Inversion

In this lecture we will learn about a remarkable transformation of the plane called inversion. We will see that inversion can be extremely helpful in problems where many circles are involved - it often enables one to simplify such problems considerably.

0.1 Definition

Inversion should be thought of as a kind of reflection in a circular mirror. This isn't precise, but this analogy can give you a good feeling of how inversion behaves.

Before we define what inversion is precisely, we will define the notion of symmetry with respect to a circle.

Let C be a circle with center O and radius R. We will call points P and P' symmetric with respect to the circle C if

1. the point P' lies on the ray from O to P.

2. $OP \cdot OP' = R^2$.

Note that for every point P in the plane, except for the center O, there exists a unique point P' symmetric to it with respect to the circle C.

We will find it convenient when dealing with inversions, to compactify our plane by one point at infinity, called ∞. One should think of this point not as some object unrelated to the plane, but rather as a point in a new space, "the extended plane", to which all the sequences of points whose absolute values tend to infinity converge.

With this additional point in mind we see that we should define the points O and ∞ to be symmetric with respect to the circle C. Indeed, it follows from the definition that a sequence of points converges to O if and only if the absolute values of the symmetric points tend to infinity.

This supplement to the definition of symmetric points allows us to claim that for every point P in the extended plane there is a point symmetric to it with respect to the circle P and such a point is unique.

Now we can define what inversion is: inversion with respect to a circle C is the transformation of the extended plane to itself that sends every point to the point symmetric to it with respect to the circle C.

For example every point on the circle C gets inverted to itself; the interior and the exterior of the circle C are switched by the inversion.

0.2 Properties

One can see immediately that inversion is an involution, i.e. inverting twice is the same as doing nothing at all.

Another simple property is that lines passing through the point O get inverted to themselves (in fact every ray with origin O gets inverted to itself).

A less immediate property of the inversion is that lines that do not pass through the point O get inverted to circles that do pass through O (and vice versa - remember that inversion is an involution).

To prove it let l be a line in the plane. Let P denote the foot of the perpendicular from the point O to the line l. Let P' denote its image under inversion in a circle with center O. We claim then that the image of the line l under inversion is the circle l' for which the segment OP' is the diameter.

Inversion

What we have to prove is that any point Q on the line l gets mapped after inversion to some point on the circle l'. Let Q' denote the image of Q under inversion, i.e. the point on the ray OQ for which $OQ \cdot OQ' = OP \cdot OP'$.We can rewrite this equality in the form $\frac{OQ}{OP'} = \frac{OP}{OQ'}$. This, together with the fact that the angle $\angle QOP = \angle P'OQ'$ is a common angle for triangles QOP and $P'OQ'$, implies that these triangles are similar. In particular $\angle OQ'P' = \angle OPQ = 90°$, hence the point Q' lies on the circle for which OP' is the diameter.

With this knowledge at hand we can prove a remarkable property of inversion: it preserves angles!!!

What we mean by this is that if two curves intersect at some point P and form angle α at the point of intersection, then their images under inversion will intersect at point P' - the inverted image of P - and form an angle α at this point (more precisely, $-\alpha$ - the magnitude of the angle will be the

same, but if we keep track of orientations and the order of the curves, then the direction of the angle changes).

First we observe it's enough to prove the claim for the case where both curves are lines - this is so because the angle between two curves at their point of intersection is defined as the angle between the lines tangent to these curves at the point of intersection.

Now if we have two lines in the plane and the angle between them is α, then we can obtain one from the other by rotating the first one by angle α around the point O (this way it becomes parallel to the second one) and then scaling with the same center O by some factor a (equal to the quotient of the distances from O to the lines).

This means that the image under inversion of the second line can be obtained from the image of the first one by rotating around the point O by angle α and then scaling by factor $1/a$.

But we know that the images of the lines under inversion are circles that pass through the point O. From what we described above we can conclude that the angle between the circles at the point O is equal to α (the angle at O between the first circle and its copy rotated by angle α is clearly α; this angle is not affected if we further scale the rotated copy - the tangent line at the point O remains the same). But for circles, the two angles at the two points of intersection are equal up to sign! Since the angle at the other point of intersection is the one we are interested in, we are done.

Next property we of inversion we would like to discuss is the fact is that it maps any circle that doesn't pass through the point O to a circle (which, of course, doesn't pass through the point O).

Before we prove it, let's note that while what the inversion in circle C does to the plane depends very much on the center of circle C, it depends much less on its radius. Namely if C_1 and C_2 are circles with the same center O, but with different radii R_1 and R_2 and S is any shape in the plane, then the inverted images S_1' and S_2' of S in circles C_1 and C_2 are related by similarity relation: S_2' is obtained from S_1' by scaling with center O and factor R_2^2/R_1^2. This follows by checking that for any point P in S the distances OP_1' and OP_2' from O to the inverted images of P in C_1 and C_2 are related by $OP_1'0/OP_2' = R_1^2/R_2^2$.

Now we can approach the problem of proving that circles invert to circles by first choosing a convenient radius of inversion, and then proving the statement for it. For any other radius of inversion the claim will follow.

We separate two cases of the statement: when the center of inversion O

lies outside the circle we are inverting and when it is inside.

In the first case we can choose the radius of the inversion circle to be equal to the length of the tangent lines from O to the circle we are inverting. With this choice the circle of inversion becomes perpendicular to the circle we are inverting (because tangent line to any circle is perpendicular to the radius of this circle that passes through the point of tangency).

We will show that in this case, i.e. if the circle s we are inverting is perpendicular to the circle C, its inversion in C coincides with itself.

For the proof let P be one of the points of intersection of the circle s and the inversion circle C and let O be the center of C. Since the circle s is perpendicular to C and the radius OP is also perpendicular to C, the segment OP is tangent to s at P. Now if l is any ray with origin at O and Q and Q' are its intersection points with s, then by a familiar [1] property of circles $OQ \cdot OQ' = OP^2$. Thus Q' is the inversion of Q. Since the ray l was arbitrary, we proved that the image of s under inversion is s itself.

For circle s_1 that does contain the inversion center O denote by AB its diameter that contains the point O. We will choose the radius R for the inversion circle to be such that $R^2 = OA \cdot OB$. We claim that the inversion of s_1 in the circle C with center O and such radius is the circle s_2 obtained from s_1 by rotating by 180° around point O. To check this, let P_1 and P_2 denote two points on circles s_1 and s_2 respectively lying on a ray l with origin at O. We want to prove $OP_1 \cdot OP_2 = R^2$. Let P_2' denote the other point of intersection of the line containing l and the circle s_1. By definition of the circle s_2, the point P_2' can be obtained from the point P_2 by rotating by 180° around the point O. Hence $OP_2 = OP_2'$. From general properties of circles we know that $OP_1 \cdot OP_2' = OA \cdot OB$, hence indeed $OP_1 \cdot OP_2 = R^2$.

Note that we have proved slightly more: the inverted image of the a circle not containing the inversion center is a circle homothetic to the original one with positive coefficient of homothety (i.e. one can scale the original circle by a positive factor to get its inverted image). Indeed, it is obvious from what we proved that this is true if the radius of inversion circle is equal to the length of the tangents from O to the circle we are inverting (in this case the homothety factor should be 1). If we change the radius of the inversion circle, the inverted image of any shape gets scaled by a positive factor, which proves our claim.

In a similar fashion we can show that the inverted image of a circle that

[1] refer to appendix

contains the point O is a circle homothetic to the original one with negative homothety factor, i.e. can be obtained by scaling by a positive factor and rotating by 180°.

0.3 How inversions can be useful?

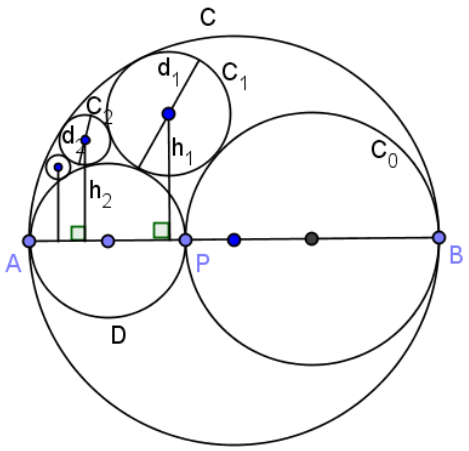
We've introduced a rather funny transformation of the plane to itself, the inversion, and proved some of its properties. Now we ask the question the reader must have asked: why? What does the inversion allow us to do that we couldn't do before?

We can't answer this question by some philosophical remark of the sort that inversions are as natural as affine transformations of the plane (wait for the next chapter though), so instead we will just introduce some problems that look extremely difficult, but become quite easy when one decides to use inversion.

The first on the list is an ancient theorem by Pappus.

Add something about who Pappus was - the last great Greek geometer etc.

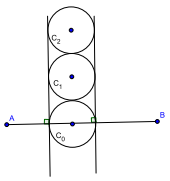
Start with a circle C. Let AB be its diameter and P - any point on it. Let D denote the circle whose diameter is AP. Let C_0 denote the circle whose diameter is PB. Let C_1 be a circle tangent to C from the inside and tangent to D and C_0 from the outside (there are two such circles, they are symmetric with respect to AB; choose the one lying in the upper half circle of C). Let C_2 be the circle tangent to C from the inside and tangent to D and C_1 from the outside; there are two such circles - choose the one that is closer to the point A. Continue in the same fashion to build C_3,C_4,... (C_{n+1} is tangent to C, D and C_n). Let h_n denote the distance from the center of circle C_n to the line AB and let d_n denote the diameter of C_n. Then $\frac{h_n}{d_n} = n$ for every n.



This result is amazing, because it doesn't depend on the point P we choose on the diameter AB at all! Moreover, it is quite difficult - it took Pappus years to prove it. Even for the first step in the construction, the circle C_1, it isn't obvious that its diameter is equal to the distance from its center to AB. For other circles the statement seems even harder!

We will now prove it quite easily using inversion.

Invert all the picture in a circle with center at point A. The line AB must get inverted to itself, since it was a line passing through the center of inversion. The circles C and D both passed through point A and didn't intersect at any other point, so they must get inverted to a pair of parallel lines. Circles C and D were orthogonal to the line AB, hence their images, the pair of parallel lines, must be orthogonal to AB. The circle C_0 must get mapped to a circle tangent to both these parallel lines and orthogonal to the line AB, hence the center of its image must lie on AB. All other circles C_i must also get mapped to circles tangent to the pair of parallel lines. In particular they all must have the same diameter, equal to the distance between the two lines.



Now notice that the inverse image of any circle not containing the center of inversion is homothetic to the original circle. In particular in our picture the quotient $\frac{h_n}{d_n}$ of the height of the center of C_n above AB to its diameter is equal to the qoutient $\frac{h'_n}{d'_n}$ of the height of the center of the image of C_n under inversion above line AB to its diameter. But in the picture after inversion it is clear that the height h'_n is equal to n times the diameter of all circles, thus proving our claim.

Exercise:

Let C be a circle with center O and radius R. Let s be a circle with radius r and center O'. Let d denote the distance between the centers O and O'. Suppose also $d > r$ so that the circle s doesn't contain O. Prove that the radius of the circle s' which is obtained from circle s by inverting in circle circle C is $R^2r/(d^2 - r^2)$.

Hint: Let AB denote the diameter of the circle s, whose extension contains the point O. Let A' and B' denote images of A and B under inversion in circle C. Prove that $A'B'$ is the diameter of s' and that $OA = d + r$, $OB = d - r$, $OA' = R^2/(d + r)$, $OB' = R^2/(d - r)$.

Exercise:

Let C denote a circle of radius 1. Let O denote its center and AB be its diameter. Let D denote the circle whose diameter is AO. Let C_0 denote the circle whose diameter is OB. Let C_1 be a circle tangent to C from the inside and tangent to D and C_0 from the outside (there are two such circles, they are symmetric with respect to AB; choose one). Let C_2 be the circle tangent to C from the inside and tangent to D and C_1 from the outside; there are two such circles - choose the one that is closer to the point A. Continue in

the same fashion to build C_3,C_4,... (C_{n+1} is tangent to C, D and C_n). Prove that the radius of C_n is $\frac{1}{n^2+2}$.

Hint: when we invert this picture in a circle with radius R and center A, the circles C and D become parallel lines orthogonal to the diameter AB and distand $R^2/2$ apart. The circles C_n invert to a chain of circles of radius $R^2/4$ tangent to both parallel lines and to each other. Show that the distance d_n to the center of inverted image of C_n satisfies $d_n^2 = (n^2 + 9/16)R^2$ and use the previous exercise to derive the desired result.

Another problem we can solve rather easily using inversion is a Porism of Steiner.

The porism claims the following.

Let C and D be two disjoint circles and for simplicity assume D is inside C. Let C_0 be any circle between them tangent to both. Construct a chain of circles as follows: circle C_1 is tangent to C, D and C_0, circle C_2 is tangent to C,D and C_1 and so on. We claim that if this chain closes after n steps, i.e. if $C_n = C_0$, then for any choice of initial circle C_0 that is tangent to both C and D the chain will close after n steps.

The proof using inversion is extremely simple. Here it is:

For concentric circles C and D the claim is obvious. If C and D are not concentric, choose an inversion that maps them to a pair of concentric circles. Since tangencies and circles are preserved by inversions, we are done!

The only non-obvious step in this argument is the fact that we can choose an inversion that maps any given pair of disjoint circles to a pair of concentric circles.

This will follow from the following lemma: for a pair of disjoint circles C and D there exists a pair of circles r, s intersecting at two points that are orthogonal to both C and D. Indeed, if r and s are two such circles and O and P ar their points of intersection, then after inversion with center O the circles r and s will get mapped to two lines intersecting at point P' (the image of P under inversion). The circles C and D must get mapped to circles orthogonal to both lines r' and s'. Now a center of a circle orthogonal to line r' lies on r'. Similarly for s'. So if C and D are orthogonal to both lines r' and s', then their centers must belong to both lines r' and s'. Hence their centers are located at P', i.e. C' and D' are concentric.

So it remains to prove the lemma: for a pair C, D of disjoint circles one can find a pair r, s of circles intersecting at two points othogonal to both C and D. How can we prove such a claim? Of course by inversion!!! Since orthogonality is preserved by inversions, proving the lemma for C and D is

equivalent to proving it for their images under some inversion.

Choose now the center of inversion on the circle C. Then after the inversion the circle C becomes a line. So it is enough to prove the lemma in case one of the circles is a line.

So suppose that C is a line and D is a circle disjoint from it. Choose r to be the line passing through center of D and orthogonal to the line C. Clearly r is orthogonal to both C and D. It remains to find another circle s having the same property. Let O be the point of intersection of r and C. Let s denote the circle whose center is at O and whose radius is equal to the length of the tangent segments from O to D. This circle s is orthogonal to the circle D, because of the choice of the radius (add reference to the general picture of orthogonal circles - you use it more than twice!) and it is orthogonal to the line C because its center lies on C. It clearly intersects the line r (since r passes through the center of s), and so we are done.

0.4 Inversion in $\mathbb{R}^n$

In this section we generalize the main properties of inversion to the Euclidean space $\mathbb{R}^n$. We will use a little bit of analytic geometry.

Choose a coordinate system in $\mathbb{R}^n$ so that the origin coincides with the center of inversion. Then the inversion with center at the origin and radius R sends the point $x \in \mathbb{R}^n$ to $\frac{x}{|x|^2}$.

How do planes look in $\mathbb{R}^n$? They are all of the form $\{x| < x, y >= c\}$ for some vector $y \in \mathbb{R}^n$ (a normal vector) and $c \in R$. Similarly a sphere with center x_0 and radius r is given by the equation $\{x| < x - x_0, x - x_0 >= r^2\}$. We can rewrite this equation in the form $\{x| < x, x > +a < x, y > +b = 0\}$ (where y is $-2x_0$ and b is $< x_0, x_0 > -r^2$).

Now the inversion of a plane $\{x| < x, y >= c\}$ is given by the equation $< \frac{x}{|x|^2}, y >= c$, or $< x, y >= c < x, x >$. If $c = 0$, then this is an equation of the plane $< x, y >= 0$. If $c \neq 0$, it is an equation of the sphere ($< x, x > -c^{-1} < x, y >= 0$). The origin 0 clearly belongs to this sphere. Hence planes passing through the origin get inverted to themselves, and those not passing through the origin get inverted into spheres passing through origin.

Now let's figure out what happens to the spheres under inversion. As in the case of the plane R^2, inversion is an involution - doing it twice doesn't change anything. Hence we already showed that the spheres that pass through the origin should get inverted to planes not passing through

the origin. So let $< x, x > + a < x, y > + b = 0$ be an equation of a sphere not passing through the point 0: this translates to $b \neq 0$. Then the points on its image under inversion satisfy $< \frac{x}{|x|^2}, \frac{x}{|x|^2} > + a < \frac{x}{|x|^2}, y > + b = 0$, or, multiplying by $< x, x >$

$$1 + a < x, y > + b < x, x >= 0$$

Since $b \neq 0$, we can rewrite it in the form $< x, x > + \frac{a}{b} < x, y > + \frac{1}{b} = 0$, which is an equation of a sphere. Thus we have proved that the spheres not passing through the center of inversion get inverted into spheres not passing through the center of inversion.

It's easy to deduce from this that circles get mapped under inversion to circles. Indeed, a circle is an intersection of several planes and a sphere. Under inversion it must get mapped to the intersection of images of these planes and spheres. But any intersection of spheres and planes is a circle (provided it is one-dimensional, which it is, since we started with a one-dimensional circle).

The only property of inversion we didn't generalize to $\mathbb{R}^n$ yet is the property that inversions preserve angles, but reverse their orientations. To see what a map does to angles, it is enough to consider only its linear approximation, its differential. If this linear approximation preserves angles, the map itself preserves them. (This is so more or less by definition of what angles between curves are: angle at a point of intersection of two curves is defined to be the angle between the tangent lines to these curves at the point of intersection. Now the differential of a map does exactly what we need here: it takes the tangent line to a curve to the tangent line to its image. Thus the map preserves angles if and only if its differential does).

Now how can we compute the linear approximation to the map $x \to \frac{x}{|x|^2}$? For points x near the unit sphere, i.e. such that $|x| = 1 + \epsilon$ for some small ϵ, the length of their image under inversion is $\frac{1}{|x|} = \frac{1}{1+\epsilon} = 1 - \epsilon + \epsilon^2 - \epsilon^3 + \ldots \approx 1 - \epsilon$. Thus near a point on the unit sphere the inversion behaves just like reflection in the tangent space at this point.

See figure ...

Since reflections preserve angles, reversing their orientations, we proved that inversion preserves angles, reversing their orientations, at any point on the sphere of inversion.

For the proof of this fact for other points, recall that inversion with two different radii, but the same center, differ only by scaling: inverting in a

sphere of radius R and then scaling by $1/R^2$ is the same as inverting in a sphere of radius 1 with the same center. So to prove that inversion in the unit sphere preserves angles at a point x_0 we first use the fact that the inversion with radius $|x_0|$ preserves angles at x_0 (this is what we proved above) and then that scaling by $1/|x_0|^2$ preserves all angles. Since the composition is equal to inversion in the unit sphere, we are done.

0.4.1 Stereographic projection

In complex analysis and sometimes in other circumstances one encounters the following map from a sphere S to plane π tangent to it at the "south pole". To find the image of a point x on the sphere under this map, one has to connect the point x to the north pole by a straight line and find the point of intersection of this line with the plane π. This point of intersection is the stereographic projection of x.

This projection has some wonderful properties, which make it very useful. First property is that it maps circles on the sphere S to circles on the plane π. The other useful property is that it preserves angles.

One can't help but notice that these properties resemble very much the properties of an inversion. And indeed, stereographic projection IS the restriction of the inversion in a sphere with center at the north pole and radius equal to the diameter of the sphere S from which we are projecting. Indeed, such an inversion maps the sphere S to a plane, passing through the south pole of S and orthogonal to the segment connecting the North pole with the south pole. So it maps S to π. Also the image of every point x on the sphere S under the described inversion lies on the ray starting at the north pole and passing through x. Thus the image of x under inversion must coincide with its image under stereographic projection (since the point of intersection of the plane π with this ray is unique).

Once we identified the stereographic projection with inversion, its properties become clear: it maps circles on the sphere to circles (or lines) on the plane and it preserves angles.

0.5 Appendix

In this appendix we will prove some of the statements from elementary geometry of the circle we used in proving the porperties of inversion in the

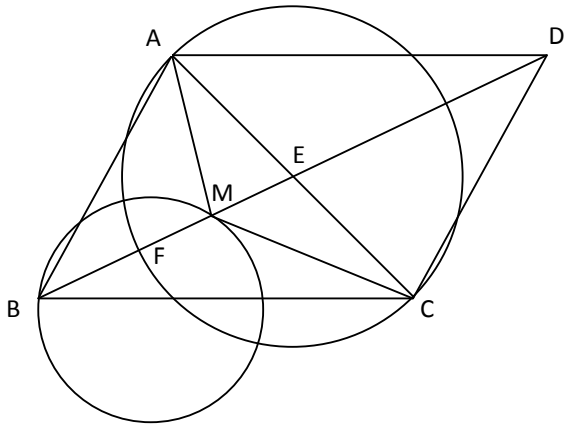
plane.

Theorem 1. *Let C be a circle with center O and let X and Y be points on Z be three points on C. Then $\angle XYZ = \frac{1}{2}\angle XOZ$. In particular the angle $\angle XYZ$ is independent of the choice of point Y on the arc XZ.*

Theorem 2. *Suppose lines l_1 and l_2 interesect at a point Z. Suppose also that the line l_1 intersects the circle C at points X_1, Y_1 and the line l_2 intersects the circle in points X_2, Y_2. Then $|ZX_1||ZY_1| = |ZX_2||ZY_2|$.*

Some solved exercises by Chang Yu and Siming Zhao

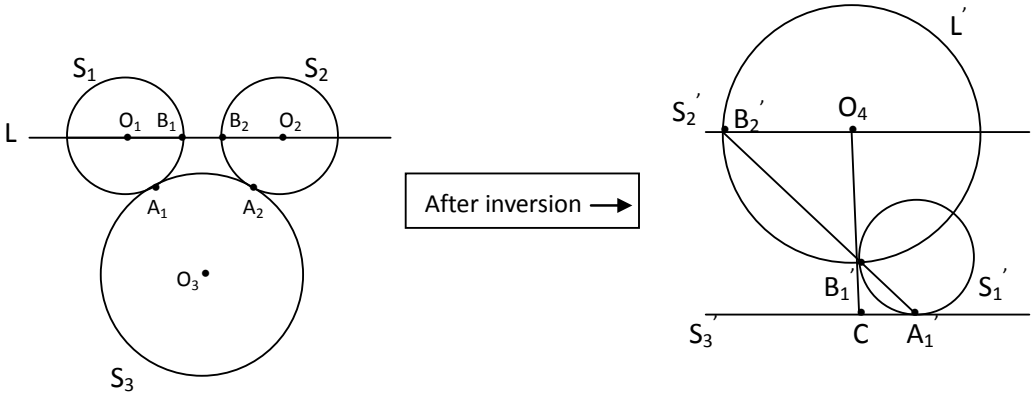
Let ABCD be a parallelogram. Let point M be the point on the diagonal BD such that such that $\angle AMC = \angle BAD$. Suppose also that the point M lies on the segment between the point B and the point of intersection of diagonals of ABCD. Prove that the circle for which AC is the diameter is orthogonal to all circles passing through points B and M.



Claim: If points B and M are inverse points respect to the inversion circle centered at E with diameter AC, Then the circle for which AC is the diameter is orthogonal to all circles passing through points B and M.
In short, we want to prove $EM \bullet EB = AE^2$

$\because \angle BAD = \angle AMC$, $\angle BAD + \angle ADC = 180^\circ$
$\therefore \angle AMC = \angle ADC = 180^\circ$
$\because$ since points A, M, C, D forms a quadrilateral and the opposite angels adds to 180°
$\therefore$ points A, M, C, and D lie on one same circle
$\therefore \angle DAC = \angle DMC$ since they share the same arc in a circle
$\therefore \angle BAE = \angle AME$
$\because \angle AEB = \angle MEA$, $\angle BAE = \angle AME$
$\therefore \angle ABE = \angle MAE$
$\therefore \triangle ABE$ similar to $\triangle MAE$
$\therefore \frac{AE}{ME} = \frac{BE}{AE}$
$\therefore EM \bullet EB = AE^2$ □

Let circles S_1 and S_2 with centers at O_1 and O_2 to be disjoint. Let circle S_3 be tangent externally to S_1 and S_2 at points A_1 and A_2 respectively. Let B_1 and B_2 be the points of intersection of the segment $O_1\ O_2$ with the circles S_1 and S_2 respectively. Show that A_1, A_2, B_1, B_2 all lie on the same circle.



Note: all inverse points after inversion are donated by the same symbol in superscript form.

Claim: If we can show A_1', A_2', B_1', B_2' lie on one line after inversion, then it means those 4 points lie on the same circle before inversion.

Fix point A_2 as inversion center, then circle S_2 and S_3 will become two parallel lines after inversion, and the tangent point A_2 will go to infinity.
Since a line can be infinitely long, a line goes through the point of infinity.
Thus, if we can prove that points A_1', B_1', B_2' lie on one line, then we finish our proof.

Construct a line Joining point O_4 and B_1', and make this line intersect S_3' at point C
$\because$ S_2' and S_3' are parallel lines
$\therefore \angle B_2'O_4B_1' = \angle A_1'CB_1'$
$\because$ points A_1' and B_1' are both tangent to circle S_1'
$\therefore A_1'C = B_1'C$
$\because$ O_4B_2' and O_4B_1' are both radius of circle L'
$\therefore O_4B_2' = O_4B_1'$
$\therefore \triangle B_2'B_1'O_4$ similar to $\triangle A_1'B_1'C$
$\because$ points O_4, B_1', C lie on one line by our construction
$\therefore$ points A_1', B_1', B_2' must lie on one line in order to make $\angle B_2'B_1'O_4 = \angle A_1'B_1'C$
$\therefore$ the four points A_1', A_2', B_1', B_2' lie on one line after inversion
$\therefore$ the four pints A_1, A_2, B_1, B_2 lie on one circle before inversion □

Mobius transformations

In this lecture we will study a very natural class of transformations of which the inversion is a particular case. We will see their relations to complex analysis and hyperbolic geometry.

0.5.1 Introduction

Our discussion of inversion in the plane was rather unsatisfying from the point of view of Klein's definition of what geometry is. Of course we introduced a very beautiful class of transformations - inversions send lines or circles to lines or circles and preserves angles up to a sign. But if we want to study geometry, which is preserved by inversions, we should embed them into a group. For instance, we know almost everything about a single inversion; but what about a composition of two of them? It is also a wonderful transformation - it preserves angles and sends lines or circles to lines or circles. And what about a composition of three? Is it an inversion in some circle?

One approach to defining "inversive" geometry - that is a geometry where the group of motions contains all inversions - is to say that a transformation is in the group of motions if it is a composition of a finite number of inversions.

Such an approach is somewhat too abstract - it seems quite complicated, given a transformation, to decide whether it is a composition of inversions or not. Of course such a transformation should send lines or circles to lines or circles and preserve angles (up to a sign), but the converse is not clear.

Another approach is to define our group of motions as the group of all transformations that send lines or circles to lines or circles and preserve angles (up to a sign). This is clearly a group that contains all inversions. Now the problem is quite opposite to the problem in the previous approach. We would like to find a collection of simple transformations with the property that every transformation in our group can be obtained as a composition of

several transformations from this collection.If we have such a collection, it is easy to verify that all our transformation satisfy some property (e.g. preserve some quantity). For this we only need to check that the simple transformations from the collection satisfy this property and that if two transformations satisfy this property, then their composition does.

In fact both a these approaches lead to the same answer. Moreover, in the plane this answer can be made very concrete if we think of the plane R^2 as the complex line C.

0.5.2 Fractional linear transformations in the extended complex line

We will start our journey into the world of Mobius geometry by a study of a very concrete group of transformations of the extended complex line that preserve angles and generalized circles.

First few definitions. The extended complex line is the complex line $\mathbb{C}$ with a point at infinity, ∞, adjoined. A generalized circle is either a circle in $\mathbb{C}$ or a line in $\mathbb{C}$ together with the point ∞.

Recall first how the arithmetic of complex numbers works:

1. To add two complex numbers z_1 and z_2 we just add the corresponding vectors (each vector joining 0 and z_i).

2. To multiply two complex numbers z_1 and z_2, we should multiply their magnitudes (i.e. the lengths of the corresponding vectors) and add their argumets (i.e. the angles the vectors make with the ray of positive real numbers): $|z_1 z_2| = |z_1|\,|z_2|$, $arg(z_1 z_2) = arg(z_1) + arg(z_2)$.

3. The conjugate of the number z is the number $\bar{z}$, which is the reflection of z in the real axis.

4. The inverse of a non-zero complex number z is the unique number $w = \frac{1}{z}$ with the property $zw = 1$: $\left|\frac{1}{z}\right| = \frac{1}{|z|}$,$arg(1/z) = -arg(z)$. It is equal to the number $\frac{\bar{z}}{|z|^2}$.

The last identity has the following geometrical meaning: the transformation $z \to 1/z$ is the composition of reflection in the real line $z \to \bar{z}$ with inversion in the unit circle $z \to z/|z|^2$ ($z/|z|^2$ is the point on the ray from the origin to z with distance from the origin equal to $1/|z|$). Since reflection and inversion both flip the signs of angles and send generalized circles to generalized circles, their composition preserves angles and sends generalized circles to generalized circles.

Here are some other examples of transformations that have the same property of preserving angles and sending generalized circles to generalized circles.

1. The translations $z \to z+c$. Translations are Euclidean transformations: they preserve everything: distances, angles, shapes etc.

2. The dilations $z \to \lambda z$ with λ real and positive (λ here is the factor of dilation).

3. The rotations $z \to \lambda z$ with λ complex number of magnitude 1 (this is the rotation about the origin by angle equal to argument of λ).

4. The transformation $z \to \lambda z$ with $\lambda \neq 0$ is a composition of dilation by $|\lambda|$ and rotation about the origin by angle $\arg(z)$.

5. The "inversion" $z \to \frac{1}{z}$. As we explained above, this is a composition of reflection in the real line with the inversion in unit circle.

By composing maps from the five examples above, we always get mappings of the form $z \to \frac{az+b}{cz+d}$ with $a, b, c, d \in \mathbb{C}, ad - bc \neq 0$. In fact any transformation of this kind can be obtained by a composition of mappings from the examples above. Indeed, if $c \neq 0$, then we can divide $az+b$ by $cz+d$ with remainder: $\frac{az+b}{cz+d} = \frac{a}{c} + \frac{bc-ad}{c} \cdot \frac{1}{cz+d}$. So the transformation $z \to \frac{az+b}{cz+d}$ is a composition of multiplication by c, translation by d, inversion, multiplication by $\frac{bc-ad}{c}$ and translation by $\frac{a}{c}$. We see from this formula the need for the condition $ad - bc \neq 0$. If it were zero, our transformation would send any point to the constant point $\frac{a}{c}$. In the case $c = 0$, it's even simpler: $\frac{az+b}{cz+d} = \frac{a}{d} \cdot z + \frac{b}{d}$.

A transformation given by $z \to \frac{az+b}{cz+d}$ with $a, b, c, d \in \mathbb{C}, ad-bc \neq 0$ is called a fractional linear transformation. We have proved that every fractional linear transformation preserves generalized circles and angles between them (with orientation).

A simple observation (that requires some checking) is that fractional linear transformations form a group under composition. That is the composition of two fractional linear transformations is fractional linear, and every transformation $z \to \frac{az+b}{cz+d}$ with $a, b, c, d \in \mathbb{C}, ad - bc \neq 0$ has an inverse $z \to \frac{dz-b}{-cz+a}$. The composition of these transformations in either order is the identity transformation.

Note that if $z_1, z_2, z_3 \in \mathbb{C}$ are distinct, then the fractional linear transformation $T^{z_1,z_2,z_3}_{0,1,\infty} : z \to \frac{z_2-z_3}{z_2-z_1} \cdot \frac{z-z_1}{z-z_3}$ sends z_1 to 0, z_2 to 1 and z_3 to ∞. So if

$u_1, u_2, u_3 \in \mathbb{C}$ is another triple of distinct complex numbers, then there is a fractional linear transformation $T^{z_1,z_2,z_3}_{u_1,u_2,u_3}$ mapping z_1 to u_1, z_2 to u_2 and z_3 to u_3. This transformation is given by the composition $\left(T^{u_1u_2u_3}_{0,1,\infty}\right)^{-1} \circ T^{z_1z_2z_3}_{0,1,\infty}$.

The next theorem gives an important geometric characterization of fractional linear transformations.

Theorem 3. *Let f be a one-to-one transformation of the extended complex line that sends generalized circles to generalized circles and preserves angles. Then f is a fractional linear transformation.*

Proof. Let z_1, z_2, z_3 be the images of points $0, 1$ and ∞ under the transformation f. There exists a fractional linear transformation $T^{z_1,z_2,z_3}_{0,1,\infty}$ that sends z_1, z_2, z_3 back to $0, 1, \infty$. The composition of f and $T^{z_1,z_2,z_3}_{0,1,\infty}$ is a transformation that fixes $0, 1$ and ∞, preserves angles and sends generalized circles to generalized circles. If we prove that this composition is the identity, it will follow that f is the inverse of $T^{z_1,z_2,z_3}_{0,1,\infty}$, so it must be fractional linear.

Now let us prove that if g is a transformation that fixes $0, 1$ and ∞, preserves angles and sends generalized circles to generalized circles, then it is the identity transformation.

The real line is the only generalized circle passing through the points $0, 1$ and ∞, so it must be sent to itself by transformation g (that is not to say that each point is fixed, only that the set is its own image).

Let z be any point not on the real line. Connect z to point 0 by a line l_0. This line is the only generalized circle that passes through the points z, 0 and ∞. Hence it must be mapped to the only generalized circle that passes through $g(z)$,0 and ∞. Since generalized circles passing through ∞ are lines, the image of this line is a line connecting $g(z)$ with 0.

Similarly the image of the line l_1 connecting the point z to the point 1 is the line connecting the point $g(z)$ to the point 1.

Now we use that g preserves angles: the angle formed by l_0 and the real line must be equal to the angle formed by line $g(l_0)$ and the real line. Hence $g(l_0)$ must coincide with l_0. Similarly the image $g(l_1)$ must coincide with the line l_1. Thus the only finite intersection point z of lines l_0 and l_1 must get mapped to the only intersection point of lines $g(l_0)$ and $g(l_1)$, namely itself.

Thus g fixes every point z not lying on the real line.

It is easy to see from this that in fact g fixes all points of the extended complex line, i.e. it is the identity mapping (e.g. we can repeat the argument above with a line parallel to the real line). □

Remark: if we want to study the transformations that preserve angles but reverse their orientations (while still sending generalized circles to generalized circles), we can proceed as follows. First apply the reflection $z \to \bar{z}$, and then apply our transformation g. The composite transformation $z \to g(\bar{z})$ preserves angles and thus is of the form $z \to \frac{az+b}{cz+d}$ with $a, b, c, d \in \mathbb{C}, ad-bc \neq 0$. Hence the original transformation g is of the form $z \to \frac{a\bar{z}+b}{c\bar{z}+d}$.

0.5.3 Hyperbolic geometry

The fractional linear transformations form a rather rich family of transformations: for instance every three points can be mapped to three other given points by such a transformation. This is much more freedom than what we had for the Euclidean motions, where all we can do is map any one point to any other point.

Even if we restrict our view to transformations that map some given circle to itself, we still get a rich enough family. To see this, we can think of the transformations that send the real line to itself. If the numbers a, b, c, d in the transformation $z \to (az+b)/(cz+d)$ are all real, then any real number z stays on the real line. So we get a three-dimensional family of transformations preserving the real line (three-, not four-, because we can always multiply all numbers a, b, c, d by a constant and get the same transformation).

In fact this family of transformations can be used as a basis for a geometry which is in many respects similar to the Euclidean geometry. More precisely, we can study the geometry on the upper half plane $Im(z) > 0$ where the group of motions is the group of transformations $z \to (az+b)/(cz+d)$ with real a, b, c, d and $ad-bc > 0$ (the second condition is necessary if we want our transformations to map the upper half-plane to the upper half-plane, rather than the lower one).

In this geometry one can define the "lines" to be the generalized semicircles which are orthogonal to the real axis. We already know from the properties of fractional linear transformations that these "lines" get mapped to "lines". Also it's easy to see that through any two points passes a unique "line": the semicircle passing through points P and Q and orthogonal to the real line should have center on the intersection of perpendicular bisector to the segment PQ and the real line.

We define angles in this geometry to be the usual Euclidean angles. We already know that these are preserved by the transformations in the group of motions we are talking about. In fact one can go pretty far along this road

and define many notions we are familiar with from Euclidean geometry, like distance, circles, areas etc, and prove analogues of familiar theorems about planar geometry.

This geometry is called hyperbolic geometry. In hyperbolic geometry all axioms that were used by Euclid to define the usual planar geometry hold, besides one - the axiom that states that for every line l and any point P not on the line there exists exactly one line which is parallel to l and passing through P. In hyperbolic geometry the corresponding axiom should be changed to "for every line l and point P not on this line there exist two distinct lines l' and l'' passing through point P and parallel to l such that any line between them is also parallel to l".

We are not going to study hyperbolic geometry in detail. We will instead study spherical geometry, which has a very simple model, but many features of hyperbolic geometry that are not present in the Euclidean one.

0.6 How to construct things using ruler and compass

One of the questions in which the ancients were interested was which constructions can be performed using only a ruler and a compass. What we mean by a ruler is just a tool that enables one to construct the straight line through two given points. A compass is a tool that constructs a circle given its center and a point on it. These are abstractions of the physical objects "ruler" and "compass," so they can do nothing else - the ruler can't be used to measure distances or build right angles.

Here is an example of a ruler and compass construction:

Given two points A and B we can construct the midpoint of the segment AB by the following procedure. Connect the points A and B by a segment. Construct circle $C_A(B)$ with center at A passing through point B. Construct circle $C_B(A)$ with center at B passing through point A. Connect the two intersection points of these two circles by a straight line. The intersection of this line and the segment AB is the midpoint of AB, because the picture is symmetric with respect to the reflection that interchanges A and B.

Note that in this construction we found not only the midpoint of AB, but also the perpendicular bisector of the segment AB.

asyinclude7PerpendicularBisector

We can use the constructions we know to build up more sophisticated ones. For example given three non-collinear points A, B and C we can construct the circle passing through them. Indeed, its center is located at the intersection point of perpendicular bisectors of AB and BC. Since we know how to construct these, we can construct the center of the circle we need, and then construct the circle itself using compass.

Once we know that we can construct the circumscribed circle of a triangle, it's natural to ask about the inscribed one. Since the center of the inscribed circle is the point of intersection of angle bisectors, all we have to do is to learn how to construct a bisector of a given angle $\angle ABC$.

AngleBisector

This is easy: first we construct a circle with center at B and radius BA. Let A' be its intersection point with the line BC. Now construct the intersection point of circles with centers ant A and A' and radii AB and $A'B$ respectively. One of the intersection points is B. Call the other one B'. The line BB' is the angle bisector of $\angle ABC$.

Our next task will be to divide a given segment into n equal parts (where n is some given number). If we knew how to pass a line parallel to a given one through a given point, the following procedure would work. To divide the segment AB, pass any line through point A which does not contain the point B. On this line mark n segments AA_1, A_1A_2,...,$A_{n-1}A_n$ of equal length (this can be done easily using a compass). Now connect A_n to B by a straight line and pass lines parallel to it through all other points A_i. Let B_i be the intersection points of these lines with segment AB. Then points $B_1, \ldots, B_n = B$ subdivide the segment AB into n equal parts.

DivideSegmentToNParts

The only component we are still missing is the construction of a line parallel to a given line l passing through a given point A. We can construct such a line as follows. Choose any point B on the line l. Let point C be the other point of intersection of the circle centered at A with radius AB and the line l. The angle bisector of the angle formed by the ray AB and the ray AC is parallel to the line l.

Another natural question to ask is whether we can subdivide a given angle into n equal parts. The answer to this question is much less obvious than what we have yet done, and we shall devote the next chapter to answering it. It turns out that most angles cannot be trisected, for example. And questions like whether some given angle (say 90 degrees) can be subdivided into 17 equal parts are rather deep and required the genius of a young Gauss

to answer.

One of the most complicated constructions known to ancients was the construction of a circle tangent to three given ones. This problem is known as Appolonius' problem.

We will study this problem only in the case that not all three circles intersect each other. To be covered:

- Appolonius problem - constructing a circle tangent to three given circles.
- Using inversion to simplify Appolonius problem.