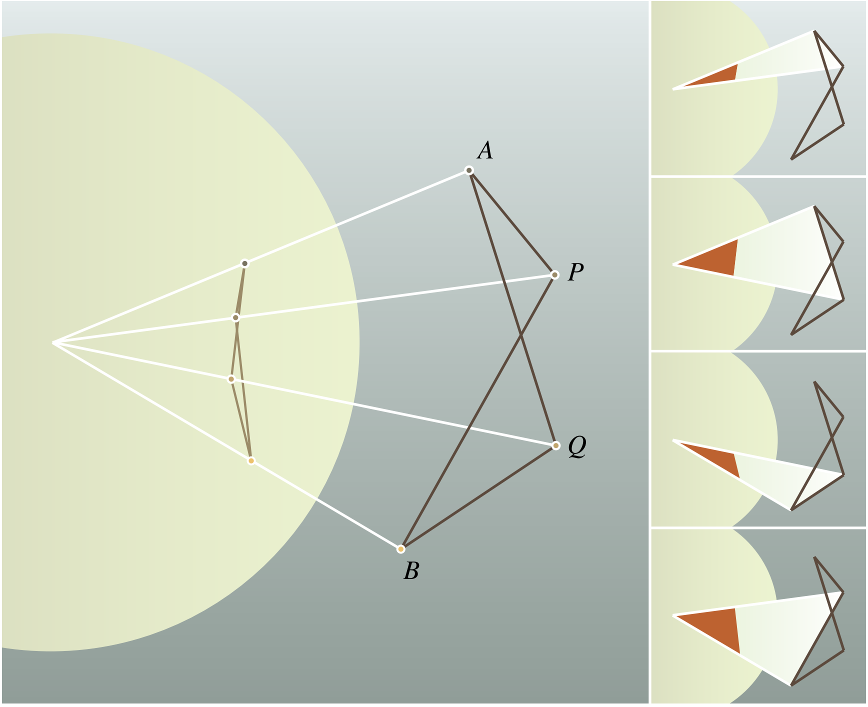
The cross ratio is invariant under inversion. That is, for any inversion σ, and points A, B, P, and Q,

$$[AB, PQ] = [\sigma(A)\sigma(B), \sigma(P)\sigma(Q)].$$

Proof. By the similarity property,

$$\frac{|\sigma(A)\sigma(P)|}{|AP|} = \frac{|O\sigma(P)|}{|OA|} \qquad \frac{|\sigma(B)\sigma(Q)|}{|BQ|} = \frac{|O\sigma(Q)|}{|OB|}$$

$$\frac{|\sigma(A)\sigma(Q)|}{|AQ|} = \frac{|O\sigma(Q)|}{|OA|} \qquad \frac{|\sigma(B)\sigma(P)|}{|BP|} = \frac{|O\sigma(P)|}{|OB|}$$



so

$$\frac{|\sigma(A)\sigma(P)|}{|AP|} \cdot \frac{|\sigma(B)\sigma(Q)|}{|BQ|} \cdot \frac{|AQ|}{|\sigma(A)\sigma(Q)|} \cdot \frac{|BP|}{|\sigma(B)\sigma(P)|}$$

$$= \frac{|O\sigma(P)|}{|OA|} \cdot \frac{|O\sigma(Q)|}{|OB|} \cdot \frac{|OA|}{|O\sigma(Q)|} \cdot \frac{|OB|}{|O\sigma(P)|} = 1.$$

Multiplying across,

$$\frac{|\sigma(A)\sigma(P)|}{|\sigma(A)\sigma(Q)|} \cdot \frac{|\sigma(B)\sigma(Q)|}{|\sigma(B)\sigma(P)|} = \frac{|AP|}{|AQ|} \cdot \frac{|BQ|}{|BP|}.$$

and so

$$[\sigma(A)\sigma(B), \sigma(P)\sigma(Q)] = [AB, PQ].$$

□

We will see the cross ratio again. It is an essential tool for building non-Euclidean geometry.

Exercises

1. Verify the equation for stereographic projection from the south pole that is given in the chapter:

$$\Phi_S(x,y,z) = \left(\frac{rx}{r+z}, \frac{ry}{r+z}\right).$$

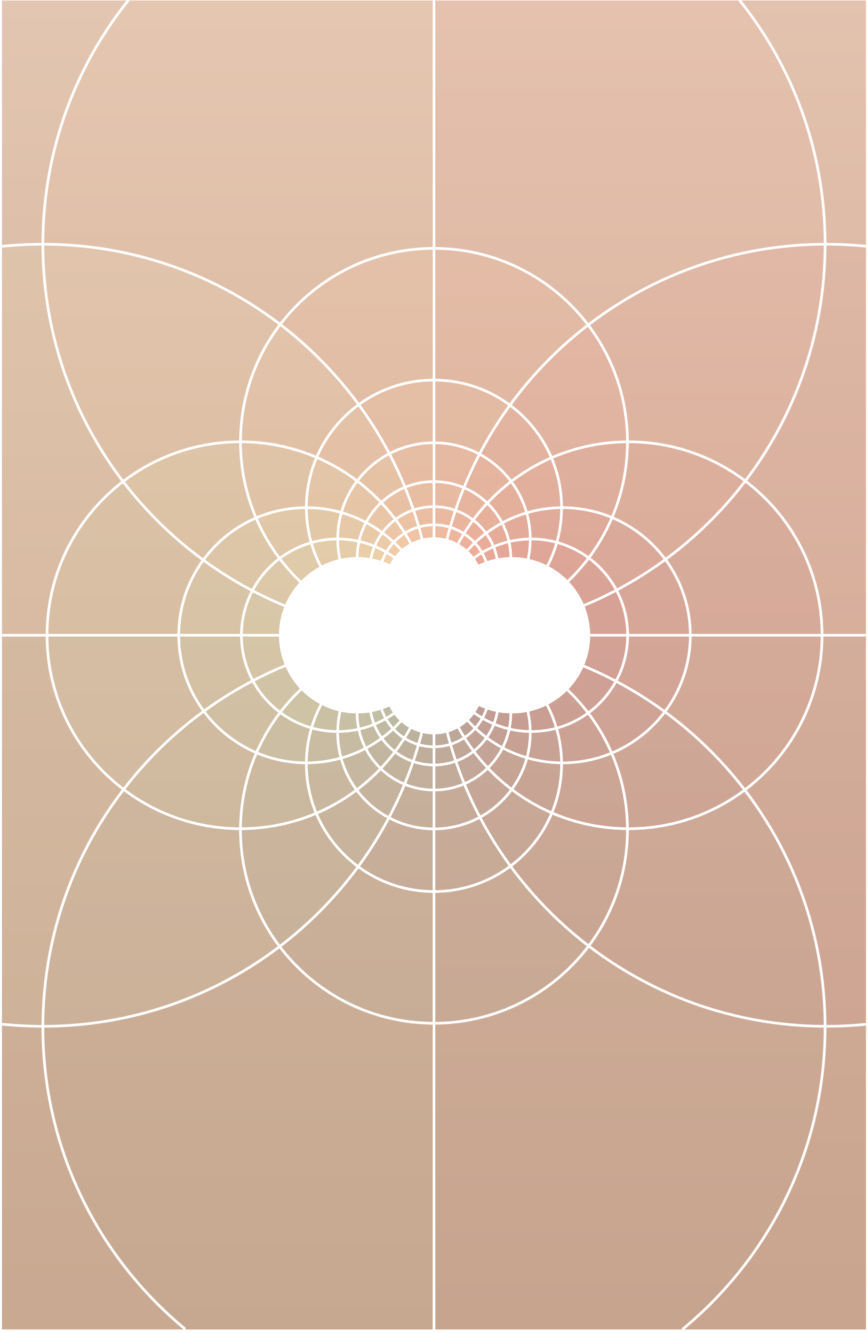
2. Verify that if two circles intersect at points P and Q, then their angle of intersection as measured at P is the same as the angle of intersection as measured at Q.

3. Complete the proof that inversion is conformal. There are two cases to consider: (a) where both ℓ_1 and ℓ_2 are circles; and (b) where one is a circle and the other is a line.

4. There are $4! = 24$ permutations of the four letters in the cross ratio. Some of those rearrangements ultimately give the same result:

$$[PQ,AB] = \frac{|PA|}{|PB|} \cdot \frac{|QB|}{|QA|} = [AB,PQ],$$

but others do not. Determine which of the 24 permutations *do* yield the same result.



36 INVERSION II

Matrix/vector arithmetic is the natural language of isometries, but it does not do so well when it comes to describing inversion. For that, it is better to translate the problem into the language of complex arithmetic. We will start off this lesson with a review of that complex arithmetic. I assume that readers who have made it this far have some experience working with complex numbers– if not, then this cursory overview is probably not sufficient– our needs here are pretty minimal, but they are not non-existent. Any standard text on complex numbers will get you up to speed in next to no time.

Complex numbers, complex arithmetic

A complex number has the form $a+bi$ where a and b are real numbers and i is a solution to the equation $x^2=-1$. The set of complex numbers $\mathbb{C}$ contains all the real numbers in the form $a+0i$, but since the square of any real number is positive, i is not itself a real number. Thus $\mathbb{C}$ is properly an extension of the real numbers. There is a bijection between complex numbers and points (or vectors) in $\mathbb{R}^2$ via

$$a+bi \longleftrightarrow (a,b).$$

This correspondence is what allows us to translate problems in $\mathbb{R}^2$ into problems in $\mathbb{C}$. Why would we want to do that? Well, the basic advantage of $\mathbb{C}$ over $\mathbb{R}^2$ is that $\mathbb{C}$ is a field– any two numbers in it can be added, subtracted, multiplied, and (except in the case of 0) divided. In contrast, while the vectors of $\mathbb{R}^2$ are equipped with addition, subtraction, and scalar multiplication, there is no natural way to multiply or divide vectors. It is the multiplication and division operations that make it worth the effort.

Addition and subtraction in $\mathbb{C}$ are essentially the same as vector addition and subtraction. Multiplication in $\mathbb{C}$ is just "FOIL" together with the fact that $i^2 = -1$. Division is done by multiplying by the "complex conjugate".

COMPLEX ARITHMETIC

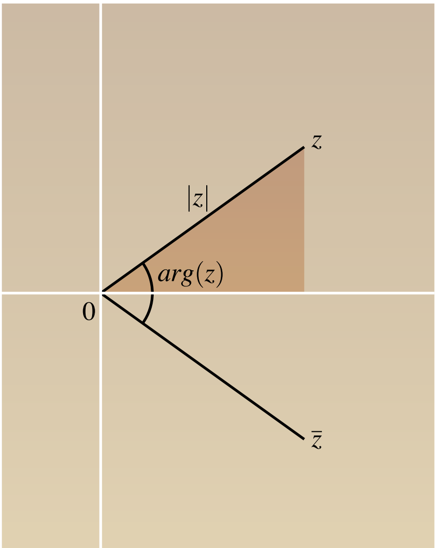
$$(a+bi)+(c+di) = (a+c)+(b+d)i$$
$$(a+bi)-(c+di) = (a-c)+(b-d)i$$
$$(a+bi)(c+di) = ac+adi+bci+bdi^2 = (ac-bd)+(ad+bc)i$$
$$\frac{a+bi}{c+di} = \frac{a+bi}{c+di}\cdot\frac{c-di}{c-di} = \frac{ac+bd}{c^2+d^2}+\frac{bc-ad}{c^2+d^2}i.$$

The *complex conjugate* of $z = a+bi$ (mentioned above) is $\overline{z} = a-bi$. The *norm* (or length or absolute value) of a complex number $z = a+bi$ is its distance from 0,

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

The *argument* of a complex number z is the measure of the angle that it forms with the real axis (as measured in the counterclockwise direction), so

$$\tan(\arg(a+bi)) = b/a.$$



Argument, norm, and complex conjugate.

The standard presentation of a complex number in the form $a+bi$ is distinctly rectangular in its construction. Complex numbers can also be expressed in a "polar form"– if $r=|z|$ and $\theta=\arg(z)$, then $a=r\cos\theta$ and $b=r\sin\theta$, so

$$a+bi=(r\cos\theta)+(r\sin\theta)i=r(\cos\theta+i\sin\theta).$$

For our purposes, this polar form is really just a stepping stone toward the ultimate goal– an "exponential form". If you have only ever been exposed to *real-valued* functions, then the trigonometric functions $\sin x$ and $\cos x$ probably seem vastly different from the exponential function e^x. For instance, $\sin x$ and $\cos x$ are bounded and periodic; the exponential function is neither bounded nor periodic. In the more expansive world of complex numbers, though, there are deep connections between these three functions. The easiest way to see those connections is by looking at their Taylor series.

Taylor series: a quick and dirty review

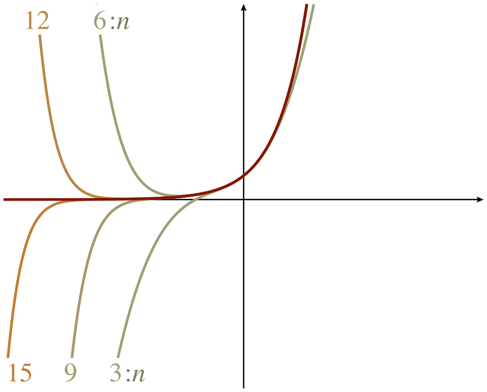
Let $f(x)$ be a function whose derivatives at a point a are all defined. The n^{th} Taylor polynomial of $f(x)$, expanded about the point a, is a specific degree n polynomial p_n that approximates $f(x)$ in a region right around a. Its coefficients are calculated by matching the function value and the first n derivatives at a of p_n with those of $f(x)$. Now all these derivatives at a give *local* information about the function right around the a (they tell us whether the function is increasing or decreasing, concave up or concave down). It makes sense that taking more derivatives would improve that approximation around a and perhaps extend the region for which the approximation is "fairly close". Taken to its natural extreme, then, if we want the best approximation, we've got to let $n\to\infty$, and look at the Taylor *series* p_∞ that approximates $f(x)$. Matching up derivatives gives the formula

$$p_\infty(x)=\sum_{n=0}^{\infty}\frac{p^{(n)}(a)}{n!}(x-a)^n.$$

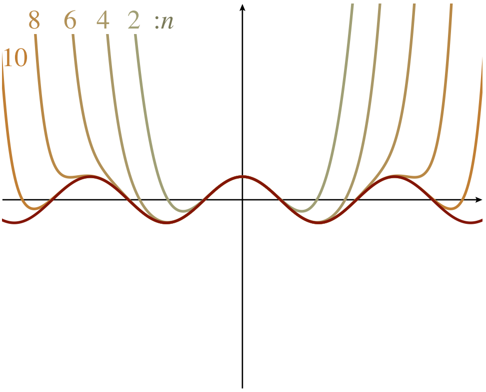
Even with an infinite sum, there is in general no guarantee that $p_\infty(x)$ will be a good approximation of $f(x)$ as you move away from a (in fact, there is now the additional question of whether the series converges at all).

Here's the good news: the Taylor series of e^x, $\sin x$, and $\cos x$ do converge to exactly the function value for all x (no matter what a value is chosen). The Taylor series expansions about $a = 0$ for these functions are

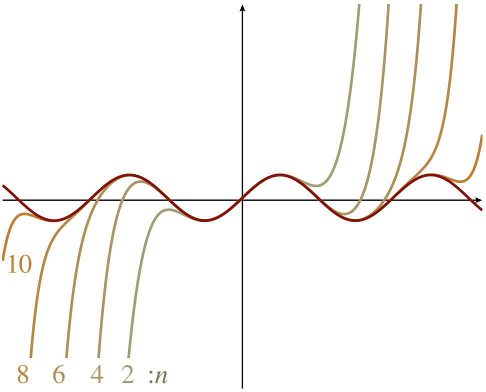
$$e^x = \sum_{n=0}^{\infty} \frac{1}{n!} x^n$$



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



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



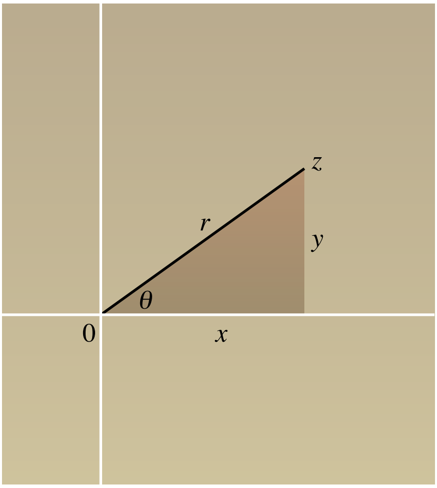
Now let's see how that allows us to relate the sine and cosine functions to the exponential. Cosine is an even function and sine is an odd function, so if we take the series expansion of $e^{i\theta}$ and segregate the even powers from the odd powers:

$$\begin{aligned}
e^{i\theta} &= \sum_{n=0}^{\infty} \frac{(i\theta)^n}{n!} \\
&= \sum_{n=0}^{\infty} \frac{(i\theta)^{2n}}{(2n)!} + \sum_{n=0}^{\infty} \frac{(i\theta)^{2n+1}}{(2n+1)!} \\
&= \sum_{n=0}^{\infty} \frac{i^{2n}\theta^{2n}}{(2n)!} + \sum_{n=0}^{\infty} \frac{i \cdot i^{2n}\theta^{2n+1}}{(2n+1)!} \\
&= \sum_{n=0}^{\infty} \frac{(-1)^n\theta^{2n}}{(2n)!} + i\sum_{n=0}^{\infty} \frac{(-1)^n\theta^{2n+1}}{(2n+1)!} \\
&= \cos\theta + i\sin\theta.
\end{aligned}$$

Therefore the polar form of a complex number z can be rewritten in an exponential form

$$z = r(\cos\theta + i\sin\theta) = re^{i\theta}.$$

All the rules of exponents still apply, so this is a very powerful alternative to the rectangular form for a complex number.



RECT. $z = x + iy$

TRIG. $z = r\cos\theta + ir\sin\theta$

EXP. $z = re^{i\theta}$

The geometry of complex arithmetic

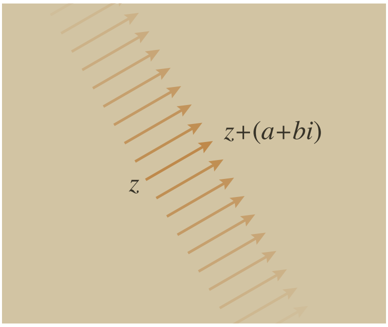
Adding the complex number $z = a + bi$ to another complex number w has the effect of translating w by the vector $\langle a, b\rangle$. Subtracting z from w has a similar effect, but the translation is in the opposite direction. For multiplication and division it is best to look at the exponential form: write $z = re^{i\theta}$ and $w = se^{i\phi}$. Then

$$zw = re^{i\theta} \cdot se^{i\phi} = rse^{i(\theta+\phi)}.$$

The effect of multiplying by z, then, is to scale from the origin by r and to rotate counterclockwise around the origin by θ. Division works similarly,

$$w/z = se^{i\phi}/re^{i\theta} = (s/r)e^{i(\phi-\theta)},$$

but this time the scaling is by $1/r$ and the rotation by θ is in the clockwise direction. For this reason, some Euclidean isometries can be described quite naturally in terms of complex arithmetic.

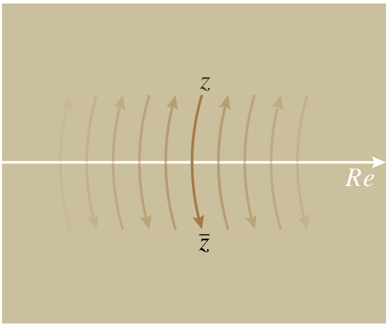


The translation

$$t\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x+a \\ y+b \end{pmatrix}$$

becomes

$$t(z) = z + (a+bi).$$

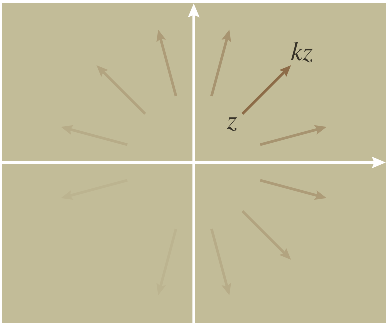


The reflection across the real (x-) axis

$$s\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ -y \end{pmatrix}$$

becomes

$$s(z) = \bar{z}.$$

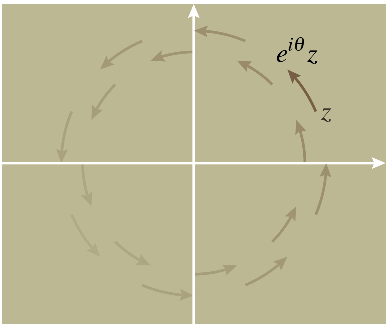


The dilation by k about the origin

$$d\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}kx\\ky\end{pmatrix}$$

becomes

$$d(z)=kz.$$

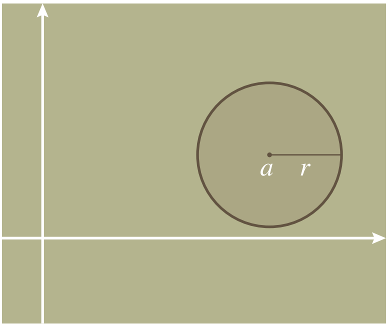


The rotation by θ about the origin

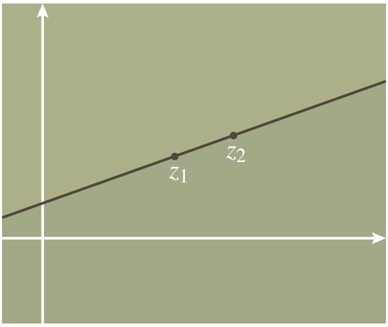
$$r\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}$$

becomes

$$r(z)=e^{i\theta}\cdot z.$$



For any complex number a and positive real number r, the equation $|z-a|=r$ describes a circle with center a and radius r.



For any two complex numbers z_1 and z_2, the function $r:\mathbb{R}\to\mathbb{C}$ defined by

$$r(t)=z_1+t(z_2-z_1)$$

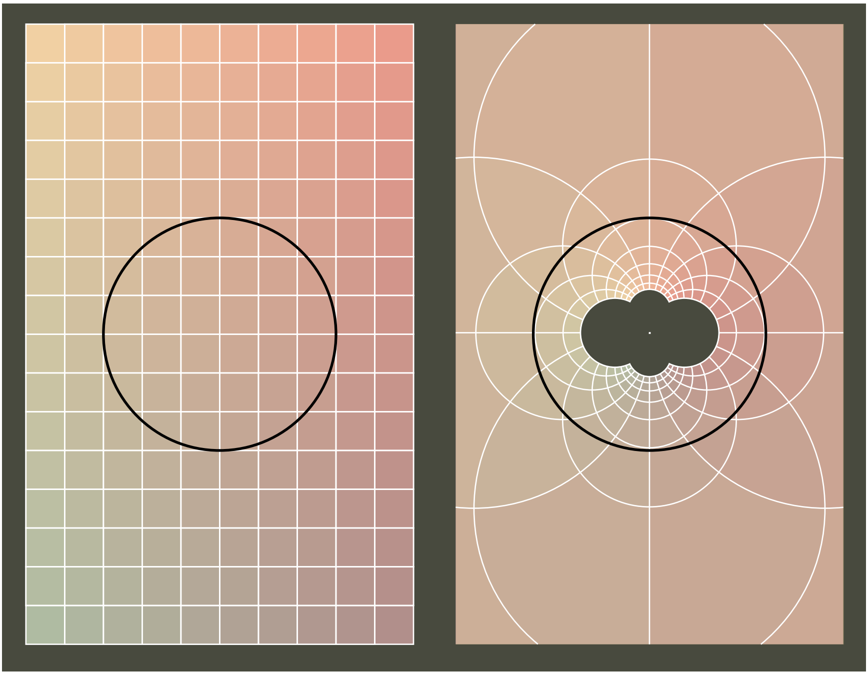
describes the line through z_1 and z_2.

The whole point of this, remember, was to find a workable equation for inversion.

THM: AN EQUATION FOR INVERSION

The inversion σ across $|z| = r$, the circle of radius r centered at the origin, is given by the equation

$$\sigma(z) = r^2/\bar{z}.$$



Inverting a grid.

Proof. Write $z = Re^{i\theta}$. According to the definition of inversion, $\sigma(z)$ is on the ray from the origin passing through z and its distance from the origin is r^2/R. The points on this ray all have an argument of θ. Therefore

$$\sigma(z) = \frac{r^2}{R}e^{i\theta} = \frac{r^2}{Re^{-i\theta}} = r^2/\bar{z}.$$

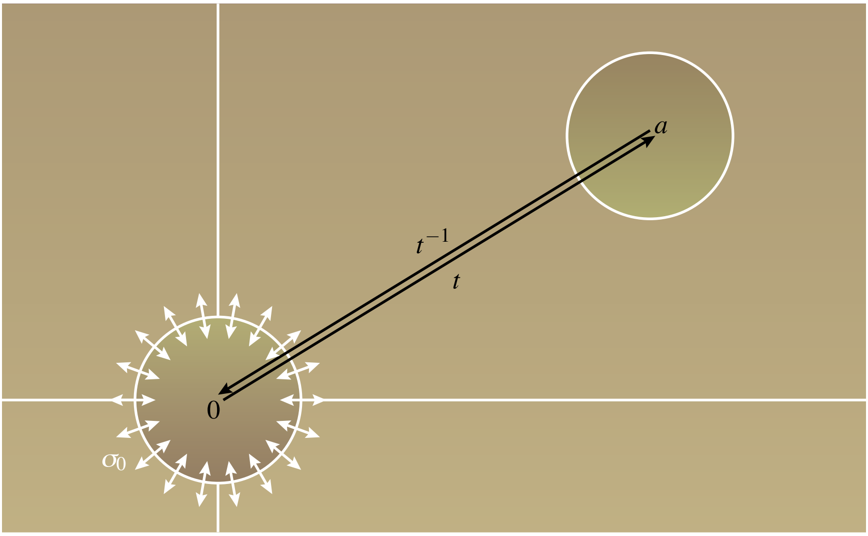
□

More generally, we can use a change of coordinates to find the equation of an inversion across a circle that is not centered at the origin.

COR: GENERAL FORM FOR AN INVERSION

The inversion σ across $|z-a| = r$, the circle of radius r centered at a, is given by the equation

$$\sigma(z) = \frac{r^2}{\overline{z-a}} + a.$$



Proof. To use the previous formula, we need to work with a change of coordinates that repositions the origin at a. We can use the translation $t(z) = z + a$. If we label σ_0 as the inversion across the circle of radius r centered at the origin, then

$$\begin{aligned}\sigma(z) &= t \circ \sigma_0 \circ t^{-1}(z)\\ &= t \circ \sigma_0(z-a)\\ &= t\left(\frac{r^2}{\overline{z-a}}\right)\\ &= \frac{r^2}{\overline{z-a}} + a.\end{aligned}$$

□

Properties of the norm and conjugate

A lot of the arithmetic of complex numbers plays on a few simple properties of the norm and the conjugate. I am providing the proof of a couple of these properties but leaving the rest to you.

THM: PROPERTIES OF THE CONJUGATE
For complex numbers z, z_1, and z_2

$$\begin{aligned}
&\overline{(\overline{z})} = z\\
&\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2}\\
&\overline{z_1 - z_2} = \overline{z_1} - \overline{z_2}\\
&\text{If } z = re^{i\theta}\text{, then } \overline{z} = re^{-i\theta}.\\
&\overline{z_1 \cdot z_2} = \overline{z_1} \cdot \overline{z_2}\\
&\overline{z_1/z_2} = \overline{z_1}/\overline{z_2} \quad (\text{if } z_2 \neq 0)
\end{aligned}$$

Proof. Let me just take the claim that $\overline{z_1 \cdot z_2} = \overline{z_1} \cdot \overline{z_2}$. Writing $z_1 = r_1 e^{i\theta_1}$ and $z_2 = r_2 e^{i\theta_2}$, then

$$\begin{aligned}
\overline{z_1 \cdot z_2} &= \overline{r_1 e^{i\theta_1} r_2 e^{i\theta_2}}\\
&= \overline{r_1 r_2 e^{i(\theta_1+\theta_2)}}\\
&= r_1 r_2 e^{-i(\theta_1+\theta_2)}\\
&= r_1 e^{-i\theta_1} r_2 e^{-i\theta_2}\\
&= \overline{z_1} \cdot \overline{z_2}.
\end{aligned}$$

□

THM: PROPERTIES OF THE NORM
For complex numbers z, z_1, and z_2

$$\begin{aligned}
&z\overline{z} = |z|^2\\
&|\overline{z}| = |z|\\
&|z_1 \cdot z_2| = |z_1| \cdot |z_2|\\
&|z_1/z_2| = |z_1|/|z_2| \quad (z_2 \neq 0)\\
&|z_1 \pm z_2| \leq |z_1| + |z_2|.
\end{aligned}$$

Proof. For example, the first one is easy to verify: write $z = re^{i\theta}$. Then

$$z\overline{z} = re^{i\theta} \cdot re^{-i\theta} = r^2 = |z|^2.$$

□

Exercises

1. Let r be the counterclockwise rotation by $\pi/2$ around the point $1+2i$. Write a complex equation describing r.

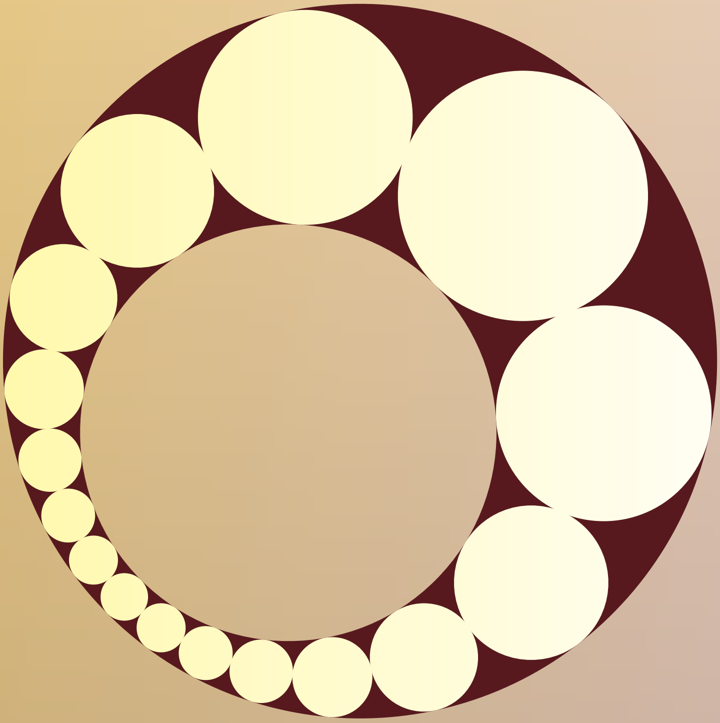
2. Let s be the reflection across the line $r(t) = t + (1+t)i$. Write a complex equation describing s.

3. Find the equation for the inversion through the circle with radius 2 and center $1+5i$.

4. Let $\mathcal{C}_1$ be the circle with radius one centered at the complex number a_1 and let $\mathcal{C}_2$ be the circle with radius one centered at the complex number a_2. Let i_1 be the inversion across $\mathcal{C}_1$ and i_2 be the inversion across $\mathcal{C}_2$. Describe the fixed point(s) of the composition map $i_2 \circ i_1$ in terms of a_1 and a_2.

5. Verify the remaining properties of the conjugate.

6. Verify the remaining properties of the norm.



37 INVERSION – APPLICATIONS

In the last two lessons we came to some understanding of the basic workings of inversion. In the long run, this should smooth the transition into non-Euclidean geometry. Before we make that transition, though, let's take a brief vacation, and look at two nice little theorems that can be proved with the help of a well-placed inversion. Both results involve a chain of mutually tangent circles. In the first, the chain of circles is inside a shape called an arbelos; in the second, it is wedged between two circles.

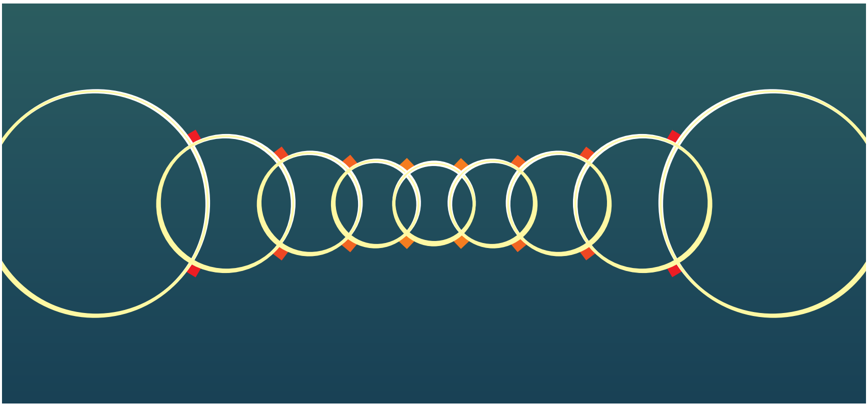
A typical inversion proof begins by applying a particular inversion to what appears to be a complicated picture. The inversion transforms that complicated picture into a simpler one with much more apparent symmetry than the original. With this newfound symmetry, the rest of the proof is easy. So if that's all there is to it, then the trick is to find the right inversion to start with. A good place to start looking is with *orthogonal circles*.

Orthogonal circles

Recall that the angle between two intersecting circles is measured by the angle between their tangent lines.

DEF: ORTHOGONAL CIRCLES

Two intersecting circles $\mathcal{C}_1$ and $\mathcal{C}_2$ are *orthogonal* if the angle between them is a right angle.



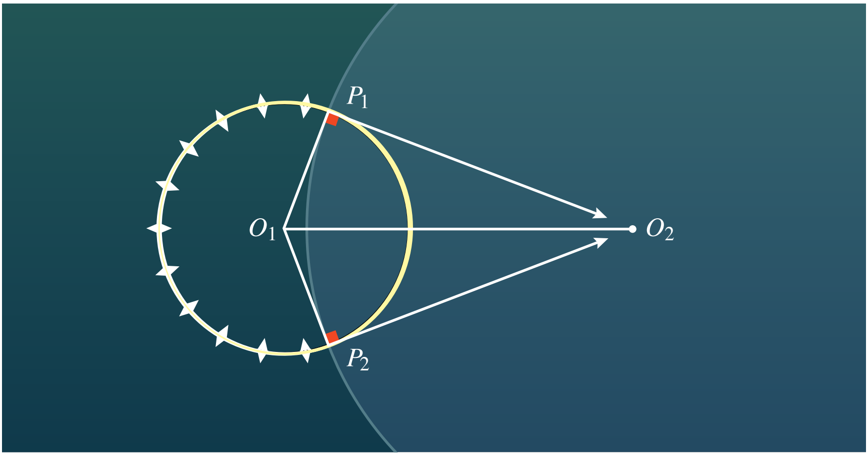
A chain of pairwise orthogonal circles.

The reason why orthogonal circles may be worth a look is this:

> THM: ORTHOGONAL CIRCLES ARE INVARIANT UNDER INVERSION
> Suppose that $\mathcal{C}_1$ and $\mathcal{C}_2$ are orthogonal circles and that σ is the inversion across $\mathcal{C}_1$. Then $\sigma(\mathcal{C}_2) = \mathcal{C}_2$.

Proof. Orthogonal circles will intersect twice, and both points of intersection are fixed by σ (since they are on $\mathcal{C}_1$). So we know a couple of points on $\sigma(\mathcal{C}_2)$ already, but we can close the door on this problem by finding the center of $\sigma(\mathcal{C}_2)$. Labels:

P_1 , P_2: the intersections of $\mathcal{C}_1$ and $\mathcal{C}_2$;
O_1, O_2: the centers of $\mathcal{C}_1$ and $\mathcal{C}_2$; and
Q: the center of $\sigma(\mathcal{C}_2)$.



Remember that σ is conformal– the angle between $\mathcal{C}_1$ and $\mathcal{C}_2$ is a right angle, so the angle between $\sigma(\mathcal{C}_2)$ and $\sigma(\mathcal{C}_1) = \mathcal{C}_1$ must be a right angle too. That means Q must be on both the line through P_1 that is perpendicular to O_1P_1 and on the line through P_2 that is perpendicular to O_2P_2. Well, only one point meets those criteria– it is O_2. So $\sigma(\mathcal{C}_2)$ is a circle centered at $Q = O_2$ and passing through P_1 and P_2– $\sigma(\mathcal{C}_2)$ is $\mathcal{C}_2$. □

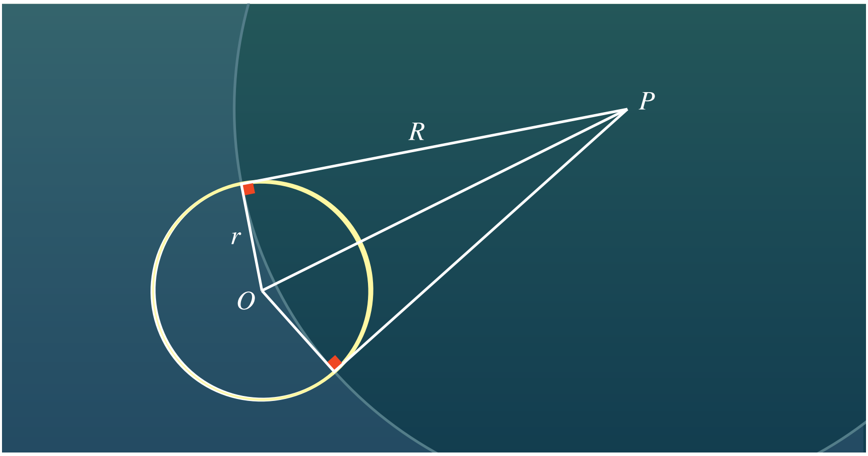
Note that while this theorem does say that $\mathcal{C}_2$ is invariant, it does not say that all the points of $\mathcal{C}_2$ are fixed. In fact, σ fixes only two points of $\mathcal{C}_2$– the points of intersection of $\mathcal{C}_1$ and $\mathcal{C}_2$. More subtly, while $\sigma(\mathcal{C}_2)$ still has center O_2, $\sigma(O_2) \neq O_2$.

If these orthogonal circles are going to play key roles in our proofs, we need to have some idea about how prevalent they are. The next two results address that issue.

THM: ON THE PREVALENCE OF ORTHOGONAL CIRCLES
Let $\mathcal{C}$ be a circle and P be a point outside $\mathcal{C}$. Then there is a unique circle centered at P which is orthogonal to $\mathcal{C}$.

Proof. It is easier to play this argument in reverse. Say that $\mathcal{C}$ has center O and radius r, and suppose that P is in fact the center of a circle orthogonal to $\mathcal{C}$. What would its radius R be? Well, inside of every pair of orthogonal circles is a right triangle. Two of its vertices are the centers of the circles; the third is one the points of intersection of the two circles. By the Pythagorean Theorem,

$$r^2 + R^2 = |OP|^2 \implies R = [|OP|^2 - r^2]^{1/2}.$$



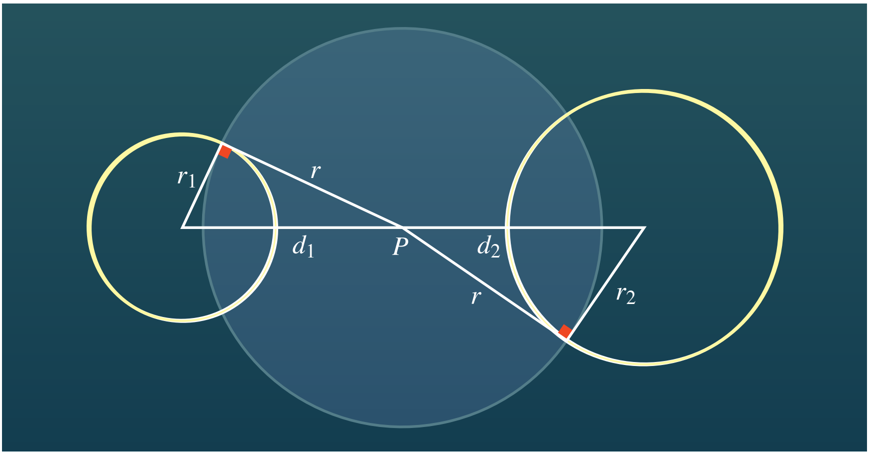
Since P is outside $\mathcal{C}$, $|OP|$ is greater than r, so this equation does have a solution. But the Pythagorean Theorem is a bi-directional– that is, since this equation does have a solution, this right triangle can be constructed with hypotenuse OP. In fact, it can be constructed in two ways– one on each side of OP. In either case, the leg with length R is the radius of the one circle that centered at P and orthogonal to $\mathcal{C}$. □

What about *pairs* of circles– given circles $\mathcal{C}_1$ and $\mathcal{C}_2$, are there any circles that are orthogonal to both? Generally the answer to this question is yes, but not always. The exception is this: if $\mathcal{C}_1$ and $\mathcal{C}_2$ are concentric circles (that is, they have the same center), then there are no circles orthogonal to both $\mathcal{C}_1$ and $\mathcal{C}_2$. [This is actually a particularly good, rather than a particularly bad, case: if $\mathcal{C}_1$ and $\mathcal{C}_2$ are concentric, then it is the *lines* through their mutual center that will intersect both circles at right angles.] But as long as $\mathcal{C}_1$ and $\mathcal{C}_2$ are not concentric, there are circles orthogonal to both. They are more scarce now, though, and the conditions for a point P to be the center of an orthogonal circle are more demanding. Label

r_1: the radius of $\mathcal{C}_1$,
r_2: the radius of $\mathcal{C}_2$,
d_1: the distance from the center of $\mathcal{C}_1$ to P,
d_2: the distance from the center of $\mathcal{C}_2$ to P.

Suppose $\mathcal{C}$ is a circle with radius r and center P which is orthogonal to both $\mathcal{C}_1$ and $\mathcal{C}_2$. By the Pythagorean Theorem,

$$r^2 + r_1^2 = d_1^2 \quad \& \quad r^2 + r_2^2 = d_2^2.$$



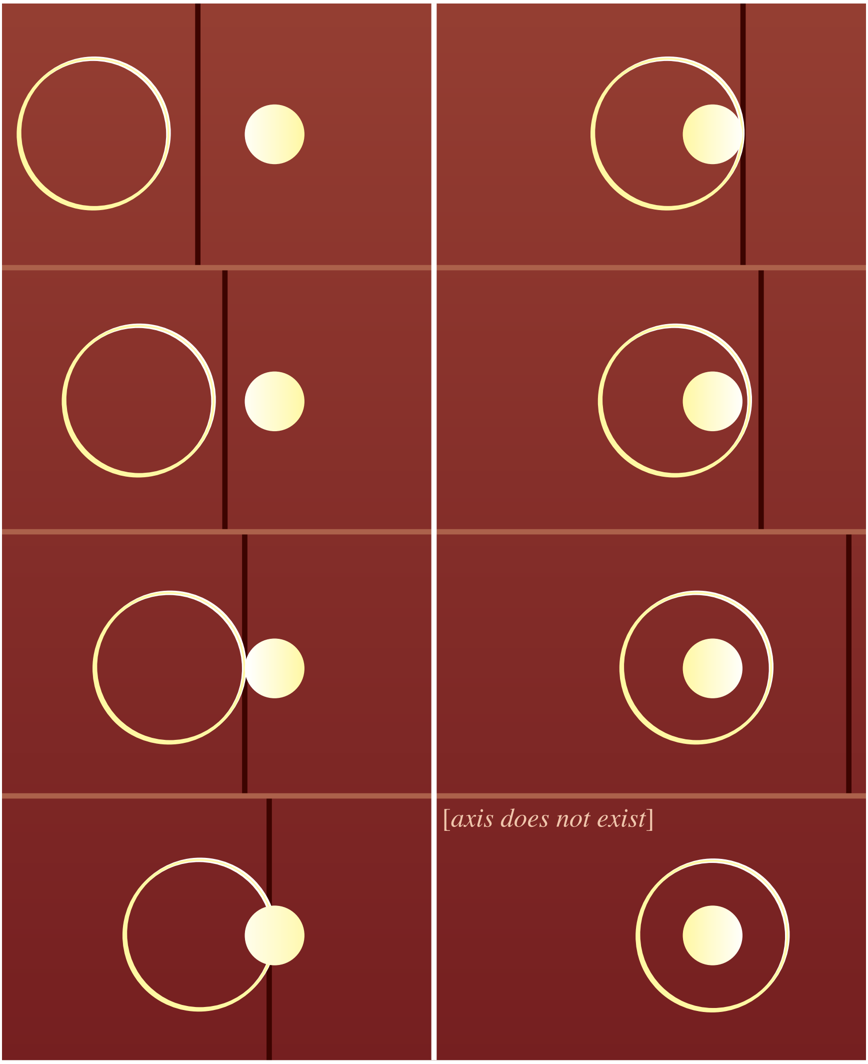
Solving for r^2 in this pair, and then setting them equal,

$$d_1^2 - r_1^2 = d_2^2 - r_2^2 \implies r_1^2 - r_2^2 = d_1^2 - d_2^2.$$

Since both r_1 and r_2 are given by the circles $\mathcal{C}_1$ and $\mathcal{C}_2$, there are only certain combinations of d_1 and d_2 which will make this equation work. The points that satisfy this condition form the *radical axis* of $\mathcal{C}_1$ and $\mathcal{C}_2$. That is all well and good, but what does the radical axis look like?

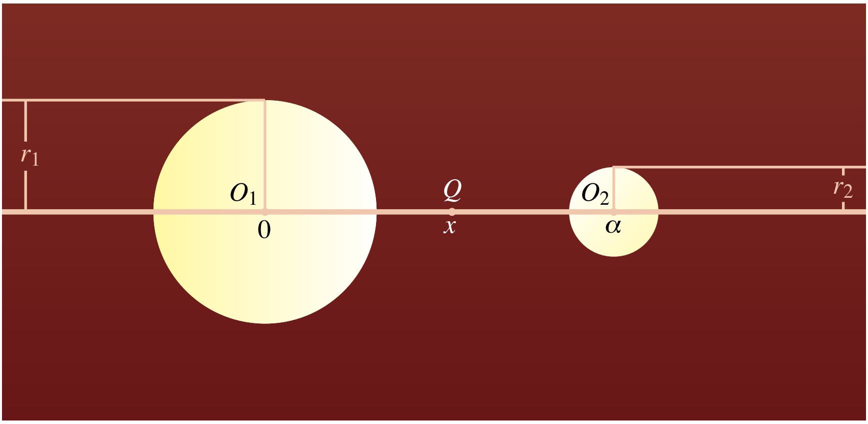
THM: THE RADICAL AXIS

The radical axis of a pair of non-concentric circles is a line that is perpendicular to the line through the circle's centers.



The radical axis: examples.

Proof. Let O_1 and O_2 be the centers of the two circles in question. If the radical axis really is a line perpendicular to O_1O_2 as claimed, then it must intersect O_1O_2. Let's start by looking for that intersection. To do so, set up a coordinate system measuring signed distance along the line O_1O_2. Center the coordinate system at O_1, so that O_1 is at coordinate 0, and label the corresponding coordinate for O_2 as α.



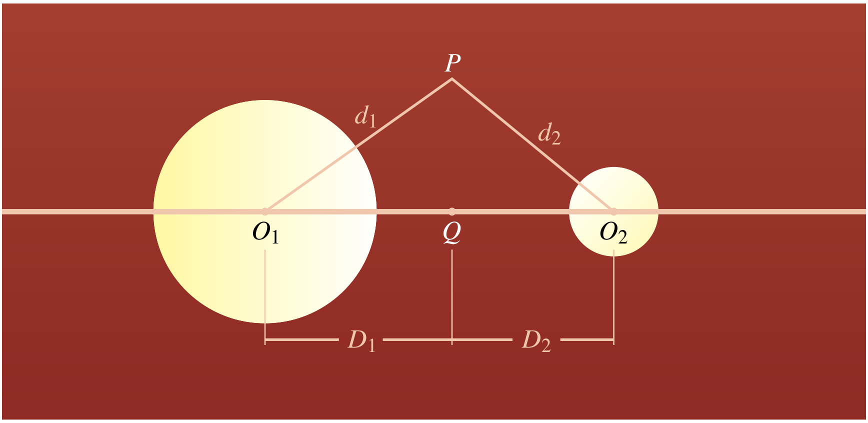
In order for the point at coordinate x to be on the radical axis, it must satisfy the equation

$$\begin{aligned} r_1^2 - r_2^2 &= x^2 - (x-\alpha)^2 \\ &= x^2 - (x^2 - 2\alpha x + \alpha^2) \\ &= 2\alpha x - \alpha^2. \end{aligned}$$

Now solve this equation for x to get $x = (r_1^2 - r_2^2 + \alpha^2)/(2\alpha)$. As long as the circles are not concentric, α will not be zero and this equation will have a (unique) solution. Therefore, exactly one point of $\leftarrow O_1O_2 \rightarrow$ is on the radical axis– call this point Q. Let's label some distances too: $D_1 = |QO_1|$ and$D_2 = |QO_2|$. Label the line which passes through Q and is perpendicular to O_1O_2 as ℓ. Of course,

$$D_1^2 - D_2^2 = r_1^2 - r_2^2,$$

since Q is on the radical axis. We want to show that the other points that satisfy this condition are all on ℓ (it is really an "if and only if" statement, which I think will be apparent in the proof). Take another point P and put $d_1 = |O_1P|$, and $d_2 = |O_2P|$.



The picture when Q is between circle centers.

Look at the two triangles $\triangle O_1QP$ and $\triangle O_2QP$. According to the Law of Cosines,

$$\begin{aligned}
d_1^2 &= D_1^2 + |PQ|^2 - 2D_1|PQ|\cos(\angle O_1QP) \\
d_2^2 &= D_2^2 + |PQ|^2 - 2D_2|PQ|\cos(\angle O_2QP).
\end{aligned}$$

Therefore

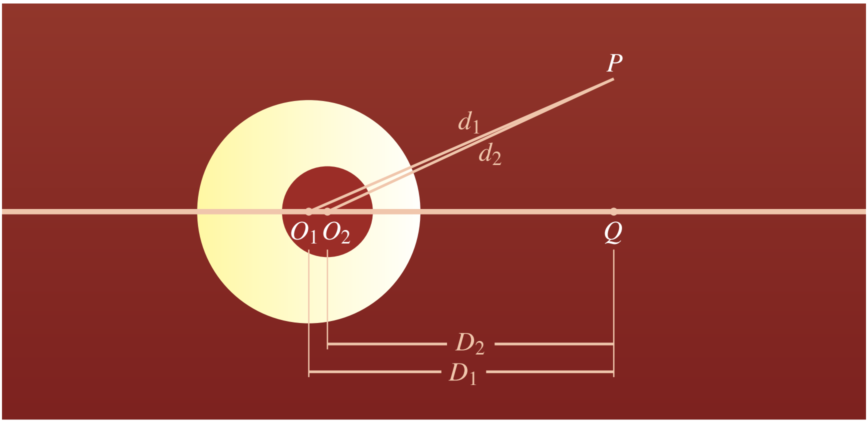
$$\begin{aligned}
d_1^2 - d_2^2 = &\left[D_1^2 + |PQ|^2 - 2D_1|PQ|\cos(\angle O_1QP)\right] \\
&- \left[D_2^2 + |PQ|^2 - 2D_2|PQ|\cos(\angle O_2QP)\right] \\
= &(D_1^2 - D_2^2) - 2|PQ| \cdot \left[D_1\cos(\angle O_1QP) - D_2\cos(\angle O_2QP)\right].
\end{aligned}$$

In order for P to be on the radical axis, $d_1^2 - d_2^2$ needs to equal $r_1^2 - r_2^2$, but the term $(D_1^2 - D_2^2)$ already equals this by itself. Therefore, P will be on the radical axis when

$$d_1\cos(\angle O_1QP) - d_2\cos(\angle O_2QP) = 0.$$

There are now two scenarios to consider, depending upon the position of P relative two O_1 and O_2. If P is between O_1 and O_2, then $\angle O_1QP$ and $\angle O_2QP$ are supplementary, so

$$\cos(\angle O_2QP) = -\cos(\angle O_1QP).$$



The picture when Q is not between circle centers.

Then

$$\begin{aligned} &d_1 \cos(\angle O_1QP) - d_2 \cos(\angle O_2QP) \\ &\qquad = d_1 \cos(\angle O_1QP) + d_2 \cos(\angle O_1QP) \\ &\qquad = (d_1 + d_2) \cos(\angle O_1QP). \end{aligned}$$

Since $d_1 + d_2 > 0$, the only way this can be zero is if $\cos(\angle O_1QP) = 0$; that is, $(\angle O_1QP) = \pi/2$. If P is not between O_1 and O_2, then $\angle O_1QP$ and $\angle O_2QP$ are equal, so

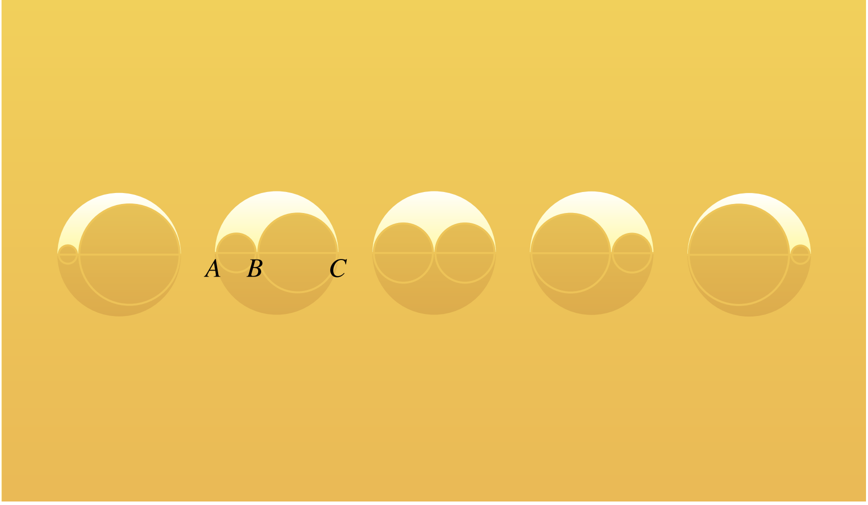
$$\begin{aligned} &d_1 \cos(\angle O_1QP) - d_2 \cos(\angle O_2QP) \\ &\qquad = d_1 \cos(\angle O_1QP) - d_2 \cos(\angle O_1QP) \\ &\qquad = (d_1 - d_2) \cos(\angle O_1QP). \end{aligned}$$

In this case, d_1 and d_2 cannot be equal because then the circles would be concentric. Therefore $d_1 - d_2 \neq 0$, so again $\cos(\angle O_1QP) = 0$; that is, $(\angle O_1QP) = \pi/2$. Either way, then, the angle at Q is a right angle. That places P on ℓ. □

That wraps up the preliminaries. Orthogonal circles will play a key role in our model for non-Euclidean geometry, but for now it is on to the theorems.

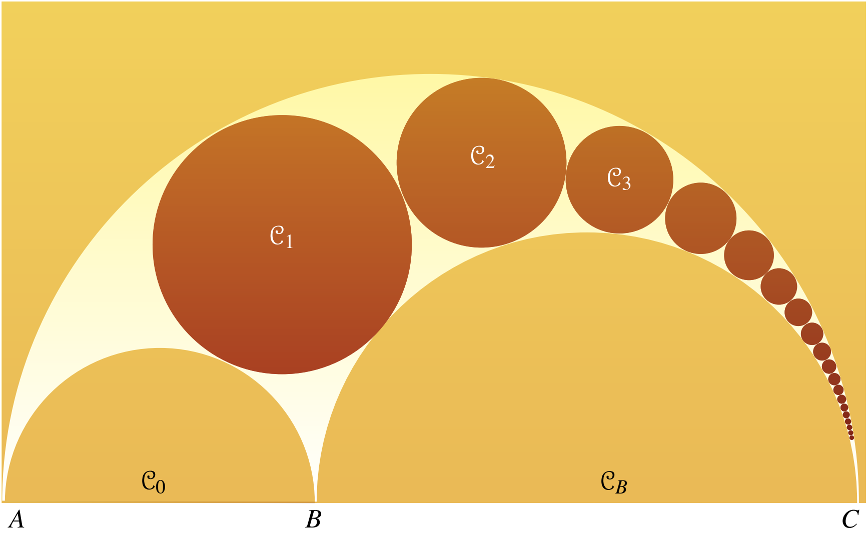
The arbelos

An *arbelos* is a shape built from three semicircles. Let me give the construction. Start with a semicircle with radius AC. Locate a point B anywhere between A and C. Now form two more semicircles on the same side of AC, with diameters AB and BC respectively. The resulting shape is called an *arbelos*.



Five arbeloses (arbeli?)

As with triangles, it is pretty common to use the term to mean either the edges (the semicircles themselves) or the interior region bounded by them. This shape has a long history in classical geometry, going all the way back to the ancient Greeks. The name comes from them: apparently arbelos is a Greek word for a particular type of knife that was used by shoemakers. Its curved blade resembled the geometric shape that now bears its name. A lot of interesting relationships have been found inside the arbelos, mainly in the form of hidden tangent circles. Our next theorem is in that vein– it presents a chain of mutually tangent circles.



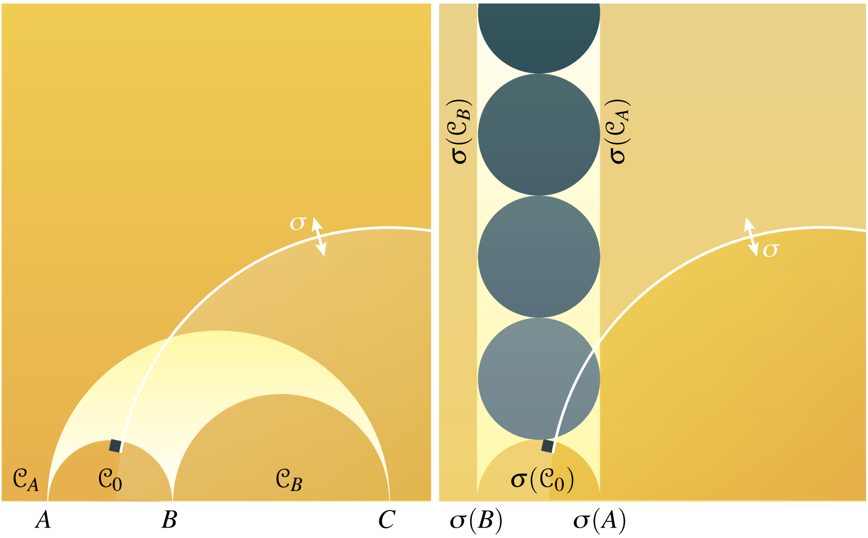
THM: A CHAIN OF CIRCLES IN THE ARBELOS

Given an arbelos formed from three circles, with diameters AC, AB, and BC, where $A * B * C$, label

$\mathcal{C}_A$: the semicircle with radius AC
$\mathcal{C}_B$: the semicircle with radius BC
$\mathcal{C}_0$: the semicircle with radius AB

Then there is a circle $\mathcal{C}_1$ that is tangent to each of $\mathcal{C}_A$, $\mathcal{C}_B$, and $\mathcal{C}_0$; there is a circle C_2 that is tangent to $\mathcal{C}_A$, $\mathcal{C}_B$, and $\mathcal{C}_1$; there is a circle $\mathcal{C}_3$ that is tangent to $\mathcal{C}_A$, $\mathcal{C}_B$, and $\mathcal{C}_2$; and in general, for any $n \geq 1$, there is a circle $\mathcal{C}_n$ that is tangent to $\mathcal{C}_A$, $\mathcal{C}_B$, and $\mathcal{C}_{n-1}$.

Proof. This proof is easy– once you have the right inversion. Recall the previous discussion of orthogonal circles: since C is outside the circle $\mathcal{C}_0$, there must be a circle centered at C which is orthogonal to $\mathcal{C}_0$. Let σ be the inversion across that circle. Then $\sigma(\mathcal{C}_0)$ is C_0 (but note that σ interchanges A and B). Both $\mathcal{C}_A$ and $\mathcal{C}_B$ pass through C, so σ inverts them to lines: $\sigma(\mathcal{C}_A)$ is the line through B which is perpendicular to AB, and $\sigma(\mathcal{C}_B)$ is the line through A which is perpendicular to AB.

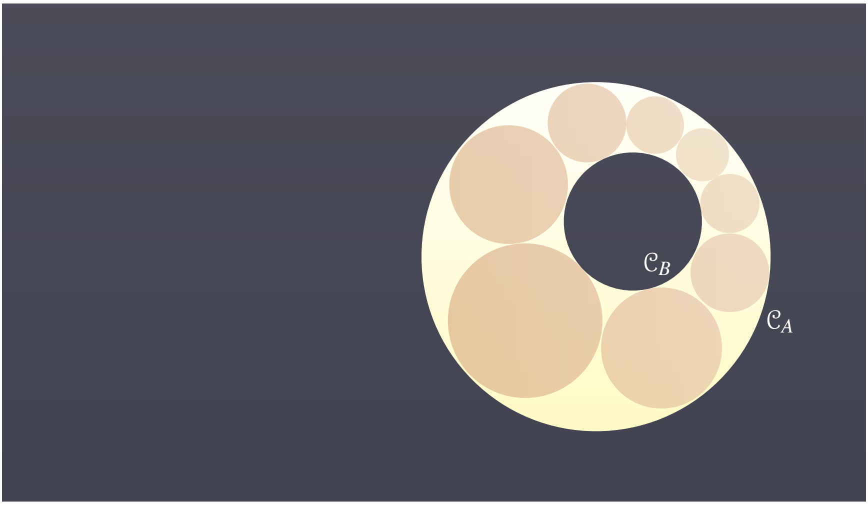


Now $\mathcal{C}_0$ is mutually tangent to two parallel lines, and in this more apparently symmetric configuration, it is easy to build a stack of circles on top of $\sigma(\mathcal{C}_0)$, each tangent to the lines $\sigma(\mathcal{C}_A)$ and $\sigma(\mathcal{C}_B)$ and the circle immediately below it. Apply σ again, sending this stack tumbling down: $\sigma(\mathcal{C}_A)$ back to $\mathcal{C}_A$, $\sigma(\mathcal{C}_B)$ back to $\mathcal{C}_B$, $\mathcal{C}_0$ staying put, but the circles stacked on top of it mapping to $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3 \ldots$. □

Steiner's porism

Jakob Steiner was a nineteenth century geometer with a particular interest in inversion and a particular disdain for analytic geometry. This next chain of circles is named in his honor. Start with two circles $\mathcal{C}_A$ and $\mathcal{C}_B$, with $\mathcal{C}_B$ contained entirely in the interior of $\mathcal{C}_A$. Now we will build a chain of mutually tangent circles between $\mathcal{C}_A$ and $\mathcal{C}_B$. First choose a circle $\mathcal{C}_1$ which is tangent to both $\mathcal{C}_A$ and $\mathcal{C}_B$. Then:

- There is a circle $\mathcal{C}_2$ which is tangent to $\mathcal{C}_A$, $\mathcal{C}_B$, and $\mathcal{C}_1$.
- There is a circle $\mathcal{C}_3$ which is tangent to $\mathcal{C}_A$, $\mathcal{C}_B$, and $\mathcal{C}_2$.
- In general, there is a circle which is tangent to $\mathcal{C}_A$, $\mathcal{C}_B$, and $\mathcal{C}_{n-1}$.



A Steiner chain between two circles.

This is very similar to the chain of circles constructed in the arbelos. The difference this time is that eventually the circles inside this Steiner chain will loop around back to $\mathcal{C}_1$. The natural question to ask is: when the chain does get back around to $\mathcal{C}_1$, will it join up perfectly? Will the last circle in the chain be tangent to $\mathcal{C}_1$? That sure would be nice, but in general it does *not* happen. However, if conditions are so that the chain does close perfectly, then this will happen no matter what circle $\mathcal{C}_1$ you use as the starting point. In other words, whether the chain closes perfectly depends only on $\mathcal{C}_A$ and $\mathcal{C}_B$, not on $\mathcal{C}_1$. In the course of this discussion, I have made two claims that need proof.

THM: STEINER'S PORISM

Suppose circle $\mathcal{C}_B$ is contained in the interior of $\mathcal{C}_A$, and that a third circle $\mathcal{C}_1$ in $\mathcal{C}_A$ is tangent to both $\mathcal{C}_A$ and $\mathcal{C}_B$.

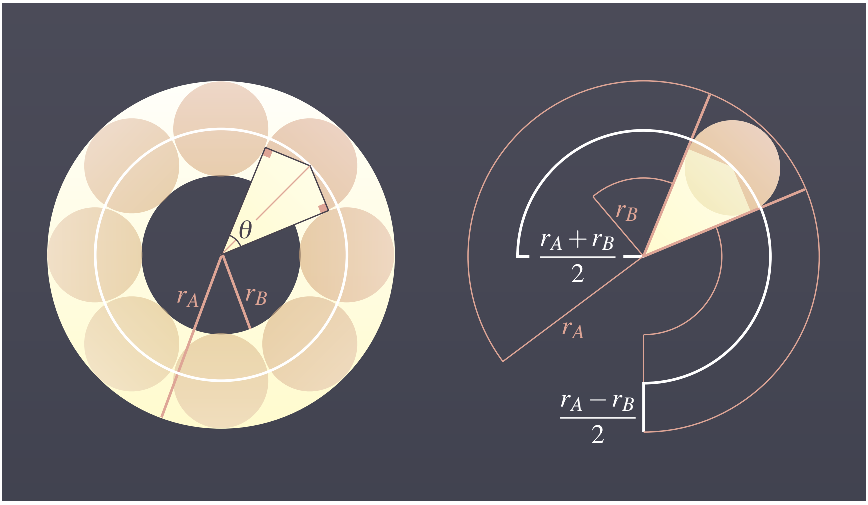
1. There is a chain of circles $\mathcal{C}_2, \mathcal{C}_3, \ldots$, where each C_n is tangent to $\mathcal{C}_A$, $\mathcal{C}_B$, and $\mathcal{C}_{n-1}$.

2. If $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N\}$ is such a chain of circles and $\mathcal{C}_N$ is tangent to $\mathcal{C}_1$, then for any such chain of circles $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N\}$, $\mathcal{D}_N$ will be tangent to $\mathcal{D}_1$.

Proof. There is one scenario where these statements are fairly obvious– when $\mathcal{C}_A$ and $\mathcal{C}_B$ are concentric. In that case, let O be the mutual centers of the circles, and let r_A and r_B be the respective radii. Then each circle in the chain $\mathcal{C}_i$ has its center on a circle halfway between $\mathcal{C}_A$ and $\mathcal{C}_B$– on the circle with center O and radius $(r_A + r_B)/2$ to be precise. All of the circles in the chain are the same size– they all have radii of $(r_A - r_B)/2$. Regarding the question of whether the chain will close up neatly, look at the angle θ at O that any one of these circles subtends. If you slice that angle in half, there is a right triangle in there– one with an opposite side of $(r_A - r_B)/2$ and a hypotenuse of $(r_A + r_B)/2$. Therefore

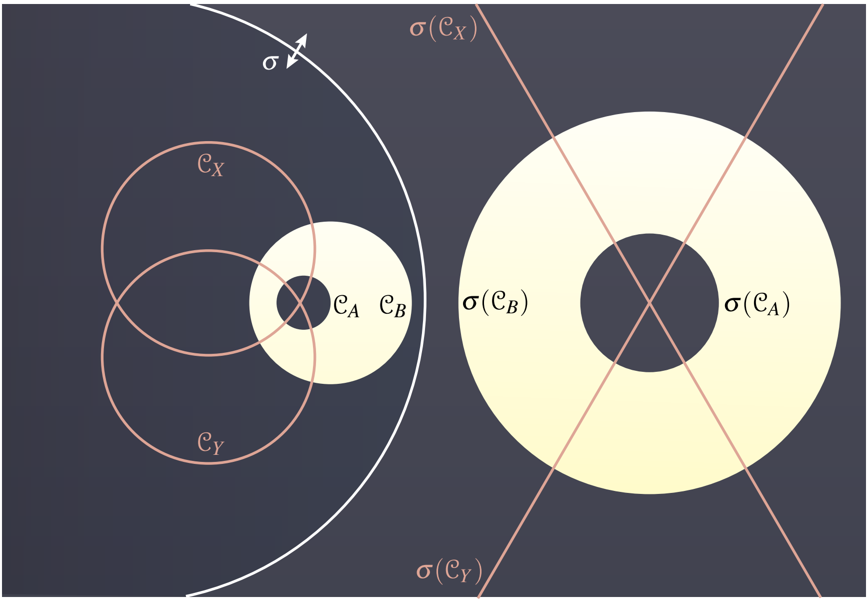
$$\sin(\theta/2) = \frac{r_A - r_B}{r_A + r_B} \implies \theta = 2\sin^{-1}\left(\frac{r_A - r_B}{r_A + r_B}\right).$$

Then the question of whether the chain closes up is just: is 2π divisible by θ? And of course the answer to this question depends only on r_A and r_B, not on where the starting circle $\mathcal{C}_1$ is.



The concentric case.

If $\mathcal{C}_A$ and $\mathcal{C}_B$ are not concentric– that is they have distinct centers O_A and O_B– then the circles in the chain will not all be the same size, as they are forced to adjust to fit in between $\mathcal{C}_A$ and $\mathcal{C}_B$. That makes both claims a bit more difficult. Fortunately, there is an inversion out there to help– an inversion that maps $\mathcal{C}_A$ and $\mathcal{C}_B$ to a concentric configuration.



Inversion turns the general configuration into a concentric one.

The right inversion is an inversion across a circle whose center is the intersection of two circles C_X and C_Y which are both orthogonal to $\mathcal{C}_A$ and $\mathcal{C}_B$. First of all, we know that there are circles that are orthogonal to both $\mathcal{C}_A$ and $\mathcal{C}_B$. They are all centered along the radical axis. Of course, not any two of those circles will intersect each other, so you will need to choose carefully. As long as you choose two circles $\mathcal{C}_X$ and $\mathcal{C}_Y$ whose centers on the radical axis are the same small distance ε on either side of O_AO_B, then $\mathcal{C}_X$ and $\mathcal{C}_Y$ will intersect. I will leave the details of the precise placement of these two circles as an exercise. Once you have them, let σ be an inversion across a circle centered at the intersection of $\mathcal{C}_X$ and $\mathcal{C}_Y$. Then $\sigma(\mathcal{C}_X)$ and $\sigma(\mathcal{C}_Y)$ are intersecting lines, and $\sigma(\mathcal{C}_A)$ and $\sigma(\mathcal{C}_B)$ are circles perpendicular to $\sigma(\mathcal{C}_X)$ and $\sigma(\mathcal{C}_Y)$. That means both $\sigma(\mathcal{C}_X)$ and $\sigma(\mathcal{C}_Y)$ run through the centers of $\sigma(\mathcal{C}_A)$ and $\sigma(\mathcal{C}_B)$. Of course, there is only one point on both $\sigma(\mathcal{C}_X)$and $\sigma(\mathcal{C}_Y)$. That point has to be the center of both $\sigma(\mathcal{C}_A)$ and $\sigma(\mathcal{C}_B)$. Therefore, $\sigma(\mathcal{C}_A)$ and $\sigma(\mathcal{C}_B)$ are concentric. We know we can construct a chain of circles between $\sigma(\mathcal{C}_A)$ and $\sigma(\mathcal{C}_B)$, and that whether it closes or not depends only on $\sigma(\mathcal{C}_A)$ and $\sigma(\mathcal{C}_B)$. Apply σ again to a chain of circles between $\sigma(\mathcal{C}_A)$ and $\sigma(\mathcal{C}_B)$– the result is a chain of circles between $\mathcal{C}_A$ and $\mathcal{C}_B$, and whether that chain closes up neatly does not depend upon the location of the first circle in the chain. □

Exercises

1. Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be two non-concentric circles. Prove that if the circles intersect, then their radical axis passes through the intersection point(s). Prove that if the circles don't intersect, then their radical axis lies outside both of them.

2. Given $A * B * C$, form an arbelos by removing half-circles with diameters AB and BC from the half-circle with diameter AC. Label these half-circles as $\mathcal{C}_0$, $\mathcal{C}_B$, and $\mathcal{C}_A$ (as in the proof of the arbelos chain). The line ℓ which passes through B and is perpendicular to AC intersects $\mathcal{C}_A$ at a point D. The circle with radius BD intersects both $\mathcal{C}_0$ and $\mathcal{C}_B$ one more time– label those points E and F. Prove that the quadrilateral $BEDF$ is a rectangle.

3. The proof of Steiner's porism uses a pair of intersecting orthogonal circles. We had previously proved that those orthogonal circles do exist– their centers are on the radical axis Prove that it is always possible to find a pair of such orthogonal circles that *do* intersect.

4. Let σ be the inversion across circle $\mathcal{C}$ with center O, and let P be a point other than O. Find a compass and straight edge construction of $\sigma(P)$. Hint: the distance from O to $\sigma(P)$ is the key, and that is governed by the equation $|O\sigma(P)| = r^2/|OP|$. To get that kind of ratio, consider a configuration of right triangles where $\triangle ABC$ has right angle at C, and D is the foot of the altitude from C.

5. Given a circle $\mathcal{C}$ and two points P and Q in the interior of $\mathcal{C}$ which are not on the same diameter, give a compass and straight edge construction of the circle which passes through both P and Q, and is orthogonal to $\mathcal{C}$.