NOTES FOR MAT4510 FALL 2015

JOHN ROGNES

Contents

Date: November 27th 2015.

1. August 19th lecture

1.1. **Quick overview.** We will use Bjørn Jahren's lecture notes "Geometric Structures in Dimension Two". The five chapters are about:

(Ch1) Hilbert's axiom system for planar geometry (some connection with MAT2500 – Geometry).

(Ch2) Hyperbolic geometry (2-dimensional theory; the 3-dimensional theory is an active field of research).

(Ch3) Classification of surfaces (higher-dimensional classifications rely on tools from MAT4530 – Algebraic Topology I and MAT4540 – Algebraic Topology II).

(Ch4) Geometry on surfaces (the distinction between topological and differential manifolds is developed in all dimensions in MAT4520 – Manifolds).

(Ch5) Differential geometry (Riemannian geometry is developed fully in MAT4590 – Differential Geometry).

Let us start with a more detailed survey of the contents of this course.

1.2. **Synthetic geometry.** Euclid's "Elements" from ca. 400 BC, treats the following subjects in thirteen books:

Plane geometry (the Pythagorean theorem), number theory (the irrationality of $\sqrt{2}$, the existence of infinitely many primes), solid geometry (construction of the Platonic polyhedra).

Start with undefined terms (point, line, contains, equal, congruent) and some postulated axioms about them.

(Eu1) Given any two points A and B, with A not equal to B, there exists a unique (one and only one) line ℓ that contains A and B.

Define the segment of points between A and B, and the ray from A passing through B.

(Eu2) Can construct segments of any given length along any given ray.

Define the circle with center A and radius congruent to the segment from A to B.

(Eu3) Can construct a circle with any given segment as a radius.

Define angle, supplementary angle and right angle.

(Eu4) Any two right angles are congruent.

Define parallel lines (as lines that do not meet).

(Eu5) For every point A and line ℓ that does not contain A, there exists a unique line m that contains A and is parallel to ℓ.

Deduce other statements, called propositions, about the undefined and defined terms, as logical consequences of the axioms and the previously proved propositions. The body of statements that can be established as proven propositions is known as "synthetic geometry".

Particularly noteworthy propositions, perhaps having short statements compared to the length of the proof needed to establish them, are called theorems.

1.3. **The Pythagorean theorem.** Consider a triangle $\triangle ABC$, with $\angle BCA$ a right angle. Construct a square $\square ADEB$ with side AB, a square $\square BFGC$ with side BC, and a square $\square CHIA$ with side CA. Then the area of $\square ADEB$ is equal to the sum of the areas of the squares $\square BFGC$ and $\square CHIA$, in the sense that $\square ADEB$ can be divided into finitely many triangles, and $\square BFGC$ and $\square CHIA$ can be divided into finitely many triangles, and there is a one-to-one correspondence between these two lists of triangles such that each triangle in one list is congruent to the corresponding triangle in the other list.

The proof assumes that we already have established how to construct the perpendicular from a point to a line, that two triangles with congruent sides, angles and sides (in order) are congruent, and that two times the area of a triangle is equal to the area of a rectangle with the same base and height as the triangle.

Draw the perpendicular from C through AB, intersecting AB in J and intersecting DE in K. We claim that the area of rectangle $\square ADKJ$ is equal to the area of square $\square CHIA$. In the same way the area of rectangle $\square BJKE$ is equal to the area of square $\square BFGC$. Since the area of $\square ADEB$ is the sum of the areas of the rectangles $\square ADKJ$ and $\square BJKE$, this will complete the proof.

To prove the claim, draw the lines CD and BI. The points B, C and H lie on the same line, parallel to AI, so the triangle $\triangle ABI$ has base AI and height AC. Hence two times the area of $\triangle ABI$ is equal to the area of the square $\square CHIA$.

The right angles $\angle CAI$ and $\angle DAB$ are congruent. Hence adding the angle $\angle BAC$ to each of these gives congruent angles $\angle BAI$ and $\angle DAC$. Furthermore the segments AI and AC are congruent, since $\square CHIA$ is a square, and the segments AB and AD are congruent, since $\square ADEB$ is a square. Hence the triangles $\triangle ABI$ and $\triangle ADC$ are congruent.

Finally, the points C, J and K lie on the same line, parallel to AD, so the triangle $\triangle ADC$ has base AD and height AJ. Hence two times the are of $\triangle ADC$ is equal to the area of the rectangle $\square ADKJ$.

In symbols, $\square CHIA \cong 2 \cdot \triangle ABI \cong 2 \cdot \triangle ADC \cong \square ADKJ$, as claimed.

Q.E.D.

This proof should illustrate how theorems are deduced from previously established propositions, using the undefined terms, the postulated axioms, and logical reasoning.

Euclid's books were used for over two thousand years, but Euclid's presentation is nonetheless incomplete by modern mathematical standards. Some assumptions are not made explicit.

Consider the following fallacious proof that every triangle is isosceles, i.e., has two sides of equal length. The proof has been attributed to Charles Dodgson, better known by his pen name Lewis Carroll ("Alice's Adventures in Wonderland" and "Through the Looking-Glass").

Consider a triangle $\triangle ABC$. We will prove that AB is congruent to AC. Consider the angle bisector ℓ of $\angle BAC$ and the side bisector m of BC. If these are parallel (or equal), then $\triangle ABC$ is isosceles. (We omit this part of the proof, which is not fallacious.) Otherwise, they meet in a point D. Draw the perpendicular from D to AB meeting the latter line in E, and draw the perpendicular from D to AC meeting the latter line in F.

Since D lies on the angle bisector ℓ, $\angle BAD = \angle EAD$ is congruent to $\angle CAD = \angle FAD$, and the right-angled triangles $\triangle AED$ and $\triangle AFD$ have the same hypotenuse AD, these two triangles are congruent. In particular, the segment AE is congruent to the segment AF, and the segment DE is congruent to the segment DF.

Since D lies on the side bisector m, segment BD is congruent to segment CD. Hence the right-angled triangles $\triangle BED$ and $\triangle CFD$ have congruent sides $DE \cong DF$ and congruent hypotenuses $BD \cong CD$, hence must be congruent. In particular, the remaining sides BE and CF must be congruent.

Hence the sides AB and AC satisfy $AB+BE = AE \cong AF = AC+CF$. Canceling $BE \cong CF$ we deduce that $AB \cong AC$, so $\triangle ABC$ is isosceles.

Q.E.D.(?)

What is the hidden, erroneous, assumption?

1.4. **Incidence geometries.** More famously, Euclid's axiom Eu5, the parallel postulate, was for a long time suspected to be superfluous, in the sense that it might possibly be deduced from the other axioms. This expectation was proven to be faulty early in the 19th century, when models were found for other geometries than the Euclidean one, where each of the undefined terms is given some concrete interpretation, and the first four of Euclid's axioms are true statements in the model, but the parallel postulate is not correct.

One of these models had been right under the nose of the Arabic astronomers and European explorers, namely in terms of the spherical geometry and trigonometry needed for navigation at sea. If one interprets "point" as a point on the surface of a sphere, such as the earth, and "line" as a great circle on that sphere, with origin at the center of the earth, the Euclid's axioms Eu1–Eu4 are almost satisfied. The only difficulty is that two distinct points A and B do not determine a unique line (= great circle) through them if the two points happen to be antipodal, i.e., on opposite sides of the earth. Hence the uniqueness clause in Eu1 fails for spherical geometry.

Presumably the navigators thought of spherical geometry as a part of solid, or 3-dimensional, geometry, rather than as planar, 2-dimensional geometry. Probably they would also not think that it could make sense to interpret the word "line" as something other than a straight line in 2- or 3-dimensional space. This required a separation of the idea that there is only one "geometry", representing the physical world around us, and the insight that synthetic geometry consists of formal reasoning about undefined terms, which might not be directly tied to that one interpretation. Even if motion along a great circle is locally the shortest path between two points on the surface of the earth, and such motion is not perceived as involving any turning to the left or to the right relative to the direction of travel, it must have been difficult to think that the word "line" could meaningfully be used about a path that is curved in the surrounding 3-dimensional space.

Nonetheless, by making a small change to the spherical model for geometry, an interpretation of "points", "lines", "congruence" and the other undefined terms of synthetic geometry can be given, satisfying Eu1–Eu4. The trick is called projective geometry, where a "point" is reinterpreted to mean a pair of antipodal points on the sphere, and a "line" is reinterpreted to mean the collection of points lying on a great circle on the sphere. Now Eu1–Eu4 hold, but Eu5 does not. For any great circle ℓ on the sphere and any pair $\{A, A'\}$ of antipodal points, not lying on ℓ, there does not exist any great circle m through $\{A, A'\}$ that does not meet m. In fact, any pair of great circles will intersect; there are no parallel lines in this geometry.

This example also shows that the parallel postulate is not superfluous; the example of projective geometry shows that there is no way that one can deduce Eu5 from Eu1-Eu4, since this would lead to a contradiction when the axioms are interpreted in terms of this model.

This geometry is also called elliptic geometry, because the perpendicular lines ℓ and m through the ends of a segment AB do not remain at constant distance, but come closer as one moves away from AB, and meet in a point C at a finite distance. Notice then that the triangle $\triangle ABC$ has two right angles, so that the sum of the angles in a spherical triangle is greater than two right angles. The excess is related to the area of the triangle, and can be expressed by saying that this is a positively curved geometry. (We will discuss curvature more precisely in the course.)

At a small scale near a point, the curvature of a sphere becomes less and less noticeable as the radius of the sphere increases. One measure of curvature is related to the square of the radius, and so one might hope that a sphere with purely imaginary radius might model a geometry with negative curvature. It turns out that no such complex valued interpretation is needed. Frameworks for hyperbolic geometry, where the distance between the perpendiculars ℓ and m at the ends of a segment AB increases to infinity as one moves away from AB, were found by Bolyai, Gauss and Lobachevsky, in some order. One model for hyperbolic geometry, due to Beltrami but often called the Poincaré disc model $\mathbb{D}$, is given by the open unit disc in the Euclidean plane. A "hyperbolic point" is a point in this open disc, but a "hyperbolic line" is the part of a Euclidean line or a Euclidean circle that lies in this disc, subject to the condition that the line or circle is orthogonal to the unit circle, i.e., the circle of the open unit disc.

This is also a model for Euclid's axioms Eu1-Eu4, but now the parallel postulate Eu5 fails in a different way from in the elliptic case. Namely, given a hyperbolic line ℓ and a hyperbolic point A not on ℓ, there is not just one but infinitely many hyperbolic lines m through A that do not meet ℓ, i.e., that are parallel to ℓ. The sum of the angles in a hyperbolic triangle $\triangle ABC$ will always be less than two right angles. The defect is reflected in the fact that this is a negatively curved geometry.

1.5. **Betweenness and Congruence.** Returning to the fallacious proof, that every triangle is isosceles, the problem is not related to the parallel postulate, but has to do with the ordering of points on a line, i.e., what it means for a point to lie between two other points. An accurate drawing will show that if E lies on the far side of B, as seen from A, so that B is between A and E, then F will not lie on the far side of C, as seen from A, but rather will lie between A and C. So if $AB + BE = AE$ we will have $AF + CF = AC$, not $AC + CF = AF$. Thus $AB + BE = AE \cong AF$ and $AF + CF = AC$, with $BE \cong CF$, but now that the signs have changed we cannot cancel two equal terms to deduce that $AB = AC$.

To precisely axiomatize planar geometry, Hilbert therefore replaced Euclid's axioms Eu1-Eu4 with a more precise set of axioms. First there are axioms concerning "incidence". These only concern points and lines and the property that a given point may or not may lie on a given line. If it does, we say that the point and line are incident.

Next there are axioms concerning "betweenness". These concern the property that three points A, B and C on a line come in a given order, so that B is between A and C. We write $A * B * C$ to express this.

Thereafter there are axioms for "congruence". When is a line segment congruent to another line segment, and when is an angle congruent to another angle?

Nearing the end, there are axioms related to constructions with compass and straightedge, ensuring that a line and a circle, or two circles, meet in exactly two points when this is reasonable.

Up to this point, it is not clear whether the points on a line, with a chosen origin (0) and unit length ($[0, 1]$) correspond to all real numbers, or only a suitable subfield. An completeness axiom similar to Dedekind's construction of the real numbers will pin this down.

Finally, Hilbert adds an axiom to replace Euclid's parallel postulate. Given a point A and a line ℓ, with A not on ℓ, if one postulates the existence of a unique parallel m to ℓ through A one recovers the classical Euclidean plane, modeled by $\mathbb{R}^2$. If one instead postulates that there exist at least two parallels m and m' to ℓ through A, then there are in fact infinitely many, and one recovers the hyperbolic plane $\mathbb{H}^2$, modeled by the Poincaré disc $\mathbb{D}$. If one postulates that no parallel line exists, one recovers the elliptic geometry, modeled by the projective plane $\mathbb{R}P^2$ of pairs of antipodal points on the unit sphere S^2, or equivalently, the space of lines through the origin in $\mathbb{R}^3$.

1.6. **The hyperbolic plane.** We will spend a good part of the course developing a couple of equivalent models for the hyperbolic plane. Sticking with the Poincaré disc model $\mathbb{D}$ for now, the points are the elements of the unit disc and the lines are the intersections of Euclidean lines or circles that meet the boundary of the disc at right angles. The notion of incidence, i.e., whether a point lies on a given line or not, is then the evident one. The notion of betweenness is also the obvious one, since each hyperbolic line is topologically equivalent (= homeomorphic) to an open interval, so the complement of each point consists of two connected components.

The notion of congruence requires more work, especially for segments. We need to determine when two hyperbolic segments AB and CD are congruent. This will allow us to introduce a notion of linear content, or length, so that AB and CD are congruent if and only if they are of the same length. It turns out that there is a group of isometries $\gamma \colon \mathbb{D} \to \mathbb{D}$, i.e., length-preserving bijections, such that AB is congruent to CD if and only if there exists such a γ with $\gamma(AB) = CD$. Hence each "local congruence" between line segments is realized by a "global congruence" of the entire geometry.

This situation is more familiar in the Euclidean and the elliptic cases. The isometries of the Euclidean plane $\mathbb{R}^2$ are the so-called Euclidean motions, which are bijections $\gamma \colon \mathbb{R}^2 \to \mathbb{R}^2$ of the form $\gamma(x) = Rx + t$ with $R \in O(2)$ an orthogonal 2×2 matrix (representing a rotation or a reflection) and $t \in \mathbb{R}^2$ a vector (representing a translation). Two Euclidean segments AB and CD are congruent, or equally long, if and only if there exists an Euclidean motion γ with $\gamma(AB) = CD$. More precisely, R must be chosen to turn AB in the same direction as CD, and

t must be chosen to move $R(A)$ to C. Then $\gamma(A) = C$ and $\gamma(AB)$ points in the same direction as CD, so $\gamma(B) = D$ if and only if AB is equally long as CD.

The isometries of the projective plane $\mathbb{R}P^2$ are the rotations and reflections $\gamma(x) = Rx$, with $R \in O(3)$ an orthogonal 3×3-matrix. Multiplication by R defines a linear bijection $\mathbb{R}^3 \to \mathbb{R}^3$, but since R preserves distances, it restricts to a bijection $S^2 \to S^2$ of the unit sphere. It takes antipodal points to antipodal points, hence induces a bijection $\gamma\colon \mathbb{R}P^2 \to \mathbb{R}P^2$, and this is the rotation or reflection that we have in mind. Again, two spherical segments (arcs, parts of great circles) AB and CD are congruent, or equally long, if and only if there exists a rotation (or reflection) γ with $\gamma(AB) = CD$. We can first rotate A to C, about an axis orthogonal to the great circle through A and C. Thereafter, by rotating around the axis through C we may bring the image of B to D if and only if AB and CD are equally long.

In the same way, we will construct a group of operations on the unit disc $\mathbb{D}$, i.e., bijections $\gamma\colon \mathbb{D} \to \mathbb{D}$, such that two hyperbolic segments AB and CD are congruent if and only if $\gamma(AB) = CD$ for one of these particular bijections. These operations γ will then be distance-preserving bijections of $\mathbb{D}$, i.e., isometries of the hyperbolic plane. If we identify $\mathbb{R}^2$ with $\mathbb{C}$ in the usual way, so that $\mathbb{D}$ corresponds to the complex numbers z of modulus $|z| < 1$, these isometries can be written on one of the forms

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

or

$$\gamma(z) = \frac{a\bar{z} + b}{c\bar{z} + d}$$

for suitable $a, b, c, d \in \mathbb{C}$. (The coefficients must be chosen so that γ maps $\mathbb{D}$ bijectively to $\mathbb{D}$.) These maps are called fractional linear transformations, or Möbius transformations.

Once we have constructed this group of Möbius transformations, we have well-defined notions of congruence and length for hyperbolic segments. It turns out that each Möbius transformation is conformal, i.e., preserves angles, so we also get well-defined notions of congruence and angle measure of hyperbolic angles. In fact, the hyperbolic angle between two hyperbolic lines meeting at A will be the same as the Euclidean angle between their respective Euclidean tangent lines at A. With these definitions, we can verify that the Poincaré disc satisfies all of Hilbert's axioms for plane hyperbolic geometry.

2. August 21st lecture

2.1. **Homogeneous spaces.** Each of the three geometries X we now have discussed, namely the Euclidean plane $\mathbb{R}^2$, the projective plane $\mathbb{R}P^2$ and the hyperbolic plane $\mathbb{D}$, are homogeneous, in the sense that there is a group $G = \{\gamma\colon X \to X\}$ of isometries that acts transitively on the set X of points, meaning that for any two points $A, B \in X$ there exists an isometry $\gamma \in G$ with $\gamma(A) = B$.

In the Euclidean case $G = O(2) \ltimes \mathbb{R}^2$ is the semi-direct product of $O(2)$ and $\mathbb{R}^2$, with respect to the standard action of $O(2)$ on $\mathbb{R}^2$. In the elliptic case $G = PO(3) = O(3)/\{\pm I\}$ is the quotient of $O(3)$ by its center $\{\pm I\}$. In the hyperbolic case G contains $PSL_2(\mathbb{R}) = SL_2(\mathbb{R})/\{\pm I\}$ as a subgroup of index two, where $SL_2(\mathbb{R})$ is the group of matrices

$$\gamma = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

with $a, b, c, d \in \mathbb{R}$ and $ad - bc = 1$. There is a relation between these matrices and the Möbius transformations mentioned above, but this connection is more easily seen with another model for the hyperbolic plane than the unit disc model.

Selecting a point $A \in X$, we may consider the subgroup $H \subset G$ of isometries η that fix A, i.e., such that $\eta(A) = A$. Then the map $G \to X$ taking γ to $\gamma(A)$ will take each right coset $\gamma H = \{\gamma\eta \mid \eta \in H\}$ to the same point in X, and therefore induces a bijection

$$G/H \xrightarrow{\cong} X$$

taking γH to $\gamma(A)$. We call H the stabilizer group of A. Such spaces G/H, with G a Lie group and H a closed subgroup, are called homogeneous spaces. Restricting to the orientation-preserving isometries, we get bijections

$$SO(2) \ltimes \mathbb{R}^2/SO(2) \cong \mathbb{R}^2 ,$$

$$SO(3)/SO(2) \cong S^2$$

(the projective case is a little more complicated) and

$$SL_2(\mathbb{R})/SO(2) \cong \mathbb{D} .$$

2.2. **Trigonometry.** Just as two points determine a line and a segment, three points A, B, C will determine a triangle $\triangle ABC$, with three sides AB, BC, CA and three angles $\angle BAC$, $\angle CBA$, $\angle ACB$. As is familiar from the Euclidean case, these quantities are not unrelated. The study of the relations between these lengths and angles is the subject of trigonometry. In the Euclidean case, the key relations are the sine rule, the cosine rule, and the fact that the sum of the three angles equals two right angles. We shall determine the corresponding laws of hyperbolic trigonometry, including a formula for the area contained within a hyperbolic triangle.

2.3. **Differential geometry.** A very different point of view on the role of lines in geometry is found in Riemann's lectures on the foundations of geometry, which relies on the differential calculus of Newton and Leibniz. We start by extending the notion of length from being defined for line segments to being defined for more general curves. In the Euclidean case, a curve in the plane may be parametrized by a function

$$\omega \colon [a, b] \to \mathbb{R}^2$$

where $[a, b]$ is some interval. For each $t \in [a, b]$, the value $\omega(t)$ is then a point on the curve, and as a set the curve consists of all these values. If

$$a = c_0 < c_1 < \cdots < c_n = b$$

is a partition of the interval, we may consider the part of the curve from $\omega(c_{i-1})$ to $\omega(c_i)$ for each $1 \le i \le n$. We require that the length of that part of the curve shall be greater than or equal to the length of the line segment between these two points, i.e.,

$$\|\omega(c_i) - \omega(c_{i-1})\|$$

where $\|(x, y)\| = \sqrt{x^2 + y^2}$ is the Euclidean norm. Hence the length of ω should be greater or equal than the sum

$$\sum_{i=1}^{n} \|\omega(c_i) - \omega(c_{i-1})\| .$$

Letting n and the partition $\{c_i\}_{i=0}^n$ vary, we get a set of such sums, and if this set has a least upper bound, we say that the curve ω is rectifiable, and define its length to be that supremum:

$$\sup_{a=c_0<\cdots<c_n=b} \sum_{i=1}^{n} \|\omega(c_i) - \omega(c_{i-1})\| .$$

For Riemannian geometry, we limit attention to the curves that are continuously differentiable, i.e., such that the tangent vector $\omega'(t)$ is defined for each $t \in [a, b]$, and such that the rule $t \mapsto \omega'(t)$ defines a continuous function. Such a curve is also called a C^1 curve. When the partition is fine enough, the sum

$$\sum_{i=1}^{n} \|\omega(c_i) - \omega(c_{i-1})\|$$

is closely approximated by the sum

$$\sum_{i=1}^{n} \|\omega'(d_i)\|(c_i - c_{i-1})$$

for any choice of points $d_i \in [c_{i-1}, c_i]$ for $1 \le i \le n$. Since the function $t \mapsto \|\omega'(t)\|$ is assumed to be continuous, it is Riemann integrable. It follows that the curve is rectifiable, and that its length is given by the integral

$$\int_a^b \|\omega'(t)\| \, dt \, .$$

2.4. **Lines as shortest paths.** Instead of starting with the primitive concept of a line in $\mathbb{R}^2$, we may instead start with the measure $\|\omega'(t)\|$ of the length of each tangent vector, and use the formula above to define the length of any C^1 curve. The line segment from A to B in $\mathbb{R}^2$ can then be characterized as the shortest path from A to B, i.e., the image of the C^1 curve $\omega \colon [a, b] \to \mathbb{R}^2$ from $\omega(a) = A$ to $\omega(b) = B$ of minimal length. (It is necessary to prove that such a curve exists, and that it is uniquely determined up to reparametrization. The corresponding problem for surfaces of minimal area is much more difficult.)

A line in $\mathbb{R}^2$ is then the image of a C^1 curve $\omega \colon \mathbb{R} \to \mathbb{R}^2$ that is locally of minimal length, i.e., such that for each $t \in \mathbb{R}$ there is an interval $[a, b]$ with $a < t < b$ such that the restricted curve $\omega|_{[a,b]} \colon [a, b] \to \mathbb{R}^2$ is the shortest path from $\omega(a)$ to $\omega(b)$. Such a path is called a complete geodesic, and this approach lets us recognize the lines in $\mathbb{R}^2$ as the complete geodesics in the plane equipped with the Euclidean length measure for tangent vectors.

The same approach can be applied in the elliptic and hyperbolic cases. A C^1 curve $\bar{\omega} \colon [a, b] \to \mathbb{R}P^2$ can be represented by a curve

$$\omega \colon [a, b] \to S^2$$

on the unit sphere S^2 inside $\mathbb{R}^3$, and each tangent vector $\omega'(t)$ to the curve, which lies in the tangent plane to S^2 at $\omega(t)$, can be viewed as a vector in $\mathbb{R}^3$. We declare its length $\|\omega'(t)\|$ to be given by the Euclidean norm $\|(x, y, z)\| = \sqrt{x^2 + y^2 + z^2}$, and define the length of $\bar{\omega}$, as well as that of ω, to be the same integral

$$\int_a^b \|\omega'(t)\| \, dt$$

as before, except that now each $\omega'(t)$ is a tangent vector to S^2 in $\mathbb{R}^3$, while earlier it was a tangent vector in $\mathbb{R}^2$. The shortest paths between two points on $\mathbb{R}P^2$ then turn out to be the segments of great circles, i.e., the elliptic line segments. Furthermore, the complete geodesics parametrize the great circles. This lets us recognize the elliptic lines in $\mathbb{R}P^2$ as the complete geodesics in the projective plane equipped with the length measure for tangent vectors inherited via S^2 from $\mathbb{R}^3$.

In the hyperbolic case, each C^1 curve $\omega \colon [a, b] \to \mathbb{D}$ can be viewed as a curve in $\mathbb{R}^2$, but in order to recognize the hyperbolic line segments as the shortest paths between two points, we cannot simply define the hyperbolic length of a tangent vector $\omega'(t)$ in $\mathbb{D}$ to be the same as its Euclidean length as a tangent vector in $\mathbb{R}^2$. A scaling factor of $2/(1 - \|\omega(t)\|^2)$ turns out to be needed, which grows to infinity as $\omega(t)$ approaches the boundary of the unit disc. This leads to the formula

$$\int_a^b \frac{2\|\omega'(t)\|}{1 - \|\omega(t)\|^2} \, dt$$

for the hyperbolic length of the curve ω, where now $\|\omega'(t)\|$ refers to the Euclidean length of $\omega'(t)$. With this modified measure of length, or norm, for tangent vectors in $\mathbb{D}$, the hyperbolic line segments become the shortest paths between their endpoints, and the hyperbolic lines are the complete geodesics.

Note that in the elliptic case the complete geodesics are closed loops of finite length, while in the hyperbolic case they are simple curves of infinite length.

2.5. **Riemannian manifolds.** At this point, it becomes possible to vastly generalize the scope of geometry. For any differentiable manifold, i.e., a topological space M that is locally homeomorphic to $\mathbb{R}^n$ for some n, equipped with some additional structure to make sure that we can talk about an n-dimensional vector space T_pM of tangent vectors to M at each point p of M, we may consider C^1 curves $\omega\colon [a,b] \to M$ in M, having tangent vectors $\omega'(t) \in T_{\omega(t)}M$ for each $t \in [a,b]$. A Riemannian metric on M is one more additional structure, namely an inner product pairing on each tangent space T_pM, which permits us to speak about the length $\|\omega'(t)\|$ of each of these tangent vectors. A Riemannian manifold is a differentiable manifold equipped with a Riemannian metric. We define the length of the curve ω in the Riemannian manifold M to be the integral

$$\int_a^b \|\omega'(t)\|\, dt.$$

This in turn allows us to identify the shortest curves in M between two points p and q. The images of these curves can now play the roles of generalized line segments in a geometric theory on M. The complete geodesics for this length measure now play the role of (full) lines. They may no longer satisfy the axioms of Euclid or Hilbert, but they still carry a lot of spatial, geometric information. The study of the geometry of Riemannian manifolds is called Riemannian geometry.

2.6. **Surfaces in $\mathbb{R}^3$.** In the case $n = 2$, a 2-dimensional manifold is usually called a surface. We shall study surfaces $M \subset \mathbb{R}^3$ contained in ordinary Euclidean 3-space, with differentiable structure defined in such a way that each tangent vector to M can be viewed as a vector in $\mathbb{R}^3$, i.e., that $T_pM \subset \mathbb{R}^3$ for each $p \in M$. (This generalizes the case of $S^2 \subset \mathbb{R}^3$ that we discussed earlier, in the context of elliptic geometry.) We can then let each tangent plane T_pM inherit the Euclidean inner product from the containing space $\mathbb{R}^3$, so that the length $\|\omega'(t)\|$ of a tangent vector to M in p is set to be equal to the Euclidean length of $\omega'(t)$ viewed as a vector in $\mathbb{R}^3$. In this case we say that M has the Riemannian metric inherited from the ambient space $\mathbb{R}^3$.

For example, we might consider the torus $M = T^2$ embedded in $\mathbb{R}^3$ as the surface of rotation obtained by rotating the circle of radius r in the xz-plane, with origin at $(R, 0, 0)$, about the z-axis. Here we assume $0 < r < R$, to ensure that M is topologically equivalent to the Cartesian product $S^1 \times S^1$ of a circle with itself. This surface is like the boundary of a doughnut. Some closed geodesics can be recognized on this surface, but the general picture is complex and fascinating. Near the outer perimeter the geometry is positively curved, like that of a sphere. Here the sum of angles in a small geodesic triangle is larger than two right angles. Near the inner perimeter, it is negatively curved, like that of a hyperboloid. Here the sum of angles in a small geodesic triangle is smaller than two right angles. In a region near the top and bottom the curvature is close to zero.

Alternatively, we might consider the surface of a glass or bowl in $\mathbb{R}^3$. In most cases this surface is topologically equivalent to S^2, but the inherited Riemannian metric will be very different, as will the structure of the collection of geodesic curves.

We might also consider the surface of a cup, or mug, in $\mathbb{R}^3$. If the cup has exactly one handle, the surface is topologically equivalent to $T^2 = S^1 \times S^1$, but the Riemannian metric will depend heavily on the shape of the cup.

Continuing, we might consider the surface of a mug with two or more handles. If there are g handles, this is called a surface of genus g. In the inherited Riemannian metric, it will have parts of positive curvature at the outside of the handles, and at the bottom of the mug (both inside and outside). It will have parts of negative curvature at the inside of the handles, and along the upper rim.

2.7. **Abstract surfaces.** More generally, we might consider abstractly defined 2-dimensional manifolds, or surfaces, that are not explicitly presented as subspaces of $\mathbb{R}^3$. For example, we might consider the unit square $I^2 = [0,1] \times [0,1]$ in the plane, with the sides identified so that $(x, 0) \sim (x, 1)$ for all $x \in [0,1]$, and $(0, y) \sim (1, y)$ for all $y \in [0,1]$. We can define the length $\|\omega'(t)\|$ of the tangent vector to a curve as the Euclidean length of the corresponding tangent vector in $\mathbb{R}^2$. This makes sense because the identifications $(x, 0) \sim (x, 1)$ and $(0, y) \sim (1, y)$ are realized by Euclidean motions, namely $(x, y) \mapsto (x, y + 1)$ and $(x, y) \mapsto (x + 1, y)$, respectively, so the definition of length is compatible with the implicit identifications. Locally, this geometry is flat, i.e., just like that of the Euclidean plane. We therefore call this Riemannian manifold the flat torus. Note that this is rather different than the Riemannian geometry on the boundary of a doughnut, inherited from the surrounding Euclidean space.

The flat torus can also be realized by making identifications of opposite edges in a regular hexagon.

A surprising theorem of the late Abel prize laureate John Nash asserts that allowing abstractly defined surfaces and abstract Riemannian metrics does in fact not add any generality compared to only considering surfaces realized as subspaces of $\mathbb{R}^3$, with the inherited Riemannian metric. In other words, any abstract Riemannian manifold can be C^1 isometrically embedded in a Euclidean space. (One can be more precise about the dimensions of these manifolds and the surrounding space.)

For example the flat torus can be realized as a subspace of $\mathbb{R}^3$, in such a way that the lengths of tangent vectors, hence also the lengths of curves, are the same in that subspace as in the abstract flat model $I^2/{\sim}$. (A picture approximating such an embedding was recently produced by the Hevea project.) One can even arrange that this subspace is contained in an arbitrarily small ball.

The genus g surface, for $g \geq 2$, can also be realized as an identification space, now starting with a polygon with $4g$ edges, which have to be identified in pairs in a specific order. For simplicity let us assume $g = 2$, so that we are considering the boundary of a mug with two handles. Starting with an octagon $ABCDEFGH$, the side AB is identified with DC, the side BC is identified with ED, the side EF is identified with HG and the side FG is identified with AH. Topologically this produces a manifold M that is topologically a genus 2 surface. However, if we started with a regular Euclidean octagon, it is not possible to give M a Riemannian metric in a way that is respected by the identifications. All eight corners are identified to one point p in M, and going once around p in M corresponds to looping through all eight angles in the octagon, in the order $\angle A$, $\angle D$, $\angle C$, $\angle B$, $\angle E$, $\angle H$, $\angle G$ and $\angle F$. Since each interior angle in the Euclidean octagon is one-and-a-half right angle (also known as 135 degrees), this adds to twelve right angles, which is three times as much as a full rotation should be for any inner product on the plane.

A solution is to replace the Euclidean octahedron with a hyperbolic octahedron, where the angle sum in a polygon can be made smaller by bringing the corners of the polygon closer to the boundary "at infinity" of the hyperbolic plane. It is possible to put the eight corners A to H as in a regular Euclidean octahedron, inside the open unit disc $\mathbb{D}$ model for the hyperbolic plane, and to replace the eight Euclidean line segments AB to HA with the corresponding eight hyperbolic line segments. Then the inner angles in this regular hyperbolic octahedron will be smaller than 135 degrees, and can be arranged to be exactly half a right angle, i.e., 45 degrees. If the identifications are made from this part of the hyperbolic plane, the total angle around the conjoined point p will now be four right angles, as desired. Furthermore each identification can be realized by a hyperbolic isometry, i.e., a Möbius transformation, and therefore the genus 2 surface M can be given the structure of a Riemannian manifold in such a way that it is locally just like a piece of the hyperbolic plane. In particular this model in everywhere negatively curved, unlike the boundary of a mug with two handles, which had some positively curved and some negatively curved parts.

This illustrates a uniformization theorem. Every (closed, connected) surface can made geometric, i.e., given a Riemannian metric so that each small open piece of the surface is isometric to a small open piece of one of the three model geometries that we have mentioned: the positively curved elliptic geometry, the flat Euclidean geometry, or the negatively curved hyperbolic geometry. These correspond to the cases of genus $g = 0$, genus $g = 1$ and genus $g \geq 2$, respectively.

2.8. **Theorema Egregium and Gauss-Bonnet.** An important result in the differential geometry of surfaces is Gauss' Theorema Egregium (Latin for "remarkable theorem"), which shows that curvature is an intrinsic property of a surface with a Riemannian metric, i.e., that it does not depend on how this surface is or may be isometrically embedded in an ambient space. (This concerns two times continuously differentiable embeddings, hence is not contradicted by Nash' theorem.)

Another important result is the Gauss–Bonnet theorem, which shows that the integral of the curvature over a surface with a Riemannian metric is a topological invariant of the surface, namely 2π times the Euler characteristic of the surface, or equivalently $2\pi \cdot (2-2g)$ if the surface has genus g.

For instance, a sphere of radius r has genus 0 and Euler characteristic 2. The curvature at each point is $1/r^2$, and the surface area is $4\pi r^2$, so the integral of the curvature over the area is $4\pi r^2/r^2 = 4\pi = 2\pi \cdot 2$. Making the corresponding calculation for the surface of a glass or bowl, or any convex body, gives the same result.

A torus has genus 1 and Euler characteristic 0. The flat torus has curvature 0 at each point, so the integral of the curvature over the surface is $0 = 2\pi \cdot 0$. Less obviously, we get the same result if we consider the boundary of a doughnut with the metric inherited from $\mathbb{R}^3$, or the boundary of a mug with one handle. Here the curvature is sometimes positive and sometimes negative, and when integrated the positive and negative contributions cancel perfectly, leaving 0 as the answer.

A surface of genus $g \geq 2$ can be realized as an identification space from a regular hyperbolic $4g$-gon, and has constant negative curvature at all points. By Gauss–Bonnet, the product of the curvature and the area is 2π times the negative number $2-2g$. Making the same calculation for a mug with $g \geq 2$ handles, the curvature will sometimes be positive and sometimes be negative, but in this case the negative contributions dominate, leaving a negative total answer.

3. August 26th lecture

3.1. **Dimension three.** Going up one dimension, the classification of 3-dimensional manifolds is much more complicated that the classification of surfaces. Some examples of 3-manifolds are Euclidean 3-space $\mathbb{R}^3$, the 3-sphere S^3 given as the unit sphere in $\mathbb{R}^4$, and the 3-torus $T^3 = S^1 \times S^1 \times S^1$. Each 3-manifold can be realized as a subspace of a higher-dimensional Euclidean space, $M \subset \mathbb{R}^N$ for some N. They can also be presented abstractly as identification spaces obtained by making identifications along the boundary of polyhedra.

If we start with a cube, $I^3 = [0,1] \times [0,1] \times [0,1]$, also known as a six-sided die, and identify opposite sides according to the rules $(x,y,0) \sim (x,y,1)$, $(x,0,z) \sim (x,1,z)$ and $(0,y,z) \sim (1,y,z)$, then the resulting 3-manifold is topologically equivalent to the 3-torus $T^3 = S^1 \times S^1 \times S^1$. This admits a flat Riemannian metric, locally modeled on the Euclidean metric in $\mathbb{R}^3$.

If we instead identify antipodal points on the boundary of I^3, i.e., via $(x,y,z) \simeq (1-x, 1-y, 1-z)$ when at least one of x, y or z is 0 or 1, we get a 3-manifold that is homeomorphic to S^3 with antipodal points identified, i.e., to the projective space $\mathbb{R}P^3$ of lines through the origin in $\mathbb{R}^4$. This manifold admits a Riemannian metric of constant positive curvature, locally modeled on the elliptic metric on S^3.

A famous example, called the Poincaré homology sphere, is given by starting with a dodecahedron, also known as a twelve-sided die, and identifying each pair of opposite sides (which are regular pentagons) by 1/10-th of a full turn. This manifold also admits a Riemannian metric of constant positive curvature. It is an interesting example, because a collection of important invariants of topological spaces, called the homology groups, are unable to distinguish the Poincaré homology sphere from the ordinary 3-dimensional sphere S^3. Another invariant, the fundamental group, does however distinguish between these two manifolds.

Yet another example, called the Seifert–Weber space, is given by starting with a dodecahedron and identifying each pair of opposite sides using 3/10-th of a full turn. This construction can be done using a regular hyperbolic dodecahedron, so that the identifications are made using hyperbolic isometries, in such a way that the Seifert–Weber space admits a Riemannian metric of constant negative curvature.

Finally, identifying opposite sides of a dodecahedron by 5/10-th of a full turn, i.e., a half-turn, gives another model for the positively curved projective space $\mathbb{R}P^3$.

3.2. **The geometrization conjecture.** Deep insight of William Thurston suggested that the interplay between geometry and topology is almost as close in dimension three as in dimension two. Thurston's geometrization conjecture from 1982 asserts that any closed 3-manifold M can be decomposed into pieces, by cutting it open along suitable embedded spheres $S^2 \to M$ or tori $T^2 \to M$, such that each of the resulting pieces admits a geometric structure. More precisely, there are eight possible 3-dimensional model geometries, including the positively curved elliptic geometry of the sphere S^3, the flat geometry of $\mathbb{R}^3$, and the negatively curved hyperbolic geometry of a hyperbolic 3-space $\mathbb{H}^3$, and each piece of the decomposition of M admits a Riemannian metric that corresponds to one of these model geometries.

Thurston proved that a large class of 3-manifolds admit a hyperbolic structure, i.e., a Riemannian metric that is locally isometric to the hyperbolic 3-space. In a sense this is the largest class of geometric pieces of 3-manifolds. The existence of a hyperbolic geometric structure in these pieces is a very useful tool for their classification, even if the full story remains complicated.

Switching from the everywhere negatively curved to the everywhere positively curved side, the geometrization conjecture includes as a special case the Poincaré conjecture from 1904, predicting that the only simply connected closed 3-manifold is the 3-sphere. The full geometrization conjecture, including the Poincaré conjecture, was famously proven by Grigori Perelman, starting with preprints published in 2002 and 2003. (Additional details were spelled out by other authors in the following years.)

Perelman's proof develops an idea introduced by Richard Hamilton, of starting with a Riemannian metric on the given 3-manifold, and then letting the Riemannian metric evolve over time, in such a way that the rate of change of the inner product defining the metric is given, up to a sign, by a version of the curvature of the metric known as the Ricci curvature. This time-evolving family of Riemannian metrics on the same manifold is called the Ricci flow. The sign is chosen so that the parts of the manifold with positive curvature shrink in size, so that the lengths of tangent vectors decrease to zero, while the parts of the manifold with negative curvature grow, so that tangent vectors become longer. This geometric evolution helps to decompose the manifold into pieces that can be analyzed. Since the positively curved parts shrink to points in finite time, a modification process called surgery is required at various points in the process. The details of a careful proof require many precise estimates from the theory of partial differential equations on manifolds, in an area that overlaps with geometric analysis and PDEs.

3.3. **Higher dimensions.** In higher dimensions, other more exotic phenomena enter. Up to dimension three there is little difference between topological manifolds, i.e., those reasonable spaces that are locally homeomorphic to Euclidean n-space, and differential manifolds, i.e., those where we have introduced enough additional structure to be able to talk about tangent vectors and other derivatives. Starting in dimension four, it turns out that not all topological manifolds admit differentiable structures, and sometimes the same topological manifold admits many different differential structures. John Milnor showed in 1956 that the 7-dimensional sphere S^7 admits many non-equivalent differential structures, called exotic spheres. Around 1982 Simon Donaldson developed a theory particular to dimension 4, which eventually led to a proof that there are uncountably many nonequivalent differentiable structures on $\mathbb{R}^4$. The higher-dimensional analogue of the Poincaré conjecture, that a manifold that is homotopy equivalent to S^n is also homeomorphic to S^n, was proven by Stephen Smale for $n \geq 5$ around 1960, and by Mike Freedman for $n = 4$ in 1982. Technically speaking, Smale's results were proven for manifolds with a geometric structure called a piecewise linear structure, which is intermediate between a "naked" topological manifold and a differentiable manifold.

In high dimensions, we therefore first have three categories, or contexts, of manifolds to study: the topological, the piecewise linear and the differentiable (or smooth) manifolds. This area is known as geometric topology, or manifold topology. For differentiable manifolds it is possible to add a Riemannian structure, and one is led to study differential geometry and Riemannian geometry. Many interesting results concern the relationship between geometry and topology, i.e., to what extent local geometric data determine the global topology, and to what extend topological objects admit geometric models.

3.4. **Related courses.** MAT4520 – Manifolds: About differentiable manifolds, tangent vectors, tensor fields and integration.

MAT4530 – Algebraic topology I: About the fundamental group and homology.

MAT4540 – Algebraic topology II: About cohomology and Poincaré duality for manifolds.

MAT4590 – Differential geometry: About Riemannian metrics and curvature.

MAT9560 – Lie groups: About isometry groups and their homogeneous spaces.

4. August 28th lecture

4.1. **Hilbert's axiom system.** [[Start with Chapter 1.]]

4.2. **Hyperbolic geometry.** [[Continue with Chapter 2, Section 2.1.]]

4.3. **Stereographic projection.** Let $S^2 = \{(x,y,z) \in \mathbb{R}^3 \mid x^2+y^2+z^2 = 1\}$ be the unit 2-sphere in 3-space. Let $N = (0,0,1) \in S^2$ be the "north pole". Stereographic projection is a homeomorphism

$$\Phi \colon S^2 \setminus \{N\} \to \mathbb{R}^2$$

with very good geometric properties. It is defined by identifying $\mathbb{R}^2$ with the xy-plane in $\mathbb{R}^3$, and sending a point $(x,y,z) \in S^2$ different from N to the intersection of the Euclidean line through N and (x,y,z) with the xy-plane. We can parametrize the line by

$$t \mapsto (1-t)(0,0,1) + t(x,y,z) = (tx, ty, 1-t+tz)$$

for $t \in \mathbb{R}$. The intersection with the xy-plane occurs when $1-t+tz = 0$, i.e., with $t = 1/(1-z)$, hence at the point $(x/(1-z), y/(1-z), 0)$. Thus

$$\Phi(x,y,z) = \Big(\frac{x}{1-z}, \frac{y}{1-z}\Big).$$

It is geometrically clear that rotation of S^2 about the z-axis corresponds under Φ to rotation of $\mathbb{R}^2$ about the origin, through the same angle.

The inverse to stereographic projection is a homeomorphism

$$\Psi \colon \mathbb{R}^2 \to S^2 \setminus \{N\}.$$

We often view this as a map to S^2, or to $\mathbb{R}^3$, without change in the notation. By construction, $\Psi(x,y)$ is a point on the Euclidean line through N and $(x,y,0)$. It has unit length, and is different from N. We can parametrize the line by

$$t \mapsto (1-t)(0,0,1) + t(x,y,0) = (tx, ty, 1-t).$$

The point $(tx, ty, 1-t)$ lies on S^2 if $(tx)^2 + (ty)^2 + (1-t)^2 = 1$, or equivalently, if $t^2(x^2+y^2+1) - 2t = 0$. The solution $t = 0$ corresponds to the point N; we are interested in the other solution, $t = 2/(x^2+y^2+1)$. Hence

$$\begin{aligned}\Psi(x,y) &= (2x/(x^2+y^2+1), 2y/(x^2+y^2+1), 1-2/(x^2+y^2+1)) \\ &= \Big(\frac{2x}{x^2+y^2+1}, \frac{2y}{x^2+y^2+1}, \frac{x^2+y^2-1}{x^2+y^2+1}\Big).\end{aligned}$$

By construction Φ and Φ are mutually inverse continuous maps, hence they specify a topological equivalence, or homeomorphism, between the punctured sphere $S^2 \setminus \{N\}$ and the plane $\mathbb{R}^2$.

If two differentiable curves ω_1 and ω_2 in $\mathbb{R}^2$ intersect at a point P, so that $\omega_1(t_1) = P = \omega_2(t_2)$ for some parameters t_1 and t_2, we can compare the tangent vectors $\omega_1'(t_1)$ and $\omega_2'(t_2)$ of these curves, which are tangent vectors at P. If these vectors are nonzero, there is a well-defined angle between them, and we then say that the two curves ω_1 and ω_2 intersect at P at that angle.

Similarly, if two differentiable curves η_1 and η_2 in S^2 intersect at a point Q, so that $\eta_1(t_1) = Q = \eta_2(t_2)$ for some parameters t_1 and t_2, we can compare the tangent vectors $\eta_1'(t_1)$ and $\eta_2'(t_2)$

of these curves, which are tangent vectors to S^2 at Q. They line in the tangent plane $T_Q S^2$ to S^2 at Q, which we can view as a subspace of $\mathbb{R}^3$. In particular, we talk about inner products, lengths and angles for such vectors as if they were vectors in $\mathbb{R}^3$. If these two tangent vectors are nonzero, there is a well-defined angle between them, and we then say that the two curves η_1 and η_2 intersect at Q at that angle.

We say that a differentiable map is conformal if it preserves angles, i.e., if composition with this map takes any pair of curves that meet at an angle θ to a new pair of curves that meet at the same angle θ.

5. September 2nd lecture

[[We did Exercises 3, 8 and 14 from Chapter 1.]]

Lemma 5.1. *Φ and Ψ are conformal.*

Proof. We prove that $\Psi\colon \mathbb{R}^2 \to S^2 \setminus \{N\}$ is conformal, i.e., that if ω_1 and ω_2 are curves in $\mathbb{R}^2$ that meet at a point $P = (x, y)$ at an angle θ, then the image curves $\eta_1 = \Psi \circ \omega_1$ and $\eta_2 = \Psi \circ \omega_2$ in $S^2 \setminus \{N\} \subset S^2 \subset \mathbb{R}^3$ meet at $Q = \Psi(P)$ at the same angle θ.

Since Φ and Ψ are mutually inverse differentiable maps, this will also imply that $\Phi\colon S^2 \setminus \{N\} \to \mathbb{R}^2$ preserves angles between intersecting curves, hence is conformal.

Since angles between tangent vectors in $S^2 \setminus \{N\}$ are defined to be the same as the angles between the corresponding vectors in $\mathbb{R}^3$, we may as well consider Ψ as a differentiable map $\Psi\colon \mathbb{R}^2 \to \mathbb{R}^3$. The chain rule for composition of differentiable maps tells us that if $\eta = \Psi \circ \omega$, then

$$\eta'(t) = \Psi'(P)\omega'(t)\,.$$

Here $\omega'(t)$ is the tangent vector of ω at the point $P = \omega(t)$, $\Psi'(P) = \Psi'(\omega(t))$ is the differential of Ψ at P, i.e., the linear transformation from the tangent plane of $\mathbb{R}^2$ at P to the tangent space of $\mathbb{R}^3$ at $\Phi(P)$ that best approximates Φ near P. This linear transformation is given in matrix terms as multiplication by the Jacobian matrix of Φ, i.e.,

$$\Psi'(P) = \Big(\frac{\partial \Psi_i}{\partial x_j}(P)\Big)_{i,j}$$

where $\Psi = (\Psi_1, \Psi_2, \Psi_3)$ and $(x_1, x_2) = (x, y)$. The image of the vector $\omega'(t)$ under the linear transformation $\Psi'(\omega(t))$ is the tangent vector $\eta'(t)$ of η at $Q = \Psi(P)$.

We must therefore prove that the linear transformation $\Psi'(P)$ from the tangent plane of $\mathbb{R}^2$ at P to the tangent space of $\mathbb{R}^3$ at Q preserves angles between nonzero vectors. We shall see that in fact the linear transformation takes an orthonormal basis to a pair of orthogonal vectors of equal length. Hence the linear transformation preserves angles, and scales all lengths by the same constant factor.

Rotation of the xy-plane about the origin corresponds, under the inverse stereographic projection Ψ, to rotation of the unit sphere about the z-axis. Since these rotations preserve angles, we may assume that $P = (x, 0)$ is a point on the x-axis, so that $\Psi(P) = (2x/(x^2+1), 0, (x^2-1)/(x^2+1))$ is a point on the xz-plane. We now consider the effect of the differential $\Psi'(P)$ on the two unit vectors $(1, 0)$ and $(0, 1)$ in the xy-plane.

The unit tangent vector $(1, 0)$ in the positive x-direction is the velocity vector of a curve ω_1 on the x-axis, which gets mapped to a curve η_1 in the unit circle of the xz-plane. Hence $\Psi'(P)$ maps it to a tangent vector in that plane. Let us simplify the notation by restricting attention to the x-axis and the xz-plane, omitting $y = 0$ from the notation. Then inverse stereographic projection defines a map $\psi\colon \mathbb{R} \to S^1 \subset \mathbb{R}^2$, where S^1 is the unit circle, given by

$$\psi(x) = \Big(\frac{2x}{x^2+1}, \frac{x^2-1}{x^2+1}\Big)\,.$$

We calculate

$$\psi'(x) = \Big(\frac{2(x^2+1) - (2x)(2x)}{(x^2+1)^2}, \frac{(2x)(2x) - (x^2-1)(2x)}{(x^2+1)^2}\Big) = \Big(\frac{2-2x^2}{(x^2+1)^2}, \frac{4x}{(x^2+1)^2}\Big)\,,$$

so that

$$\|\psi'(x)\|^2 = \frac{(2-2x^2)^2+(4x)^2}{(x^2+1)^4} = \frac{4x^4+8x^2+4}{(x^2+1)^4} = \frac{4}{(x^2+1)^2}$$

and

$$\|\psi'(x)\| = \frac{2}{x^2+1}\,.$$

In words, the image of $\omega_1'(t_1) = (1,0)$ is a vector $\eta_1'(t_1)$ in the xz-plane of length $2/(x^2+1)$.

The unit tangent vector $(0,1)$ in the positive y-direction is the velocity vector of the circle ω_2 in the xy-plane with center at the origin, going through the point $P = (x,0)$. Hence this circle has radius $|x|$. Inverse stereographic projection maps this curve to another circle, η_2, parallel to the xy-plane and going through the point $\Psi(P) = (2x/(x^2+1), 0, (x^2-1)/(x^2+1))$. This circle has radius $2|x|/(x^2+1)$. Since the radius of the latter circle is $2/(x^2+1)$ times as large as the radius of the former circle, it follows from the rotational symmetry of the situation that the length of the velocity vector $\eta_2'(t_2)$ is also $2/(x^2+1)$ times as large as the length of the velocity vector $\omega_2'(t_2) = (0,1)$, and that these point in the same direction. Hence the image of $\omega_2'(t_2) = (0,1)$ is the vector $\eta_2'(t_2) = (0, 2/(x^2+1), 0)$ in the positive y-direction of length $2/(x^2+1)$.

Hence the orthonormal vectors $(1,0)$ and $(0,1)$ are mapped by $\Psi'(P)$ to two vectors, one in the xz-plane and one parallel with the y-axis, both having length $2/(x^2+1)$. It follows that the linear transformation $\Psi'(P)$ scales all lengths by this number, but does not alter angles. Hence Ψ preserves angles. □

Lemma 5.2. *Consider a curve $C \subset S^2$, which may or may not contain $N = (0,0,1)$.*

(a) If $N \notin C$, then C is a circle on S^2 if and only if $\Phi(C)$ is a Euclidean circle in $\mathbb{R}^2$.

(b) If $N \in C$ then C is a circle on S^2 if and only if $\Phi(C \setminus \{N\})$ is a Euclidean line in $\mathbb{R}^2$.

Proof. A circle $C \subset S^2$ is the intersection of S^2 with a plane α in $\mathbb{R}^3$, defined by an equation $ax + by + cz = d$, where we can assume that (a,b,c) is a vector of unit length that is orthogonal to α, and where $|d| < 1$. (If $|d| = 1$ the intersection consists of a single point, and if $|d| > 1$ it is empty.) We have $N \in C$ if and only if $c = d$.

The image $\Phi(C)$ (or $\Phi(C \setminus \{N\})$) of C under stereographic projection consists of the points $(x,y) \in \mathbb{R}^2$ such that $\Psi(x,y) \in C$, i.e., the points that satisfy

$$a \cdot \frac{2x}{x^2+y^2+1} + b \cdot \frac{2y}{x^2+y^2+1} + c \cdot \frac{x^2+y^2-1}{x^2+y^2+1} = d\,,$$

which we can rewrite as

$$2ax + 2by + c(x^2+y^2-1) = d(x^2+y^2+1)$$

or as

$$(c-d)(x^2+y^2) + 2ax + 2by = c + d\,.$$

If $c \neq d$, this is the equation of a Euclidean circle with center at $(-a/(c-d), -b/(c-d)) = (a/(d-c), b/(d-c))$ and positive radius r satisfying $r^2 = (1-d^2)/(c-d)^2$. If $c = d$ this is the equation of a Euclidean line, equal to the line $ax + by = c$. Here $a^2 + b^2 = 1 - c^2 = 1 - d^2$ is positive, so this is, indeed, a line.

Conversely, any Euclidean circle or line D in $\mathbb{R}^2$ can be realized by this equation, with a, b, c and d as above and then $\Psi(D)$ or $\Psi(D) \cup \{N\}$ is a circle in S^2, according to the case $c \neq d$ or $c = d$, respectively.

This is clear when D is a line. When D is a circle in $\mathbb{R}^2$, we may consider a diameter AB of D such that the line through A and B goes through the origin. The image $\Psi(AB)$ is then a segment of a great circle on S^2. The Euclidean line segment in $\mathbb{R}^3$ from $\Psi(A)$ to $\Psi(B)$ is the diameter of a circle C on S^2. The image $\Phi(C)$ is then a circle in $\mathbb{R}^2$, by what we have already shown, containing $A = \Phi(\Psi(A))$ and $B = \Phi(\Psi(B))$. Since C is symmetric about the diameter $\Psi(A)\Psi(B)$, it follows that $\Phi(C)$ is symmetric about AB, i.e., that AB is a diameter of $\Phi(C)$. Hence $\Phi(C) = D$, since a circle is determined by its diameter. This implies that $\Psi(D) = C$, so that $\Psi(D)$ is a circle on S^2. This concludes the proof. □

[[Proceed with Section 2.2.]]

5.1. **Models for the hyperbolic plane.** We will refer to the following models for the hyperbolic plane.

- The Beltrami–Klein model $\mathbb{K} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\}$, where the $\mathbb{K}$-lines are the nonempty intersections of Euclidean lines in $\mathbb{R}^2$ with the open unit disc $\mathbb{K}$.
- The lower open hemisphere $\mathbb{B} = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1, z < 0\}$, where the $\mathbb{B}$-lines are semi-circles in S^2 meeting the boundary $\partial\mathbb{B} = \{(x, y, 0) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\}$ orthogonally.
- The Poincaré disc model $\mathbb{D} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\} = \{z \in \mathbb{C} \mid |z| < 1\}$, where the $\mathbb{D}$-lines are segments of Euclidean lines or circles in $\mathbb{R}^2$ meeting the boundary $\partial\mathbb{D} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\} = \{z \in \mathbb{C} \mid |z| = 1\}$ orthogonally.
- The right-hand open hemisphere $\mathbb{B}' = \{(x, y, z) \in \mathbb{R}^3 \mid x^2+y^2+z^2 = 1, y > 0\}$, where the $\mathbb{B}$-lines are semi-circles in S^2 meeting the boundary $\partial\mathbb{B}' = \{(x, 0, z) \in \mathbb{R}^3 \mid x^2 + z^2 = 1\}$ orthogonally.
- The upper half-plane model $\mathbb{H} = \{(x, y) \in \mathbb{R}^2 \mid y > 0\} = \{z \in \mathbb{C} \mid \operatorname{Im} z > 0\}$, where the $\mathbb{H}$-lines are segments of Euclidean lines or circles in $\mathbb{R}^2$ meeting the boundary line $\{(x, 0) \in \mathbb{R}^2\} = \mathbb{R} \subset \mathbb{C}$ orthogonally.

We are most interested in the disc and half-plane models, $\mathbb{D}$ and $\mathbb{H}$.

We use the following homeomorphisms to identify the various models.

- Vertical projection $(x, y) \leftrightarrow (x, y, -\sqrt{1 - x^2 - y^2})$ identifies $\mathbb{K}$ and $\mathbb{B}$.
- Stereographic projection $\Phi \colon \mathbb{B} \to \mathbb{D}$, with inverse $\Psi \colon \mathbb{D} \to \mathbb{B}$, identifies $\mathbb{B}$ and $\mathbb{D}$.
- Rotation through a right angle about the x-axis, $(x, y, z) \mapsto (x, -z, y)$, identifies the lower open hemisphere $\mathbb{B}$ with the right-hand open hemisphere $\mathbb{B}'$.
- Stereographic projection $\Phi \colon \mathbb{B}' \to \mathbb{H}$, with inverse $\Psi \colon \mathbb{H} \to \mathbb{B}'$, identifies $\mathbb{B}'$ and $\mathbb{H}$.

Given what we have proved about stereographic projection, it is clear that these identifications take the lines in one model to the lines of each other model. The combined identification $\mathbb{D} \cong \mathbb{B} \cong \mathbb{B}' \cong \mathbb{H}$ turns out to be given by the formula

$$z \mapsto \frac{1}{i}\frac{z + i}{z - i} = \frac{z + i}{iz + 1}$$

for $|z| < 1$.

6. September 4th lecture

6.1. **Bijections preserving lines and circles.** We shall define the notion of congruence between line segments, in each of these models, by means of a group of homeomorphisms acting transitively on the model. These homeomorphisms will then turn out to preserve all hyperbolic lengths, i.e., be hyperbolic isometries, and to map lines (= complete geodesics) to lines, in each model. In particular, the isometries of the disc model will be bijections $\gamma \colon \mathbb{D} \to \mathbb{D}$ that map $\mathbb{D}$-lines to $\mathbb{D}$-lines. Similarly, the isometries of the upper half-plane model will be bijections $\gamma \colon \mathbb{H} \to \mathbb{H}$ that map $\mathbb{H}$-lines to $\mathbb{H}$-lines.

Notice that all of these bijections $\gamma \colon \mathbb{D} \to \mathbb{D}$ and $\gamma \colon \mathbb{H} \to \mathbb{H}$ map segments of some Euclidean lines or circles in $\mathbb{R}^2 = \mathbb{C}$ to segments of segments of the same kind of Euclidean lines or circles. It turns out that these maps can be extended to take almost all Euclidean lines or circles in $\mathbb{C}$ to Euclidean lines or circles in $\mathbb{C}$, only with the exception that when a circle is sent to a line, one point gets sent to "infinity".

To compensate for this exception, we instead work with the one-point compactification $\bar{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ of $\mathbb{C}$, known as the Riemann sphere. Stereographic projection $\Phi \colon S^2 \setminus \{N\} \to \mathbb{R}^2 = \mathbb{C}$ extends to a homeomorphism $\Phi \colon S^2 \to \bar{\mathbb{C}}$. A curve in $\bar{\mathbb{C}}$ that is either a Euclidean circle in $\mathbb{C}$, or of the form $\ell \cup \{\infty\}$ where ℓ is a Euclidean line in $\mathbb{C}$, will be called a $\bar{\mathbb{C}}$-circle. (These curves correspond under Φ to the great circles in S^2.) Working in $\bar{\mathbb{C}}$, the boundary of $\mathbb{H}$ also contains the point ∞, hence equals $\partial\mathbb{H} = \bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. Stereographic projection also induces a homeomorphism $S^1 \cong \bar{\mathbb{R}}$.

Each hyperbolic isometry $\gamma\colon \mathbb{D} \to \mathbb{D}$ or $\gamma\colon \mathbb{H} \to \mathbb{H}$ will be obtained by restriction from a bijection

$$m\colon \bar{\mathbb{C}} \to \bar{\mathbb{C}}$$

that takes $\bar{\mathbb{C}}$-circles to $\bar{\mathbb{C}}$-circles. These so-called Möbius transformations turn out to be of two kinds: the fractional linear transformations (FLTs)

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

where $a, b, c, d \in \mathbb{C}$ satisfy $ad - bc \neq 0$, and their composites with the complex conjugation map $z \mapsto \bar{z}$, given by

$$n(z) = \frac{a\bar{z} + b}{c\bar{z} + d}\,,$$

with a, b, c, d as above. The first kind of map is complex differentiable, hence holomorphic. The second kind is not complex differentiable, but can be referred to as anti-holomorphic. (This means that the real differential is not complex linear, but takes multiplication by i to multiplication by $-i$.) An FLT m is thus a holomorphic Möbius transformation. The composite n of an FLT and complex conjugation is an anti-holomorphic Möbius transformation. Alternatively, we may refer to an FLT m as an even Möbius transformation, and to the composite n of an FLT and complex conjugation as an odd Möbius transformation.

6.2. **Fractional linear transformations (FLTs).**

Definition 6.1. A *complex fractional linear transformation* (FLT) is a meromorphic function $m\colon \bar{\mathbb{C}} \to \bar{\mathbb{C}}$ given by the formula

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

for some $a, b, c, d \in \mathbb{C}$ with $ad - bc \neq 0$. We write $\text{Möb}^+ = \text{Möb}^+(\mathbb{C})$ for the set of all FLTs. In view of the following lemma, this is a (non-abelian) group.

Lemma 6.2. *The FLTs form a group under composition. The composite $m \circ n$ of $m(z) = (az + b)/(cz + d)$ and $n(z) = (a'z + b)/(c'z + d)$ is the FLT*

$$(m \circ n)(z) = m(n(z)) = \frac{(aa' + bc')z + (ab' + bd')}{(ca' + dc')z + (cb' + dd')}\,.$$

(Here $(aa' + bc')(cb' + dd') - (ab' + bd')(ca' + dc') = (ad - bc)(a'd' - b'c') \neq 0$.) The FLT $e(z) = (1z + 0)/(0z + 1)$ acts as the identity. The inverse of $m(z) = (az + b)/(cz + d)$ is the FLT

$$m^{-1}(z) = \frac{dz - b}{-cz + a}\,.$$

Proof. These claims are verified by direct calculations. □

Lemma 6.3. *Each FLT can be written as a composite of one or more of the following FLTs:*

- $m(z) = az = (az + 0)/(0z + 1)$ *with* $a \in \mathbb{C}$, $a \neq 0$ *(rotation and scaling);*
- $m(z) = z + b = (1z + b)/(0z + 1)$ *with* $b \in \mathbb{C}$ *(translation);*
- $m(z) = 1/z = (0z + 1)/(1z + 0)$ *(inversion).*

Hence these elements generate Möb^+ *as a group.*

Proof. If $c = 0$ so that $ad \neq 0$ we can write $m(z) = (az + b)/d = (a/d)z + (b/d)$ as $z \mapsto (a/d)z$ followed by $z \mapsto z + (b/d)$. Otherwise, we can write

$$\frac{az + b}{cz + d} = \frac{1}{c}\Big(a - \frac{ad - bc}{cz + d}\Big)$$

as the composite of $z \mapsto cz$, $z \mapsto z + d$, $z \mapsto 1/z$, $z \mapsto -(ad - bc)z$, $z \mapsto z + a$ and $a \mapsto (1/c)z$, in order. (It suffices to use $b = 1$, since $z + b = b((1/b)z + 1)$ for $b \neq 0$.) □

Lemma 6.4. *(a) An FLT is conformal, i.e., preserves angles between tangent vectors.*

(b) An FLT maps $\bar{\mathbb{C}}$-circles to $\bar{\mathbb{C}}$-circles.

Proof. (a) The differential of an FLT m takes the tangent vector $\omega'(t)$ of a curve ω in $\bar{\mathbb{C}}$ to the tangent vector $\eta'(t)$ of the composite curve $\eta = m \circ \omega$ in $\bar{\mathbb{C}}$. The linear transformation taking $\omega'(t) \in \mathbb{C}$ to $\eta'(t) \in \mathbb{C}$ is given by multiplication with the complex derivative

$$m'(z) = \frac{a(cz+d) - (az+b)c}{(cz+d)^2} = \frac{ad-bc}{(cz+d)^2}$$

at the point $z = \omega(t)$. This is a nonzero complex number [[when $ad-bc \neq 0$ and $cz+d \neq 0$]], so multiplication by it corresponds to rotation through an angle equal to the argument of $m'(z)$, and scaling by a factor equal to the modulus (= norm) of $m'(z)$. Both rotation and scaling preserves angles between vectors, hence so does the differential of m. [What happens at ∞? Two $\bar{\mathbb{C}}$-lines intersecting at ∞ also intersect, at the same angle, at some point in $\mathbb{C}$, ETC.]]

(b) It is clear that $m(z) = az$ and $m(z) = z + b$, with $a \neq 0$, each map $\bar{\mathbb{C}}$-circles to $\bar{\mathbb{C}}$-circles. It therefore suffices to verify that $m(z) = 1/z$ also maps $\bar{\mathbb{C}}$-circles to $\bar{\mathbb{C}}$-circles. [[This can be geometrically verified by realizing inversion in terms of two stereographic projections, but we instead give a calculational proof.]]

We can write the equation

$$(x-p)^2 + (y-q)^2 = r^2$$

of a circle in $\mathbb{R}^2 = \mathbb{C}$, with center (p, q) and radius $r > 0$, in terms of $z = x + iy$ as

$$z\bar{z} - p(z+\bar{z}) + iq(z-\bar{z}) + p^2 + q^2 = r^2$$

or as

$$z\bar{z} + \mu z + \bar{\mu}\bar{z} + \nu = 0$$

where $\mu = -p + iq \in \mathbb{C}$ and $\nu = p^2 + q^2 - r^2 \in \mathbb{R}$, subject to $\nu < \mu\bar{\mu}$. Multiplying by a scalar $\lambda \in \mathbb{R}$, and replacing $\lambda\mu$ and $\lambda\nu$ by μ and ν, respectively, we get the equation

$$\lambda z\bar{z} + \mu z + \bar{\mu}\bar{z} + \nu = 0\,,$$

with $\lambda, \nu \in \mathbb{R}$, $\mu \in \mathbb{C}$ and $\lambda\nu < \mu\bar{\mu}$. If $\lambda = 0$ this is the equation

$$\mu z + \bar{\mu}\bar{z} + \nu = 0$$

of a line in $\mathbb{R}^2 = \mathbb{C}$, with $\nu \in \mathbb{R}$, $\mu \in \mathbb{C}$ and $\mu \neq 0$, otherwise it is the equation of a circle.

Let $w = m(z) = 1/z$. As z ranges through the solutions of the equation above, w ranges through the solutions of the equation $\lambda/(w\bar{w}) + \mu/w + \bar{\mu}/\bar{w} + \nu = 0$. Multiplying by $w\bar{w}$, and reordering, we can rewrite this as

$$\nu w\bar{w} + \bar{\mu}w + \mu\bar{w} + \lambda = 0\,,$$

with $\nu, \lambda \in \mathbb{R}$, $\bar{\mu} \in \mathbb{C}$ and $\nu\lambda < \bar{\mu}\mu$. If $\nu = 0$ this is the equation of a line, otherwise it is the equation of a circle. Hence $m(z) = 1/z$ maps any $\mathbb{C}$-circle to another $\mathbb{C}$-circle. [[A similar calculation works for any FLT $m(z) = (az + b)/(cz + d)$.]] □

Corollary 6.5. *(a) Any FLT $m\colon \bar{\mathbb{C}} \to \bar{\mathbb{C}}$ that maps $\mathbb{D}$ to itself, so that $m(\mathbb{D}) = \mathbb{D}$, takes $\mathbb{D}$-lines to $\mathbb{D}$-lines.*

(b) Any FLT $m\colon \bar{\mathbb{C}} \to \bar{\mathbb{C}}$ that maps $\mathbb{H}$ to itself, so that $m(\mathbb{H}) = \mathbb{H}$, takes $\mathbb{H}$-lines to $\mathbb{H}$-lines.

Proof. (a) If $m(z) = (az + b)/(cz + d)$ maps $\mathbb{D}$ to itself, it must take the boundary circle $\partial\mathbb{D}$ to itself. Since it preserves angles (is conformal), it takes any $\bar{\mathbb{C}}$-circle (line or circle) that meets $\partial\mathbb{D}$ at a right angle to another $\bar{\mathbb{C}}$-circle that meets $\partial\mathbb{D}$ at a right angle. Restricting to the part in $\mathbb{D}$, it follows that m takes $\mathbb{D}$-lines to $\mathbb{D}$-lines.

(b) If $m(z) = (az + b)/(cz + d)$ maps $\mathbb{H}$ to itself, it must take the boundary circle $\partial\mathbb{H} = \bar{\mathbb{R}}$ to itself. Since it preserves angles (is conformal), it takes any $\bar{\mathbb{C}}$-circle (line or circle) that meets $\partial\mathbb{H}$ at a right angle to another $\bar{\mathbb{C}}$-circle that meets $\partial\mathbb{H}$ at a right angle. Restricting to the part in $\mathbb{H}$, it follows that m takes $\mathbb{H}$-lines to $\mathbb{H}$-lines. □

6.3. **The projective linear group.** As an aside, we note that the group Möb$^+$ of FLTs is closely related to the group

$$GL_2(\mathbb{C}) = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \mid a, b, c, d \in \mathbb{C}, ad - bc \neq 0 \right\}$$

of invertible complex 2 × 2 matrices, with group structure given by matrix multiplication

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix} = \begin{bmatrix} aa' + bc' & ab' + bd' \\ ca' + dc' & cb' + dd' \end{bmatrix} .$$

The neutral element is the identity matrix $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. The rule

$$\phi \colon GL_2(\mathbb{C}) \to \text{Möb}^+$$

sending the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ to the FLT $m(z) = (az + b)/(cz + d)$ is then a surjective group homomorphism. The kernel, $D = \ker(\phi) = \phi^{-1}(e)$, consists of the matrices that are sent to the identity FLT, i.e., such that $(az + b)/(cz + d) = z$ for all z. This holds if and only if $a = d$ and $b = c = 0$, so that

$$D = \left\{ \lambda I = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \mid \lambda \in \mathbb{C}, \lambda \neq 0 \right\}$$

consists of the (invertible) diagonal matrices. Hence the matrices that correspond to the same FLT as $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ are precisely the products

$$\begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} \lambda a & \lambda b \\ \lambda c & \lambda d \end{bmatrix}$$

for $\lambda \neq 0$. We therefore get a group isomorphism

$$\bar{\phi} \colon GL_2(\mathbb{C})/D \xrightarrow{\cong} \text{Möb}^+ .$$

This kernel D, of diagonal matrices, equals the center of $GL_2(\mathbb{C})$, so the quotient group on the left hand side is often called the projective linear group $PGL_2(\mathbb{C})$. Hence the group of FLTs, or even/holomorphic Möbius transformations, is isomorphic to this projective linear group.

6.4. **Action on three points.** In order to determine which FLTs map $\mathbb{D}$ to $\mathbb{D}$, or map $\mathbb{H}$ to $\mathbb{H}$, it is illuminating to analyze to what extent we can prescribe the values of an FLT on a given set of points in $\bar{\mathbb{C}}$.

Lemma 6.6. *Given three distinct points z_1, z_2 and z_3 in $\bar{\mathbb{C}}$ there is one and only one FLT $m \colon \bar{\mathbb{C}} \to \bar{\mathbb{C}}$ such that $m(z_1) = 1$, $m(z_2) = 0$ and $m(z_3) = \infty$.*

If all three points lie in $\bar{\mathbb{R}}$, then m may be expressed with real coefficients, i.e., as $m(z) = (az + b)/(cz + d)$ with $a, b, c, d \in \mathbb{R}$ and $ad - bc \neq 0$.

Proof. If z_1, z_2 and z_3 all lie in $\mathbb{C}$, the FLT

$$m(z) = \frac{z - z_2}{z_1 - z_2} \frac{z_1 - z_3}{z - z_3}$$

satisfies $m(z_1) = 1$, $m(z_2) = 0$ and $m(z_3) = \infty$, as required. There are three other cases:

- If $z_1 = \infty$, let $m(z) = (z - z_2)/(z - z_3)$.
- If $z_2 = \infty$, let $m(z) = (z_1 - z_3)/(z - z_3)$.
- If $z_3 = \infty$, let $m(z) = (z - z_2)/(z_1 - z_2)$.

In each case, $m(z_1) = 1$, $m(z_2) = 0$ and $m(z_3) = \infty$, as required.

This establishes existence. The claim about the real case follows by inspection of the formulas.

To prove uniqueness, suppose that $n \colon \bar{\mathbb{C}} \to \bar{\mathbb{C}}$ is also an FLT such that $n(z_1) = 1$, $n(z_2) = 0$ and $n(z_3) = \infty$. Then the composite FLT $\ell = m \circ n^{-1}$ satisfies $\ell(1) = 1$, $\ell(0) = 0$ and $\ell(\infty) = \infty$. The last condition implies $\ell(z) = az + b$ for some $a, b \in \mathbb{C}$, and the first two conditions imply

$a = 1$ and $b = 0$, so that $\ell = e$ is the identity. It follows that $m = m \circ n^{-1} \circ n = e \circ n = n$, proving uniqueness. □

Corollary 6.7. *Given three distinct points z_1, z_2 and z_3 in $\bar{\mathbb{C}}$, and (again) three distinct points w_1, w_2 and w_3 in $\bar{\mathbb{C}}$, there is one and only one FLT $m\colon \bar{\mathbb{C}} \to \bar{\mathbb{C}}$ such that $m(z_1) = w_1$, $m(z_2) = w_2$ and $m(z_3) = w_3$.*

If all six points lie in $\bar{\mathbb{R}}$, then m may be expressed with real coefficients, i.e., as $m(z) = (az + b)/(cz + d)$ with $a, b, c, d \in \mathbb{R}$ and $ad - bc \neq 0$.

Proof. Choose FLTs ℓ and n so that $\ell(z_1) = 1$, $\ell(z_2) = 0$, $\ell(z_3) = \infty$, $n(w_1) = 1$, $n(w_2) = 0$ and $n(w_3) = \infty$. Then $m = n^{-1} \circ \ell$ maps z_i to w_i for $i = 1, 2, 3$, and may be expressed with real coefficients if ℓ and n admit such a presentation. Uniqueness is proved in the same way as in the lemma. □

Remark 6.8. These results show that $PGL_2(\mathbb{C}) = \text{Möb}^+$ acts simply transitively on the configuration space

$$F_3(\bar{\mathbb{C}}) = \{(z_1, z_2, z_3) \in \bar{\mathbb{C}}^3 \mid z_i \neq z_j \text{ for } i \neq j\}$$

of three distinct ordered points in $\bar{\mathbb{C}}$. The map

$$\text{Möb}^+ \xrightarrow{\cong} F_3(\bar{\mathbb{C}})$$

taking m to $(m^{-1}(1), m^{-1}(0), m^{-1}(\infty))$ is a topological equivalence (homeomorphism), even if the group structure on the left hand side is not evident on the right hand side. The use of m^{-1} in place of m ensures that a triple of points (z_1, z_2, z_3) on the right hand side corresponds to the FLT m on the left hand side with $m(z_1) = 1$, $m(z_2) = 0$ and $m(z_3) = \infty$, as above.

Lemma 6.9. *Given two $\bar{\mathbb{C}}$-circles C and C', there exists an FLT m such that $m(C) = C'$. In other words,* Möb$^+$ *acts transitively on the set of $\bar{\mathbb{C}}$-lines in $\bar{\mathbb{C}}$.*

Proof. Note first that given three distinct points in $\bar{\mathbb{C}}$ there is one and only one $\bar{\mathbb{C}}$-circle through all three of them. If one of the points is at ∞, the $\bar{\mathbb{C}}$-circle is the line through the other two points. Otherwise, if all three points lie on a line, the $\bar{\mathbb{C}}$-circle must be that line. Finally, if all three points do not lie on a line, hence form a triangle, the $\mathbb{C}$-circle must be the circumcircle of that triangle. Its center lies on the perpendicular bisector of each of the three edges of the triangle.

Now choose three distinct points z_1, z_2 and z_3 on C and three distinct points w_1, w_2 and w_3 on C'. Let $m\colon \bar{\mathbb{C}} \to \bar{\mathbb{C}}$ be the FLT with $m(z_i) = w_i$ for $i = 1$, 2 and 3. Then $m(C)$ and C' are $\bar{\mathbb{C}}$-circles with three points in common. As noted above, this implies that $m(C) = C'$. □

7. September 9th lecture

7.1. Preserving the upper half-plane.

Proposition 7.1. *An FLT $m\colon \bar{\mathbb{C}} \to \bar{\mathbb{C}}$ restricts to a homeomorphism of the upper half-plane $\mathbb{H} \to \mathbb{H}$ if and only if it can be written in the form $m(z) = (az + b)/(cz + d)$ where a, b, c and d are real and $ad - bc = 1$. Such FLTs map $\mathbb{H}$-lines to $\mathbb{H}$-lines.*

Proof. If $m(\mathbb{H}) = \mathbb{H}$ then $m(\bar{\mathbb{R}}) = \bar{\mathbb{R}}$, so $m(1)$, $m(0)$ and $m(\infty)$ all lie in $\bar{\mathbb{R}}$. Hence m may be expressed with real coefficients a, b, c and d. Assuming this, the calculation

$$\begin{aligned} m(z) &= \frac{az + b}{cz + d} = \frac{(az + b)(c\bar{z} + d)}{(cz + d)(c\bar{z} + d)} \\ &= \frac{ac|z|^2 + (ad + bc)\operatorname{Re} z + bd}{|cz + d|^2} + i\frac{(ad - bc)\operatorname{Im} z}{|cz + d|^2} \end{aligned}$$

shows that m maps the upper half-plane (where $\operatorname{Im} z > 0$) to the upper half-plane (where $\operatorname{Im} m(z) > 0$) if and only if $ad - bc > 0$. In this case we may divide each of a, b, c and d by one of the real square roots of $ad - bc$, to arrange that $ad - bc = 1$.

Conversely, if a, b, c and d are real, with $ad - bc = 1$, then $m(\bar{\mathbb{R}}) = \bar{\mathbb{R}}$ and m maps $\mathbb{H}$ onto $\mathbb{H}$, so m restricts to a homeomorphism of the upper half-plane.

Each $\mathbb{H}$-line L is part of a unique $\mathbb{C}$-circle C that meets $\bar{\mathbb{R}}$ at a right angle. Any FLT m with $m(\bar{\mathbb{R}}) = \bar{\mathbb{R}}$ will map this to another $\mathbb{C}$-circle $m(C)$ that meets $\bar{\mathbb{R}}$ at a right angle. Hence any FLT m with $m(\mathbb{H}) = \mathbb{H}$ will take the $\mathbb{H}$-line L to an $\mathbb{H}$-line $m(L)$. □

Definition 7.2. Let Möb$^+(\mathbb{H})$ denote the group of FLTs that restrict to homeomorphisms of $\mathbb{H}$. It is a subgroup of the group Möb$^+$ = Möb$^+(\mathbb{C})$ of all FLTs. Let

$$SL_2(\mathbb{R}) = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \mid a, b, c, d \in \mathbb{R}, ad - bc = 1 \right\}$$

be the special linear group of real 2×2 matrices with determinant 1. The rule

$$\psi \colon SL_2(\mathbb{R}) \to \text{Möb}^+(\mathbb{H})$$

sending $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ to $m(z) = (az + b)/(cz + d)$ is a surjective group homomorphism, with kernel $\ker(\psi) = \{I, -I\}$ equal to the center of $SL_2(\mathbb{R})$. We get a group isomorphism

$$\bar{\psi} \colon PSL_2(\mathbb{R}) \xrightarrow{\cong} \text{Möb}^+(\mathbb{H})$$

where $PSL_2(\mathbb{R}) = SL_2(\mathbb{R})/\{\pm I\}$ is the *projective special linear group* over $\mathbb{R}$.

Lemma 7.3. *Given two $\mathbb{H}$-lines L and L', there exists an FLT m preserving $\mathbb{H}$ such that $m(L) = L'$. In other words,* Möb$^+(\mathbb{H})$ *acts transitively on the set of $\mathbb{H}$-lines in $\mathbb{H}$.*

Proof. Let C and C' be the $\bar{\mathbb{C}}$-circles containing L and L', respectively. These meet $\bar{\mathbb{R}}$ orthogonally, say at $\{z_1, z_2\}$ and $\{w_1, w_2\}$, respectively. Choose points $z_3 \in L$ and $w_3 \in L'$. There is then a unique FLT $m(z) = (az + b)/(cz + d)$ with $m(z_1) = w_1$, $m(z_2) = w_2$ and $m(z_3) = w_3$, which satisfies $m(L) = L'$. It maps the $\bar{\mathbb{C}}$-line $\bar{\mathbb{R}}$, which meets C at right angles at z_1 and z_2, to a $\bar{\mathbb{C}}$-line $m(\bar{\mathbb{R}})$, which meets C' at right angles at w_1 and w_2. This implies that $m(\bar{\mathbb{R}}) = \bar{\mathbb{R}}$. Since $m(z_3) = w_3$, with both z_3 and w_3 in $\mathbb{H}$, it follows that $m(\mathbb{H}) = \mathbb{H}$, so m preserves $\mathbb{H}$. □

Remark 7.4. We will find similar results for FLTs that restrict to homeomorphisms $\mathbb{D} \to \mathbb{D}$ in Section 2.7.

7.2. **The cross-ratio.**

Definition 7.5. Given three distinct points $z_1, z_2, z_3 \in \bar{\mathbb{C}}$, the uniquely determined FLT m with $m(z_1) = 1$, $m(z_2) = 0$ and $m(z_3) = \infty$ is a function $\bar{\mathbb{C}} \to \bar{\mathbb{C}}$ that is denoted

$$m(z) = [z, z_1, z_2, z_3] .$$

The extended number $[z, z_1, z_2, z_3] \in \bar{\mathbb{C}}$ is called the *cross-ratio* of z, z_1, z_2 and z_3. (Another notation is $(z, z_1; z_2, z_3)$.)

Remark 7.6. In the real case, when z and z_1 lie between z_2 and z_3, the point z divides the segment from z_2 to z_3 into pieces of length $z - z_2$ and $z_3 - z$, with ratio $(z - z_2)/(z_3 - z)$. Likewise, z_1 divides the same segment into pieces of length $z_1 - z_2$ and $z_3 - z_1$, with ratio $(z_1 - z_2)/(z_3 - z_1)$. The quotient

$$\frac{z - z_2}{z_3 - z} : \frac{z_1 - z_2}{z_3 - z_1}$$

equals the cross-ratio $[z, z_1, z_2, z_3]$.

Lemma 7.7. *Given three distinct points z_1, z_2 and z_3 in $\bar{\mathbb{C}}$, and an FLT m, the cross-ratio satisfies*

$$[z, z_1, z_2, z_3] = [m(z), m(z_1), m(z_2), m(z_3)]$$

for each point $z \in \bar{\mathbb{C}}$. In particular, if m is the FLT with $m(z_1) = w_1$, $m(z_2) = w_2$ and $m(z_3) = w_3$, we have

$$[z, z_1, z_2, z_3] = [m(z), w_1, w_2, w_3] .$$

Proof. Let n be the FLT mapping $m(z_1)$, $m(z_2)$ and $m(z_3)$ to 1, 0 and ∞, in that order. Then $\ell = n \circ m$ is the FLT mapping z_1, z_2 and z_3 to 1, 0 and ∞. By definition, $[z, z_1, z_2, z_3] = \ell(z)$ and $[m(z), m(z_1), m(z_2), m(z_3)] = n(m(z))$. Hence these cross-ratios are equal. □

Remark 7.8. The cross-ratio defines a map

$$[-,-,-,-]\colon F_4(\bar{\mathbb{C}}) \to \bar{\mathbb{C}} \setminus \{1,0,\infty\} = \mathbb{C} \setminus \{0,1\}$$

from the configuration space $F_4(\bar{\mathbb{C}})$ of four distinct ordered points in $\bar{\mathbb{C}}$, to the complement of $\{0,1\}$ in $\mathbb{C}$. The group Möb^+ of FLTs acts freely on this configuration space, and the cross-ratio is constant on each orbit. Hence there is an induced bijection

$$F_4(\bar{\mathbb{C}})/\,\text{Möb}^+ \xrightarrow{\cong} \mathbb{C} \setminus \{0,1\}$$

from the orbit space for the action of Möb^+ on $F_4(\bar{\mathbb{C}})$, taking the orbit

$$\{(m(z), m(z_1), m(z_2), m(z_3)) \mid m \in \text{Möb}^+\}$$

to the cross-ratio $[z, z_1, z_2, z_3]$. It follows that any function of four distinct points in $\bar{\mathbb{C}}$, assumed to be invariant under the action by FLTs, can be expressed in terms of the cross-ratio. In this sense, the cross-ratio is the universal invariant of configurations of four distinct ordered points in $\bar{\mathbb{C}}$, with respect to the action by holomorphic Möbius transformations.

[[See Proposition 2.2.10 for more about the cross-ratio.]]

7.3. **Möbius transformations.** There are more $\bar{\mathbb{C}}$-circle preserving homeomorphisms $\bar{\mathbb{C}} \to \bar{\mathbb{C}}$ than the FLTs. One such homeomorphism σ is given by complex conjugation: $\sigma(z) = \bar{z}$ for $z \in \mathbb{C}$ and $\sigma(\infty) = \infty$. Another is given by reflection in the imaginary axis: $\rho(z) = -\bar{z}$ for $z \in \mathbb{C}$ and $\rho(\infty) = \infty$.

Definition 7.9. A *complex Möbius transformation* is a homeomorphism $\bar{\mathbb{C}} \to \bar{\mathbb{C}}$ given by one of the formulas

$$m(z) = \frac{az+b}{cz+d} \qquad \text{or} \qquad n(z) = \frac{a\bar{z}+b}{c\bar{z}+d}$$

where $a, b, c, d \in \mathbb{C}$ and $ad - bc \neq 0$. We write Möb = Möb($\mathbb{C}$) for the set of all complex Möbius transformations.

Lemma 7.10. *Each Möbius transformation is either holomorphic, and equals an FLT $m(z) = (az+b)/(cz+d)$, or it is anti-holomorphic, and of the form $n(z) = (a\bar{z}+b)/(c\bar{z}+d)$. In each case we may assume that $ad - bc = 1$.*

Proof. For the first claim, the key thing to check is that $\sigma(z) = \bar{z}$ is not holomorphic (but anti-holomorphic). For the second claim, note that we can divide a, b, c and d by one of the complex square roots of $ad - bc$. □

Lemma 7.11. Möb *is the group of homeomorphisms $\bar{\mathbb{C}} \to \bar{\mathbb{C}}$ generated by the FLTs and complex conjugation.* Möb^+ *is a subgroup of index two, and there is a split extension*

$$1 \to \text{Möb}^+ \to \text{Möb} \to \{e, \sigma\} \to 1$$

with σ acting on Möb^+ *by taking $m(z) = (az+b)/(cz+d)$ to $m^\sigma(z) = (\bar{a}z+\bar{b})/(\bar{c}z+\bar{d})$.*

Proof. It is clear that each complex Möbius transformation can be written as m or $n = m \circ \sigma$, where $m(z) = (az+b)/(cz+d)$ is an FLT and σ is complex conjugation. We must prove that Möb is closed under composition. This follows from the relation $\sigma \circ m = m^\sigma \circ \sigma$. □

Remark 7.12. It is possible to prove that Möb is the whole group of $\bar{\mathbb{C}}$-circle preserving homeomorphisms $\bar{\mathbb{C}} \to \bar{\mathbb{C}}$. See e.g. Theorem 2.21 on page 51 in J. W. Anderson, Hyperbolic Geometry, Springer-Verlag (2007).

Definition 7.13. Let Möb($\mathbb{H}$) be the group of Möbius transformations that restrict to homeomorphisms of $\mathbb{H}$. It is a subgroup of the full group Möb = Möb($\mathbb{C}$) of Möbius transformations, and contains the group $\text{Möb}^+(\mathbb{H})$, of FLTs preserving $\mathbb{H}$, as a subgroup of index two.

Lemma 7.14. *There is a split extension*

$$1 \to \text{Möb}^+(\mathbb{H}) \to \text{Möb}(\mathbb{H}) \to \{e, \rho\} \to 1$$

with ρ acting on $\text{Möb}^+(\mathbb{H})$ *by taking $m(z) = (az+b)/(cz+d)$ (with $a, b, c, d \in \mathbb{R}$, $ad - bc = 1$) to $m^\rho(z) = (az - b)/(-cz + d)$.*

Proof. The key facts are that $\rho \in \text{Möb}(\mathbb{H})$ is not in $\text{Möb}^+(\mathbb{H})$, and the relation $\rho \circ m = m^\rho \circ \rho$ holds:

$$-\frac{a\bar{z}+b}{c\bar{z}+d} = \frac{a(-\bar{z})-b}{-c(-\bar{z})+d}\,.$$

□

Definition 7.15. We sometimes write $\text{Möb}^- = \text{Möb}^-(\mathbb{C})$ for the coset $\text{Möb} \setminus \text{Möb}^+$ of anti-holomorphic (or odd) Möbius transformations. Similarly, we write $\text{Möb}^-(\mathbb{H})$ for the coset $\text{Möb}(\mathbb{H}) \setminus \text{Möb}^+(\mathbb{H})$ of anti-holomorphic (or odd) Möbius transformations preserving $\mathbb{H}$.

Proposition 7.16. *Each Möbius transformation preserving $\mathbb{H}$ can be written in one of the forms*

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

with a, b, c and $d \in \mathbb{R}$ and $ad - bc = 1$, or

$$m(z) = \frac{a\bar{z}+b}{c\bar{z}+d}$$

with a, b, c and $d \in \mathbb{R}$ and $ad - bc = -1$. In each case the presentations are unique, up to replacing (a, b, c, d) by $(-a, -b, -c, -d)$.

Proof. We already checked this in the holomorphic case. In the anti-holomorphic case we can write the Möbius transformation as $m = \rho \circ n$ with $n(z) = (az+b)/(cz+d)$ as in the first case. Then $m(z) = -(a\bar{z}+b)/(c\bar{z}+d) = ((-a)\bar{z} + (-b))/(c\bar{z}+d)$ with a, b, c and $d \in \mathbb{R}$ and $ad - bc = 1$. Replacing (a, b, c, d) with $(-a, -b, c, d)$ yields the claim. □

8. September 11th lecture

8.1. **Eigenvectors and fixed points.** The action of the group $GL_2(\mathbb{C})$ on $\bar{\mathbb{C}}$ by FLTs is closely related to the (usual) linear action of $GL_2(\mathbb{C})$ on

$$\mathbb{C}^2 = \{(z_1, z_2) \mid z_1, z_2 \in \mathbb{C}\}\,,$$

complex 2-dimensional space, given by

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} az_1 + bz_2 \\ cz_1 + dz_2 \end{bmatrix}.$$

For brevity we write this as $M \cdot (z_1, z_2) = (az_1 + bz_2, cz_1 + dz_2)$, where $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$.

Since $ad - bc \neq 0$, we have $M \cdot (z_1, z_2) \neq (0, 0)$ whenever $(z_1, z_2) \neq (0, 0)$. The resulting actions by $GL_2(\mathbb{C})$ on $\mathbb{C}^2 \setminus \{(0, 0)\}$ and $\bar{\mathbb{C}}$ are compatible under the projection

$$\pi \colon \mathbb{C}^2 \setminus \{(0, 0)\} \to \bar{\mathbb{C}}$$

that takes (z_1, z_2) to z_1/z_2, interpreted as ∞ when $z_2 = 0$. The compatibility means that

$$\pi(M \cdot (z_1, z_2)) = M \cdot \pi(z_1, z_2)\,,$$

or equivalently, that

$$\frac{az_1 + bz_2}{cz_1 + dz_2} = \frac{a(z_1/z_2) + b}{c(z_1/z_2) + d} = m(z_1/z_2)$$

where $m(z) = (az + b)/(cz + d)$ is the FLT associated to M.

We can analyze the action of $M \in GL_2(\mathbb{C})$ on $\mathbb{C}^2$ in terms of eigenvectors and eigenvalues. If (z_1, z_2) is an eigenvector for M, with eigenvalue λ, we have $M \cdot (z_1, z_2) = (\lambda z_1, \lambda z_2)$. Here $\lambda \neq 0$, since M is invertible. Then

$$m(z) = \pi(M \cdot (z_1, z_2)) = \pi(\lambda z_1, \lambda z_2) = \lambda z_1/\lambda z_2 = z_1/z_2 = z\,,$$

where $z = \pi(z_1, z_2)$. In other words, $z \in \bar{\mathbb{C}}$ is a fixed point for $m\colon \bar{\mathbb{C}} \to \bar{\mathbb{C}}$. The classification of M in terms of eigenvectors therefore corresponds to a classification of m in terms of fixed points.

8.2. **Classification of real FLTs.** We are principally interested in real FLTs (and Möbius transformations), preserving $\mathbb{H} \subset \bar{\mathbb{C}}$, hence will assume that a, b, c and d are real, and that $ad - bc = 1$. In other words, $M \in SL_2(\mathbb{R})$. The eigenvalues of M are then roots of the characteristic polynomial

$$\begin{aligned} p_M(t) &= \det(tI - M) = (t-a)(t-d) - bc \\ &= t^2 - (a+d)t + (ad - bc) = t^2 - (a+d)t + 1\,. \end{aligned}$$

Hence

$$\lambda = \frac{(a+d) \pm \sqrt{(a+d)^2 - 4}}{2}\,.$$

Since a, b, c and d are real, we can divide into three cases, according to the sign of $(a+d)^2 - 4$. In this discussion we assume that M is not $\pm I$, so that m is not the identity transformation.

8.2.1. *The hyperbolic case.* If $|a + d| > 2$ we say that M and m are of hyperbolic type. The matrix M has two real eigenvalues λ and λ', and two real eigenvectors (x_1, x_2) and (x'_1, x'_2), with $M \cdot (x_1, x_2) = (\lambda x_1, \lambda x_2)$ and $M \cdot (x'_1, x'_2) = (\lambda' x'_1, \lambda' x'_2)$. Since the eigenvalues are different, the eigenvectors are linearly independent, so by scaling one or both eigenvectors we may assume that $x_1 x'_2 - x'_1 x_2 = 1$.

The fixed points of the FLT $m(z) = (az + b)/(cz + d)$, i.e., the solutions of $m(z) = z$ with $z \in \bar{\mathbb{C}}$, are then the two points $x = \pi(x_1, x_2) = x_1/x_2$ and $x' = \pi(x'_1/x'_2) = x'_1/x'_2$ on the extended real line $\bar{\mathbb{R}}$.

Note that $\lambda + \lambda' = a + d$ and $\lambda\lambda' = 1$. Replacing M by $-M$, if necessary, we may assume that $a + d$ is positive, in which case λ and λ' are also positive. Interchanging λ and λ', if necessary, we may assume that $1 < \lambda < \infty$ and $0 < \lambda' = 1/\lambda < 1$.

Let $N = \begin{bmatrix} x_1 & x'_1 \\ x_2 & x'_2 \end{bmatrix}$ represent a change-of-basis, between the standard basis for $\mathbb{C}^2$ and the basis given by the eigenvectors, so that $M \cdot N = N \cdot L$, with $L = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda' \end{bmatrix}$ a diagonal matrix. Both N and L lie in $SL_2(\mathbb{R})$.

Let $n(z) = (x_1 z + x'_1)/(x_2 z + x'_2)$ and $\ell(z) = (\lambda z + 0)/(0z + \lambda') = \lambda^2 z$ be the associated FLTs in $\text{Möb}^+(\mathbb{H})$. Then $m \circ n = n \circ \ell$. Viewing the homeomorphism $n\colon \bar{\mathbb{C}} \to \bar{\mathbb{C}}$ as a change of coordinate system on $\bar{\mathbb{C}}$, the FLT $\ell(z) = \lambda^2 z$ (at the source of n) corresponds to the original FLT $m(z) = (az + b)/(cz + d)$ (at the target of n). The two fixed points 0 and ∞ of ℓ get mapped by n to the fixed points $n(0) = x$ and $n(\infty) = x'$ of m.

Up to conjugation by n, $m = n \circ \ell \circ n^{-1}$ is given by the standard hyperbolic FLT, $\ell(z) = \lambda^2 z$ with $1 < \lambda < \infty$. The sign of λ depends on the choice of M, but the value of $\eta = \lambda^2$ only depends on m.

8.2.2. *The parabolic case.* If $|a + d| = 2$ we say that M and m are of parabolic type. The matrix M has only one eigenvalue, $\lambda = (a + d)/2$, equal to 1 or -1. Replacing M with $-M$, if necessary, we may assume that $a + d$ is positive, in which case $\lambda = 1$.

If the eigenspace of M were 2-dimensional, M would be the identity matrix. Since we are assuming this not to be the case, the eigenspace of M must be 1-dimensional. Hence M has one real eigenvector (x_1, x_2) with $M \cdot (x_1, x_2) = (\lambda x_1, \lambda x_2)$, and m has one fixed point $x = x_1/x_2$, which lies in $\bar{\mathbb{R}}$.

Choose any real vector (x_1', x_2'), so that $x_1x_2' - x_1'x_2 = 1$, and let $N_1 = \begin{bmatrix} x_1 & x_1' \\ x_2 & x_2' \end{bmatrix}$. Then $M \cdot N_1 = N_1 \cdot L_1$, with N_1 and $L_1 = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}$ both in $SL_2(\mathbb{R})$. Here $\alpha = 1$, $\gamma = 0$ and $\delta = 0$, since $M \cdot (x_1, x_2) = (x_1, x_2)$ and $\det(L_1) = 1$. Hence $L_1 = \begin{bmatrix} 1 & \beta \\ 0 & 1 \end{bmatrix}$ with $\beta \in \mathbb{R}$. Let $n_1(z)$ and $\ell_1(z) = (1z + \beta)/(0z + 1) = z + \beta$ be the associated FLTs, with $m \circ n_1 = n_1 \circ \ell_1$.

In fact, we can standardize ℓ_1 a bit more. Let $k \neq 0$ and let $N_2 = \begin{bmatrix} k & 0 \\ 0 & 1/k \end{bmatrix}$. Then $L_1 \cdot N_2 = N_2 \cdot L$ with N_2 and $L = \begin{bmatrix} 1 & \beta/k^2 \\ 0 & 1 \end{bmatrix}$ in $SL_2(\mathbb{R})$. Choosing k with $k^2 = |\beta|$ we may thus arrange that $L = \begin{bmatrix} 1 & \pm 1 \\ 0 & 1 \end{bmatrix}$. Let $n_2(z) = (kz + 0)/(0z + (1/k)) = k^2 z$ and $\ell(z) = (1z \pm 1)/(0z + 1) = z \pm 1$ be the associated FLTs in $\text{Möb}^+(\mathbb{H})$, with $\ell_1 \circ n_2 = n_2 \circ \ell$.

Combining the two steps, let $N = N_1 \cdot N_2$, with associated FLT $n = n_1 \circ n_2$ in $\text{Möb}^+(\mathbb{H})$. Then $m \circ n = n \circ \ell$. Viewing $n \colon \bar{\mathbb{C}} \to \bar{\mathbb{C}}$ as a change of coordinate system, the FLT $\ell(z) = z \pm 1$ (at the source of n) corresponds to the FLT $m(z)$ (at the target of n). The single fixed point $z = \infty$ of ℓ gets mapped by n to the single fixed point $n(\infty) = x$ of m, which lies in $\bar{\mathbb{R}}$.

8.2.3. *The elliptic case.* If $|a + d| < 2$ we say that M and m are of elliptic type. The matrix M has a conjugate pair of complex eigenvalues, λ and $\bar{\lambda}$, and two complex eigenvectors (w_1, w_2) and $(\bar{w}_1, \bar{w}_2)$, with $M \cdot (w_1, w_2) = (\lambda w_1, \lambda w_2)$ and $M \cdot (\bar{w}_1, \bar{w}_2) = (\bar{\lambda}\bar{w}_1, \bar{\lambda}\bar{w}_2)$. (In each case the bar denotes complex conjugation).

The fixed points of the FLT $m(z) = (az + b)/(cz + d)$ acting on $\bar{\mathbb{C}}$ are then $w = w_1/w_2$ and $\bar{w} = \bar{w}_1/\bar{w}_2$. Interchanging (w_1, w_2) and $(\bar{w}_1, \bar{w}_2)$, if necessary, we may assume that $\operatorname{Im} w > 0$, so that $w \in \mathbb{H}$ is in the upper half-plane. Then the restriction of m to $\mathbb{H}$ has exactly one fixed point, namely w.

Let N be such that $n^{-1}(z) = (z - \operatorname{Re} w)/\operatorname{Im} w$ in $\text{Möb}^+(\mathbb{H})$, and let $L = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}$ be such that $M \cdot N = N \cdot L$. We can arrange that N, hence also L, lies in $SL_2(\mathbb{R})$. Let ℓ be the associated FLT, with $m \circ n = n \circ \ell$. Then $n^{-1}(w) = i$, so $w = n(i)$ and $\ell(i) = i$. Under the change of coordinates $n \colon \bar{\mathbb{C}} \to \bar{\mathbb{C}}$, the FLT ℓ (at the source of n) corresponds to the FLT m (at the target of n). The single fixed point $z = i$ in $\mathbb{H}$ of ℓ gets mapped to the single fixed point $n(i) = w$ in $\mathbb{H}$ of m.

From $\ell(i) = i$ we get $i\alpha + \beta = i\delta - \gamma$, so $\alpha = \delta$ and $\beta = -\gamma$. From $\alpha\delta - \beta\gamma = 1$ we get $\alpha^2 + \beta^2 = 1$, so there exists a unique angle $\theta \in [0, 2\pi)$ with $\alpha = \cos\theta$ and $\beta = \sin\theta$. Then $L = R_\theta = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$ represents clockwise rotation through the angle θ in $\mathbb{R}^2$, around the origin, while the associated FLT

$$\ell(z) = r_\theta(z) = \frac{(\cos\theta)z + \sin\theta}{(-\sin\theta)z + \cos\theta}$$

represents a "hyperbolic rotation" through an angle 2θ in the counter-clockwise direction of the upper half-plane $\mathbb{H}$, around the fixed point i. To see this, recall that

$$r_\theta'(z) = \frac{1}{((-\sin\theta)z + \cos\theta)^2} \,,$$

so $r_\theta'(i) = (\cos\theta + i\sin\theta)^2 = e^{i2\theta}$. Hence r_θ maps (differentiable) curves through i to (differentiable) curves through $r_\theta(i) = i$, and acts on their tangent vectors by multiplication by $r_\theta'(i) = e^{i2\theta}$, i.e., by rotation through the angle 2θ.

The cases $\theta = 0$ and $\theta = \pi$ are excluded by our assumption $M \neq \pm I$. Notice that $R_{\theta+\pi} = -R_\theta$, so $r_{\theta+\pi} = r_\theta$. We may therefore assume that $\theta \in (0, \pi)$. The eigenvalues of M are the same as the eigenvalues of $L = R_\theta$, i.e., $\lambda = \cos\theta + i\sin\theta$ and $\bar{\lambda} = \cos\theta - i\sin\theta$.

We summarize our findings in the following statement. (The "exactly one" uniqueness statements are left as exercises.)

Proposition 8.1. *Let $m(z) = (az+b)/(cz+d)$ in $\text{Möb}^+(\mathbb{H})$ be an FLT preserving $\mathbb{H}$, with a, b, c and $d \in \mathbb{R}$ and $ad - bc = 1$. Assume that $m \neq e$ is not the identity.*

(a) If $|a+d| > 2$ then m is of hyperbolic type, and is conjugate in $\text{Möb}^+(\mathbb{H})$ to exactly one

$$\ell(z) = \eta z$$

with $1 < \eta < \infty$. In this case m has exactly two fixed points, both of which lie in $\bar{\mathbb{R}}$.

(b) If $|a+d| = 2$ then m is of parabolic type, and is conjugate in $\text{Möb}^+(\mathbb{H})$ to exactly

$$\textit{one of } z \mapsto z+1 \textit{ and } z \mapsto z-1 \ .$$

In this case m has only one fixed point, which lies in $\bar{\mathbb{R}}$.

(c) If $|a+d| < 2$ then m is of elliptic type, and is conjugate in $\text{Möb}^+(\mathbb{H})$ to exactly one

$$\ell(z) = r_\theta(z) = \frac{(\cos\theta)z + \sin\theta}{(-\sin\theta)z + \cos\theta}$$

with $0 < \theta < \pi$. In this case m has exactly two fixed points, one in $\mathbb{H}$ and the other being its complex conjugate. □

[[The eigenvalues of M can be normalized as $\{\lambda, 1/\lambda\}$ for λ in the open ray $1 < \lambda < \infty$ in the hyperbolic case, as $\{1\}$ in the parabolic case, and as $\{\lambda, 1/\lambda\}$ for λ in the open semicircle $|\lambda| = 1$, $\operatorname{Im} z > 0$ in the elliptic case. Visualize hyperbolic, parabolic and elliptic FLTs, as in Figures 2.3.2, 2.3.1 and 2.3.3, respectively. Introduce the axis of a hyperbolic FLT, the horocycles of a parabolic FLT, and the "center" of an elliptic FLT.]]

Remark 8.2. The two standard parabolic transformations, $z \mapsto z+1$ and $z \mapsto z-1$, are conjugate in $\text{Möb}(\mathbb{H})$, since $\rho(z+1) = \rho(z) - 1$, where $\rho(z) = -\bar{z}$. Hence there is only one conjugacy class of Möbius transformations of parabolic type.

The two standard elliptic transformations r_θ and $r_{\pi-\theta}$ are also conjugate in $\text{Möb}(\mathbb{H})$, since $\rho \circ r_\theta = r_{\pi-\theta} \circ \rho$. Hence each conjugacy class of Möbius transformations of elliptic type contains r_θ for exactly one $\theta \in (0, \pi/2]$.

8.3. **Classification of real Möbius transformations.** It remains to classify the anti-holomorphic Möbius transformations that preserve $\mathbb{H}$.

Definition 8.3. For $p \in \mathbb{R}$, the Möbius transformation

$$m(z) = p - (\bar{z} - p) = 2p - \bar{z}$$

is given by *reflection* in the vertical line L given by $x = p$. The vector $m(z) - p$ is obtained from $z - p$ by reflection in the imaginary axis. The fixed point set of m is the line L, and m interchanges the two sides of the complement of L.

For $p \in \mathbb{R}$ and $r > 0$ the Möbius transformation

$$m(z) = p + r^2 \frac{z-p}{|z-p|^2} = p + \frac{r^2}{\bar{z} - p}$$

is given by *inversion* in the circle C given by $(x-p)^2 + y^2 = r^2$. The vector $m(z) - p$ is obtained from $z - p$ by scaling its length so that $|m(z) - p| \cdot |z - p| = r^2$. The fixed point set of m is the circle C, and m interchanges the inside and the outside of C.

In both cases we say that m is given by inversion in the corresponding $\mathbb{H}$-line, i.e., the part of L or C that lies in $\mathbb{H}$. Each inversion m satisfies $m \circ m = e$.

Proposition 8.4. *Let $m(z) = (a\bar{z} + b)/(c\bar{z} + d)$ in $\text{Möb}^-(\mathbb{H})$ be a Möbius transformation preserving $\mathbb{H}$, with a, b, c and $d \in \mathbb{R}$ and $ad - bc = -1$.*

(a) If $a + d = 0$, then m is an inversion, with fixed point set an $\mathbb{H}$-line. It is conjugate in $\text{Möb}(\mathbb{H})$ to the reflection $\rho(z) = -\bar{z}$ in the imaginary axis, as well as to the inversion $z \mapsto 1/\bar{z}$ in the unit circle.

(b) If $a + d \neq 0$, then m can be factored as $m = n \circ \ell$, where n is an inversion and n and ℓ commute ($n \circ \ell = \ell \circ n$). This factorization is unique, and ℓ is of hyperbolic type, with axis equal to the $\mathbb{H}$-line of inversion of n. In this case m has exactly two fixed points, both of which lie in $\bar{\mathbb{R}}$. □

The composites $m = n \circ \ell = \ell \circ n$ in case (b) are often called glide reflections: the hyperbolic transformation ℓ glides along its axis, and the inversion n reflects in that axis.

8.3.1. *The case of inversions.* Suppose that $a + d = 0$, so that $a = -d$. If $c = 0$ then $d \neq 0$ and $m(z) = (-d\bar{z} + b)/(0\bar{z} + d) = (b/d) - \bar{z}$ is reflection in the vertical line given by $x = b/2d$. If $c \neq 0$ then

$$m(z) = \frac{a\bar{z} + b}{c\bar{z} + d} = \frac{a}{c} + \frac{1/c^2}{\bar{z} - (a/c)}$$

is inversion in the circle with center a/c and radius $1/|c|$.

Let $n \in \text{Möb}^+(\mathbb{H})$ be an FLT preserving $\mathbb{H}$ that maps the imaginary axis $x = 0$ to the $\mathbb{H}$-line of inversion of m. Then $m \circ n = n \circ \ell$ where ℓ is an inversion fixing the imaginary axis, which implies that $\ell = \rho$.

[[To see that $\ell = n^{-1} \circ m \circ n$ is an inversion, note that the trace of $L = N^{-1} \cdot M \cdot N$ equals the trace of M. Alternatively use that the fixed point set is an $\mathbb{H}$-line in $\mathbb{H}$, not just two points in $\bar{\mathbb{R}}$, and compare with the composite case.]]

Alternatively, let n be an FLT preserving $\mathbb{H}$ that maps the unit circle $x^2 + y^2 = 1$ to the $\mathbb{H}$-line of inversion of m. Then $m \circ n = n \circ \ell$ where ℓ is an inversion fixing the unit circle, which implies that $\ell(z) = 1/\bar{z}$.

8.3.2. *The case of glide reflections.* Suppose, finally, that $a + d \neq 0$. We claim that in this case $m(z) = (a\bar{z}+b)/(c\bar{z}+d)$ has exactly two fixed points, both in $\bar{\mathbb{R}}$. If $c = 0$, $m(z) = (a\bar{z}+b)/d = z$ if and only if $z = b/(d - a)$ or $z = \infty$. If $c \neq 0$ then $m(\infty) = a/c$, so $z = x + iy$ is not ∞. The condition $m(z) = z$ is equivalent to $a\bar{z}+b = c|z|^2+dz$, or equivalently, to $ax+b = c(x^2+y^2)+dx$ and $-ay = dy$. The second equation implies $y = 0$, so $z = x$ is a root of $cx^2 + (d - a)x - b = 0$. Since $(d - a)^2 + 4bc = (a + d)^2 + 4 > 0$, this equation has precisely two real roots.

Let x_1 and x_2 be the two fixed points of m, both in $\bar{\mathbb{R}}$, and let L be the $\mathbb{H}$-line that intersects $\bar{\mathbb{R}}$ orthogonally at x_1 and x_2. Let n be the inversion on L. Then $\ell = n^{-1} \circ m$ and $\ell' = m \circ n^{-1}$ are both FLTs preserving $\mathbb{H}$, different from e, that fix x_1 and x_2. From the classification of FLTs preserving $\mathbb{H}$, it follows that both ℓ and ℓ' are of hyperbolic type, both with axis L. In fact $\ell|L = m|L = \ell'|L$, which implies that $\ell = \ell'$. Hence $m = n \circ \ell = \ell \circ n$, with n an inversion and ℓ hyperbolic.

[[See Proposition 2.3.3 for the proof of uniqueness.]]

9. September 16th lecture

9.1. Congruence in $\mathbb{H}$.

Definition 9.1. If z and w are two distinct points in $\mathbb{H}$, let $\overleftrightarrow{zw}$ be the uniquely determined *line* containing them, let $\overrightarrow{zw}$ be the *ray* from z containing w, and let $[z, w] = [w, z]$ be the *segment* between z and w.

If z, u and v are three distinct points in $\mathbb{H}$, with $\overleftrightarrow{zu}$ not equal to $\overleftrightarrow{zv}$, then let $\angle uzv$ be the *angle* with vertex z, consisting of the two rays $\overrightarrow{zu}$ and $\overrightarrow{zv}$.

Definition 9.2. (a) Let $[z, w]$ and $[z', w']$ be (hyperbolic line) segments in $\mathbb{H}$. We define $[z, w]$ to be *congruent* to $[z', w']$, written $[z, w] \cong [z', w']$, if and only if there exists a Möbius transformation $m \in \text{Möb}(\mathbb{H})$ preserving $\mathbb{H}$ such that $m([z, w]) = [z', w']$.

(b) Let $\angle uzv$ and $\angle u'z'v'$ be angles in $\mathbb{H}$. We define $\angle uzv$ to be *congruent* to $\angle u'z'v'$, written $\angle uzv \cong \angle u'z'v'$, if and only if there exists a Möbius transformation $m \in \text{Möb}(\mathbb{H})$ preserving $\mathbb{H}$ such that $m(\overrightarrow{zu}) = \overrightarrow{z'u'}$ and $m(\overrightarrow{zv}) = \overrightarrow{z'v'}$.

Definition 9.3. A hyperbolic line L meets $\bar{\mathbb{R}}$ in two points, p and q, say, which we call the *endpoints* of the line. If L is contained in the $\bar{\mathbb{C}}$-circle C, then $C \cap \bar{\mathbb{R}} = \{p, q\}$. The endpoints uniquely determine the line, and we write $L = (p, q) = (q, p)$.

A hyperbolic ray R meets $\bar{\mathbb{R}}$ in one point, q, say, which we call the *endpoint* of the ray. If $R = \overrightarrow{zw}$ is a ray with vertex z and endpoint q, then these points uniquely determine the ray, and we write $R = [z, q)$.

Lemma 9.4. *Given two lines $L = (p, q)$ and $L' = (p', q')$, and points $z \in L$ and $z' \in L'$, there exists a unique $m \in \text{Möb}^+(\mathbb{H})$ (an FLT preserving $\mathbb{H}$) such that $m(p) = p'$, $m(z) = z'$ and $m(q) = q'$. In particular, $m(L) = L'$.*

Proof. There is a unique FLT m with $m(p) = p'$, $m(z) = z'$ and $m(q) = q'$. In particular, $m(L) = L'$. Since $\bar{\mathbb{R}}$ is a $\bar{\mathbb{C}}$-circle that meets L at right angles at p and q, it follows that $m(\bar{\mathbb{R}})$ is a $\bar{\mathbb{C}}$-circle that meets L' at right angles at p' and q'. Since $\bar{\mathbb{R}}$ is the unique $\bar{\mathbb{C}}$-circle that meets L' at right angles at p' and q', it follows that $m(\bar{\mathbb{R}}) = \bar{\mathbb{R}}$. Since $m(z) = z'$ it follows that $m(\mathbb{H}) = \mathbb{H}$, so $m \in \text{Möb}^+(\mathbb{H})$. □

Corollary 9.5. *Given two rays $R = [z, q)$ and $R' = [z', q')$ there is a unique $m \in \text{Möb}^+(\mathbb{H})$ such that $m(R) = R'$. In particular, $m(z) = z'$ and $m(q) = q'$. In other words, $\text{Möb}^+(\mathbb{H})$ acts simply transitively on the set of rays in $\mathbb{H}$.*

Proof. The rays R and R' are contained in unique lines $L = (p, q)$ and $L' = (p', q')$, respectively. Let m be the unique FLT preserving $\mathbb{H}$ with $m(p) = p'$, $m(z) = z'$ and $m(q) = q'$. Then $m(R) = R'$, as desired. □

Lemma 9.6. *(a) An FLT preserving $\mathbb{H}$ is uniquely determined by its values at two distinct points. In other words, if $z \neq w \in \mathbb{H}$ and $m, n \in \text{Möb}^+(\mathbb{H})$ satisfy $m(z) = n(z)$ and $m(w) = n(w)$, then $m = n$.*

(b) If $[z, w] \cong [z', w']$ then there exists a unique FLT m preserving $\mathbb{H}$ with $m(z) = z'$ and $m(w) = w'$.

Proof. (a) Let $\overleftrightarrow{zw} = L = (p, q)$, with $p * z * w$ and $z * w * q$. The (hyperbolic) lines $m(L)$ and $n(L)$ both contain $m(z) = n(z)$ and $m(w) = n(w)$, hence are equal. Thus the sets of endpoints, $\{m(p), m(q)\}$ and $\{n(p), n(q)\}$, are also equal. Since m and n preserve betweenness, we must have $m(p) * m(z) * m(w)$ and $n(p) * n(z) * n(w)$, which implies that $m(p) = n(p)$ a and $m(q) = n(q)$. Hence the FLTs m and n agree on four points (p, z, w and q), and must be equal.

(b) By the definition of congruence, there exists a Möbius transformation m preserving $\mathbb{H}$ with $m([z, w]) = [z', w']$. If $m \in \text{Möb}^-(\mathbb{H})$, we may precompose m with inversion in the $\mathbb{H}$-line $\overleftrightarrow{zw}$, to arrange that $m \in \text{Möb}^+(\mathbb{H})$.

If $m(z) = w'$ and $m(w) = z'$ we may precompose m with an FLT preserving $\mathbb{H}$ and interchanging z and w. To prove that such an interchanging FLT exists, choose an FLT n preserving $\mathbb{H}$ and taking the positive imaginary axis to the line $\overleftrightarrow{zw}$. Then $n(is) = z$ and $n(it) = w$ for some $s, t > 0$. Let $\ell(u) = -st/u$, for $u \in \mathbb{H}$. Then ℓ interchanges is and it, and $n \circ \ell \circ n^{-1}$ is an FLT preserving $\mathbb{H}$ and interchanging z and w. This way we may arrange that $m(z) = z'$ and $m(w) = w'$.

Uniqueness of this m is clear from part (a). □

Lemma 9.7. *Given two rays R_1 and R_2, with common vertex z, there is a unique inversion n such that $n(R_1) = R_2$. (This implies that $n(z) = z$ and $n(R_2) = R_1$.)*

Proof. Omitted. See Lemma 2.4.5 in Jahren's book. □

9.2. **Verification of Hilbert's congruence axioms.**

Proposition 9.8. *Congruence of segments and angles in $\mathbb{H}$, as defined above, satisfies Hilbert's axioms C1 through C6:*

(C1) Given a segment $[z, w]$ and a ray R' with vertex z' there is a unique point w' on R' such that $[z, w] \cong [z', w']$.

(C2) $\cong$ is an equivalence relation on the set of segments.

(C3) If $u * v * w$, $u' * v' * w'$, $[u,v] \cong [u',v']$ and $[v,w] \cong [v',w']$, then $[u,w] \cong [u',w']$.

(C4) Given an angle $\angle uzv$ and a ray $R = [w,q)$, there are unique angles $\angle qwp_1$ and $\angle qwp_2$, on opposite sides of $\overleftrightarrow{wq}$, such that $\angle qwp_1 \cong \angle uzv \cong \angle qwp_2$.

(C5) $\cong$ is an equivalence relation on the set of angles.

(C6) Given triangles $\triangle uzv$ and $\triangle u'z'v'$, with $[z,u] \cong [z',u']$, $\angle uzv \cong \angle u'z'v'$ and $[z,v] \cong [z',v']$, we also have $[u,v] \cong [u',v']$, $\angle zuv \cong \angle z'u'v'$ and $\angle zvu \cong \angle z'v'u'$ (side-angle-side criterion for congruence).

Proof. (C1) Let $R = \overrightarrow{zw} = [z,q)$ be the ray containing $[z,w]$. Then there is a unique FLT m preserving $\mathbb{H}$ with $m(R) = R'$. In particular, $m(z) = z'$. Let $w' = m(w)$. Then $m([z,w]) = [z',w']$, so $[z,w] \cong [z',w']$.

Suppose that w'' is a second point on R' such that $[z,w] \cong [z',w'']$. There is a unique FLT n preserving $\mathbb{H}$ such that $n(z) = z'$ and $n(w) = w''$. Then $\overrightarrow{z'w'} = R' = \overrightarrow{z'w''}$, so $m(R) = R'$ and $n(R) = R'$. By simple transitivity of the action of $\text{Möb}^+(\mathbb{H})$ on rays it follows that $m = n$, so $w' = m(w) = n(w) = w''$.

(C2) This is clear, since $\text{Möb}(\mathbb{H})$ is a group.

(C3) Let $m \in \text{Möb}^+(\mathbb{H})$ be such that $m(u) = u'$ and $m(v) = v'$. Let $w'' = m(w)$, so that $[u,w] \cong [u',w'']$ and $[v,w] \cong [v',w'']$. Then $u' * v' * w'$ and $u' * v' * w''$, so w'' lies on the ray $\overrightarrow{v'w'}$. Furthermore, $[v',w'] \cong [v,w] \cong [v',w'']$, so $w' = w''$ by the uniqueness in (C1). Hence $[u,w] \cong [u',w''] = [u',w']$.

(C4) Let $m \in \text{Möb}^+(\mathbb{H})$ be such that $m([z,u)) = [w,q)$, and let $n \in \text{Möb}^-(\mathbb{H})$ be obtained by postcomposing m with inversion in the line L containing $[w,q)$. Define p_1 and p_2 by $[w,p_1) = m([z,v))$ and $[w,p_2) = n([z,v))$. Then m exhibits the congruence $\angle uzv \cong \angle qwp_1$ and n exhibits the congruence $\angle uzv \cong \angle qwp_2$. The angles $\angle qwp_1$ and $\angle qwp_2$ lie on opposite sides of L, because m and n differ by inversion in that line.

[[For uniqueness, see the proof on pages 42–43 in Jahren's book, which uses Lemma 2.4.5.]]

(C5) This is clear, since $\text{Möb}(\mathbb{H})$ is a group.

(C6) Let $m \in \text{Möb}(\mathbb{H})$ be such that $m(\overrightarrow{zu}) = \overrightarrow{z'u'}$ and $m(\overrightarrow{zv}) = \overrightarrow{z'v'}$, realizing the congruence $\angle uzv \cong \angle u'z'v'$. Then $m(z) = z'$, $[z,u] \cong [z', m(u)]$ and $[z,v] \cong [z', m(v)]$, so by the uniqueness in (C1) we can deduce that $m(u) = u'$ and $m(v) = v'$. Hence m maps $\triangle uzv$ to $\triangle u'z'v'$ and the triangles as congruent. □

10. September 18th lecture

10.1. **Distance in $\mathbb{H}$.** We now introduce a notion of distance between points in $\mathbb{H}$, so that two segments are congruent if and only if they have the same length, i.e., the same distance between their endpoints.

Definition 10.1. A *metric* on $\mathbb{H}$ is a function $d \colon \mathbb{H} \times \mathbb{H} \to \mathbb{R}$ such that

(D1) $d(z,w) \geq 0$ for all $z, w \in \mathbb{H}$, and $d(z,w) = 0$ if and only if $z = w$ (positivity),

(D2) $d(z,w) = d(w,z)$ for all $z, w \in \mathbb{H}$ (symmetry), and

(D3) $d(u,w) \leq d(u,v) + d(v,w)$ for all $u, v, w \in \mathbb{H}$ (triangle inequality).

We seek a metric on $\mathbb{H}$ that is invariant under Möbius transformations, so that:

(D5) $d(z,w) = d(z',w')$ if (and only if) there exists an $m \in \text{Möb}(\mathbb{H})$ with $m(z) = z'$ and $m(w) = w'$.

Given $z \neq w \in \mathbb{H}$ there is a unique $m \in \text{Möb}^+(\mathbb{H})$ such that $m(\overrightarrow{zw}) = [i, \infty)$, or equivalently, such that $m(z) = i$ and $m(w) = it$ for some $t > 1$. By condition (D5) we must then have $d(z,w) = d(i,it)$, so d is fully determined by the function

$$f \colon [1,\infty) \to [0,\infty)$$

given by $f(t) = d(i,it)$. Note that $f(1) = 0$. By the triangle inequality we must have $d(i,ist) \leq d(i,is) + d(is,ist)$ for all $s,t \geq 1$. If we add the condition that (hyperbolic line) segments are locally to be paths of shortest lengths, we can demand (D4) that $d(i,ist) = d(i,is) + d(is,ist)$, since is lies on the line segment $[i, ist]$, so the shortest path from i to ist should pass through is. Due to the existence of the hyperbolic FLT $z \mapsto sz$, we must have $d(i, it) = d(is, ist)$. Hence $f(st) = d(i, ist) = d(i, is) + d(is, ist) = d(i, is) + d(i, it) = f(s) + f(t)$ for all $s, t \geq 1$. It follows that f has the form $f(t) = C \ln(t)$ for some $C > 0$. It will be convenient to let $C = 1$. Note that $t = |it| = |m(w)|$, with the notations above.

Definition 10.2. For $z, w \in \mathbb{H}$ let

$$d(z, w) = \ln(|m(w)|)\,,$$

where $m \in \text{Möb}^+(\mathbb{H})$ is chosen so that $m(z) = i$ and $m(w) = it$, for some $t \geq 1$.

Proposition 10.3. *The function $d\colon \mathbb{H} \times \mathbb{H} \to \mathbb{R}$ defines a metric on $\mathbb{H}$.*

Proof. The function is clearly well defined, and (D1) is obvious.

To prove (D2), let $m \in \text{Möb}^+(\mathbb{H})$ be such that $m(z) = i$ and $m(w) = it$ with $t \geq 1$, and let $\ell \in \text{Möb}^+(\mathbb{H})$ be given by $\ell(u) = -t/u$. Then $\ell(i) = it$ and $\ell(it) = i$, so $n = \ell \circ m \in \text{Möb}^+(\mathbb{H})$ satisfies $n(w) = i$ and $n(z) = it$. Hence $d(w, z) = \ln(|n(z)|) = \ln(t) = \ln(|m(w)|) = d(z, w)$, as required.

We postpone the proof of the triangle inequality, (D3), after the following three lemmas. □

Lemma 10.4. *Let $z, w, z', w' \in \mathbb{H}$. Then $d(z, w) = d(z', w')$ if and only if there exists an $m \in \text{Möb}(\mathbb{H})$ with $m(z) = z'$ and $m(w) = w'$.*

Proof. If $d(z, w) = d(z', w')$ then there exist $n, n' \in \text{Möb}^+(\mathbb{H})$ with $n(z) = i = n'(z')$ and $n(w) = it = n'(w')$, for some $t \geq 1$. (This uses that ln is strictly increasing, so that $\ln(t) \geq 0$ determines $t \geq 1$.) Then $m = (n')^{-1} \circ n \in \text{Möb}^+(\mathbb{H})$ satisfies $m(z) = z'$ and $m(w) = w'$.

Conversely, if $m(z) = z'$ and $m(w) = w'$ for some $m \in \text{Möb}(\mathbb{H})$, we may assume that $m \in \text{Möb}^+(\mathbb{H})$ by, if necessary, precomposing m with inversion in the line $\overleftrightarrow{zw}$. If $n \in \text{Möb}^+(\mathbb{H})$ is an FLT preserving $\mathbb{H}$ such that $n(z) = i$ and $n(w) = it$, for some $t \geq 1$, then $n' = n \circ m^{-1}$ must be an FLT preserving $\mathbb{H}$ such that $n'(z') = i$ and $n'(w') = is$, for some $s \geq 1$. Here $is = n'(w') = n(m^{-1}(w')) = n(w) = it$, so $s = t$. Hence $d(z, w) = \ln(|n(w)|) = \ln(|n'(w')|) = d(z', w')$, as required. □

Lemma 10.5. *Given $z \neq w \in \mathbb{H}$ let p and q be the endpoints of the line through z and w, so that $\overleftrightarrow{zw} = (p, q)$, with $p * z * w$ and $z * w * q$. The FLT m preserving $\mathbb{H}$ that maps z to $m(z) = i$ and w to $m(w) = it$ with $t > 1$, is given by the formula*

$$m(u) = i \cdot \frac{u - p}{u - q} \cdot \frac{z - q}{z - p}\,.$$

Hence

$$d(z, w) = \ln\Big(\frac{w - p}{w - q} \cdot \frac{z - q}{z - p}\Big)\,.$$

Proof. The unique FLT sending z, p and q to 1, 0 and ∞, respectively, is given by

$$n(u) = \frac{u - p}{u - q} \cdot \frac{z - q}{z - p}$$

(extended as usual, if $p = \infty$ or $q = \infty$). It takes w to a point t between 1 and ∞. Now let $m(u) = i \cdot n(u)$. This FLT takes p, z, w and q to 0, i, it and ∞. Hence it takes $\bar{\mathbb{R}}$, which meets (p, q) orthogonally at p and q, to a $\bar{\mathbb{C}}$-circle that meets the imaginary axis orthogonally at 0 and at ∞. The latter $\bar{\mathbb{C}}$-circle must be $\bar{\mathbb{R}}$, so $m(\bar{\mathbb{R}}) = \bar{\mathbb{R}}$. Since $m(z) \in \mathbb{H}$, it follows that $m(\mathbb{H}) = \mathbb{H}$, so $m \in \text{Möb}^+(\mathbb{H})$. □

Lemma 10.6. *For $z, w \in \mathbb{H}$ we have*

$$d(z, w) \geq |\ln\Big(\frac{\operatorname{Im} w}{\operatorname{Im} z}\Big)|\,,$$

with equality if and only if $\operatorname{Re} z = \operatorname{Re} w$.

Proof. Note that if we interchange z and w, the fraction $\operatorname{Im} w/\operatorname{Im} z$ becomes inverted, so $\ln(\operatorname{Im} w/\operatorname{Im} z)$ changes sign. Due to the absolute value operator on the right hand side, both sides of the inequality remain unchanged after this interchange. We may therefore assume that $\operatorname{Im} z \le \operatorname{Im} w$, so that $\ln(\operatorname{Im} w/\operatorname{Im} z)$ is non-negative (and the absolute value sign is superfluous).

If $\operatorname{Re} z = \operatorname{Re} w$ we have $p = \operatorname{Re} z = \operatorname{Re} w$ and $q = \infty$, so

$$d(z,w) = \ln\Big(\frac{w-p}{z-p}\Big) = \ln\Big(\frac{i \operatorname{Im} w}{i \operatorname{Im} z}\Big) = \ln\Big(\frac{\operatorname{Im} w}{\operatorname{Im} z}\Big) .$$

Otherwise, $\operatorname{Re} z \ne \operatorname{Re} w$ and p and q are finite. The (strict) inequality is trivially satisfied if $\operatorname{Im} z = \operatorname{Im} w$, so we may assume that $\operatorname{Im} z < \operatorname{Im} w$.

The positive real number $(w-p)/(w-q) \cdot (z-q)/(z-p)$ equals its own modulus, so

$$d(z,w) = \ln\Big(\Big|\frac{w-p}{w-q}\Big| \cdot \Big|\frac{z-q}{z-p}\Big|\Big) = \ln\Big(\Big|\frac{w-p}{w-q}\Big| / \Big|\frac{z-p}{z-q}\Big|\Big) .$$

Let $A = \angle pqz$ and $B = \angle pqw$. Then

$$\Big|\frac{z-p}{z-q}\Big| = \tan A = \frac{\operatorname{Im} z}{|\operatorname{Re} z - q|}$$

and

$$\Big|\frac{w-p}{w-q}\Big| = \tan B = \frac{\operatorname{Im} w}{|\operatorname{Re} w - q|} .$$

Since $p * z * w * q$ we have $|\operatorname{Re} z - q| > |\operatorname{Re} w - q|$, so

$$\begin{aligned} d(z,w) &= \ln(\tan B/\tan A) = \ln\Big(\frac{\operatorname{Im} w}{|\operatorname{Re} w - q|} / \frac{\operatorname{Im} z}{|\operatorname{Re} z - q|}\Big) \\ &= \ln\Big(\frac{|\operatorname{Re} z - q|}{|\operatorname{Re} w - q|} \cdot \frac{\operatorname{Im} w}{\operatorname{Im} z}\Big) > \ln\Big(\frac{\operatorname{Im} w}{\operatorname{Im} z}\Big) . \end{aligned}$$

□

Proof of (D3) and (D4). Applying a Möbius transformation that maps u to i and w to it, it will suffice to consider the special case when $u = i$ and $w = it$, with $t > 1$. Then

$$\begin{aligned} d(u,v) + d(v,w) &\ge |\ln\Big(\frac{\operatorname{Im} v}{\operatorname{Im} u}\Big)| + |\ln\Big(\frac{\operatorname{Im} w}{\operatorname{Im} v}\Big)| \ge |\ln\Big(\frac{\operatorname{Im} v}{\operatorname{Im} u}\Big) + \ln\Big(\frac{\operatorname{Im} w}{\operatorname{Im} v}\Big)| \\ &= |\ln\Big(\frac{\operatorname{Im} v}{\operatorname{Im} u} \cdot \frac{\operatorname{Im} w}{\operatorname{Im} v}\Big)| = |\ln\Big(\frac{\operatorname{Im} w}{\operatorname{Im} u}\Big)| = d(u,w) . \end{aligned}$$

The last equality uses $\operatorname{Re} u = \operatorname{Re} w$, and this proves (D3).

We have equality if and only if $\operatorname{Re} u = \operatorname{Re} v = \operatorname{Re} w$ and if $\ln(\operatorname{Im} v/\operatorname{Im} u)$ and $\ln(\operatorname{Im} w/\operatorname{Im} v)$ have the same sign. This is equivalent to asking that $v \in [u,w]$, proving the strong form of (D4) asserting that $d(u,w) = d(u,v) + d(v,w)$ if and only if $v \in [u,w]$. □

11. September 23rd lecture

11.1. Angle measure in $\mathbb{H}$.

We also wish to introduce a numerical measure for hyperbolic angles in $\mathbb{H}$, so that

(A1) the angle measure takes positive real values,

(A2) congruent angles have the same angle measure,

(A3) the angle measure of $\angle uzw$ is the sum of the angle measures of $\angle uvz$ and of $\angle vzw$, when v is inside $\angle uzw$,

(A4) the angle measure of a right angle is equal to $\pi/2 = 90°$.

Here z, u, v and w are points in $\mathbb{H}$, and $\angle uzv$ is defined to be a right angle if it is congruent to its supplementary angle, i.e., the angle $\angle vzw$ where $u * z * w$.

These conditions will determine the angle measure, if it exists. For instance, if $\angle uzw$ is a right angle, if v is inside $\angle uzw$, and $\angle uzv \cong \angle vzw$, then the angle measure of $\angle uzv$ must be equal to $\pi/4 = 45°$. Continuing this way, the measure of a dense set of angles is determined by these conditions, and the measure of the remaining angles is determined by continuity/monotonicity.

It follows that the angle measure takes values in the interval $(0, \pi) = (0°, 180°)$, and that all such values are realized.

The fact that congruence in $\mathbb{H}$ is generated by Möbius transformations, and that these are conformal, i.e., preserve the Euclidean angle measure of the containing space $\mathbb{C}$, will enable us to simply transport the Euclidean angle measure in $\mathbb{C}$ to the (uniquely determined) hyperbolic angle measure in $\mathbb{H}$. We simply define hyperbolic angle between two hyperbolic rays $\overrightarrow{zu}$ and $\overrightarrow{zv}$ to be equal to the Euclidean angle between these two Euclidean curves, oriented from the vertex z and heading in the direction of u and v, respectively.

Proposition 11.1. *The hyperbolic angle measure satisfies (A1) through (A4). Furthermore, two hyperbolic angles are congruent if (and only if) they have the same angle measure.*

Proof. (A1) The Euclidean angle between $\overrightarrow{zu}$ and $\overrightarrow{zv}$ is evidently non-negative. It would be zero if the $\bar{\mathbb{C}}$-circles containing these hyperbolic rays were tangent at z, but since these $\bar{\mathbb{C}}$-circles meet $\bar{\mathbb{R}}$ at right angles that only happens if the $\bar{\mathbb{C}}$-circles are equal, in which case $\overrightarrow{zu}$ and $\overrightarrow{zv}$ lie on the same line, and the angle $\angle uzv$ is not defined..

(A2) If the Möbius transformation m maps $\angle uzv$ to $\angle u'z'v'$, then the Euclidean angle between $\overrightarrow{zu}$ and $\overrightarrow{zv}$ is equal to the Euclidean angle between $\overrightarrow{z'u'}$ and $\overrightarrow{z'v'}$. This is simply the fact that m is conformal. Hence the hyperbolic angle measure of $\angle uzv$ is equal to the hyperbolic angle measure of $\angle u'z'v'$.

(A3) This is clear from the Euclidean case.

(A4) If $\angle uzv$ is a hyperbolic right angle, and $u * z * w$ in $\mathbb{H}$, then $\angle uzv$ is congruent to $\angle vzw$, so the Euclidean angle A between $\overrightarrow{zu}$ and $\overrightarrow{zv}$ has the same measure as the Euclidean angle B between $\overrightarrow{zv}$ and $\overrightarrow{zw}$. Since $\overrightarrow{zu}$ and $\overrightarrow{zw}$ lie on the same $\mathbb{H}$-line, and have the same Euclidean tangent line at z, it follows that these two Euclidean angles, A and B, are supplementary, and add to $\pi = 180°$. Hence A is a right angle, and the hyperbolic measure of $\angle uzv$ is $\pi/2 = 90°$, as claimed.

It remains to prove that if $\angle uzv$ and $\angle u'z'v'$ have the same angle measure, then they are congruent. Choose Möbius transformations n and n' mapping $\overrightarrow{zu}$ and $\overrightarrow{z'u'}$ to $[i, \infty)$. Then $n(\overrightarrow{zv})$ and $n'(\overrightarrow{z'v'})$ meet $[i, \infty)$ at the same Euclidean angle. Hence they either agree, or can be mapped to one another by the reflection $\rho(z) = -\bar{z}$ in the imaginary axis. In the first case, $m = (n')^{-1} \circ n$ is a Möbius transformation showing that $\angle uzv$ and $\angle u'z'v'$ are congruent. In the second case, $m = (n')^{-1} \circ \rho \circ n$ achieves the same purpose. □

12. September 25th lecture

12.1. **Poincaré's disc model $\mathbb{D}$.** We now translate our definitions and results about the upper half-plane model $\mathbb{H} = \{z \in \mathbb{C} \mid \operatorname{Im} z > 0\}$ for the hyperbolic plane to definitions and results about the unit disc model $\mathbb{D} = \{z \in \mathbb{C} \mid |z| < 1\}$. We do this using the bijection $\xi \colon \mathbb{D} \to \mathbb{H}$, defined as the composite of inverse stereographic projection $\Psi \colon \mathbb{D} \to \mathbb{B}$, the rotation $(x, y, z) \mapsto (x, -z, y)$ from $\mathbb{B}$ to $\mathbb{B}'$, and stereographic projection $\Phi \colon \mathbb{B}' \to \mathbb{H}$. It equals the FLT that maps 1, $-i$ and i to 1, 0 and ∞, respectively, and is given by the formula

$$\xi(z) = \frac{z+i}{1+i} \cdot \frac{1-i}{z-i} = \frac{z+i}{iz+1}\,.$$

Its inverse, $\xi^{-1} \colon \mathbb{H} \to \mathbb{D}$, is given by the formula

$$\xi^{-1}(z) = \frac{z-i}{-iz+1} = \frac{iz+1}{z+i}\,.$$

(Jahren emphasizes $G = \xi^{-1}$, with $G^{-1} = \xi$.)

The mutually inverse maps ξ and ξ^{-1} induce a one-to-one correspondence between the points of $\mathbb{D}$ and the points of $\mathbb{H}$.

Since ξ and ξ^{-1} are FLTs, they preserve angles, and map $\bar{\mathbb{C}}$-circles to $\bar{\mathbb{C}}$-circles. Recall that a $\mathbb{D}$-line $L = C \cap \mathbb{D}$ is the part in $\mathbb{D}$ of a $\bar{\mathbb{C}}$-circle that meets $S^1 = \partial\mathbb{D} = \{z \in \mathbb{C} \mid |z| = 1\}$ at right angles. Since ξ maps $S^1 = \partial\mathbb{D}$ to $\bar{\mathbb{R}} = \partial\mathbb{H}$, it takes each $\mathbb{D}$-line to the part in $\mathbb{H}$ of a

$\bar{\mathbb{C}}$-circle that meets $\bar{\mathbb{R}}$ at right angles, i.e., to an $\mathbb{H}$-line. Hence ξ and ξ^{-1} induce a one-to-one correspondence between the $\mathbb{D}$-lines in $\mathbb{D}$ and the $\mathbb{H}$-lines in $\mathbb{H}$.

Clearly these bijections preserve incidence: if a point $z \in \mathbb{D}$ lies on a $\mathbb{D}$-line $L \subset \mathbb{D}$, then the point $\xi(z) \in \mathbb{H}$ lies on the $\mathbb{H}$-line $\xi(L) \subset \mathbb{H}$, and conversely.

These bijections also preserve betweenness. [[Elaborate?]]

To define congruence between line segments or angles in $\mathbb{D}$, we use the group

$$\text{Möb}(\mathbb{D}) = \{m \in \text{Möb}\,|m(\mathbb{D}) = \mathbb{D}\}$$

of Möbius transformations that preserve $\mathbb{D}$. This condition is equivalent to asking that $(\xi \circ m \circ \xi^{-1})(\mathbb{H}) = \mathbb{H}$, i.e., that $\xi m \xi^{-1} \in \text{Möb}(\mathbb{H})$ is a Möbius transform preserving $\mathbb{H}$. Hence Möb($\mathbb{D}$) and Möb($\mathbb{H}$) are conjugate subgroups of Möb = Möb($\mathbb{C}$), with

$$\text{Möb}(\mathbb{H}) = \xi\,\text{Möb}(\mathbb{D})\xi^{-1}$$

and

$$\text{Möb}(\mathbb{D}) = \xi^{-1}\,\text{Möb}(\mathbb{H})\xi\,.$$

In particular, the rule $m \mapsto \xi m \xi^{-1}$ induces a group isomorphism $\text{Möb}(\mathbb{D}) \cong \text{Möb}(\mathbb{H})$. Let $\text{Möb}^+(\mathbb{D})$ and $\text{Möb}^-(\mathbb{D})$ denote the index two subgroup and coset of Möb($\mathbb{D}$) corresponding to $\text{Möb}^+(\mathbb{H})$ and $\text{Möb}^-(\mathbb{H})$ in Möb($\mathbb{H}$), respectively. Here $\text{Möb}^+(\mathbb{D})$ consists of the FLTs preserving $\mathbb{D}$. An example of an element in $\text{Möb}^-(\mathbb{D})$ is complex conjugation: $\sigma(z) = \bar{z}$ for $z \in \mathbb{D}$.

We now define congruence between segments of $\mathbb{D}$-lines, and between angles in $\mathbb{D}$, as we did in $\mathbb{H}$, but using the group Möb($\mathbb{D}$) in place of the group Möb($\mathbb{H}$). The maps ξ and ξ^{-1} will then preserve congruence, i.e., take congruent segments to congruent segments, and take congruent angles to congruent angles.

Since ξ preserves angles, the hyperbolic angle measure in $\mathbb{D}$ will be inherited from the Euclidean angle measure in the containing space $\mathbb{C}$, just as the hyperbolic angle measure in $\mathbb{H}$ was inherited from the Euclidean angle measure in $\mathbb{C}$.

Proposition 12.1. *Each FLT (holomorphic Möbius transformation) preserving $\mathbb{D}$, i.e., each element $m \in \text{Möb}^+(\mathbb{D})$, can be written in the form*

$$m(z) = \frac{\alpha z + \beta}{\bar{\beta} z + \bar{\alpha}}$$

with $|\alpha|^2 - |\beta|^2 = 1$. This presentation is unique, up to replacing (α, β) with $(-\alpha, -\beta)$.

Each anti-holomorphic Möbius transformation preserving $\mathbb{D}$, i.e., each element $m \in \text{Möb}^-(\mathbb{D})$, can be written in the form

$$m(z) = \frac{\alpha \bar{z} + \beta}{\bar{\beta} \bar{z} + \bar{\alpha}}$$

with $|\alpha|^2 - |\beta|^2 = 1$. This presentation is unique, up to replacing (α, β) with $(-\alpha, -\beta)$.

Proof. The group $\text{Möb}^+(\mathbb{D}) = \xi^{-1}\,\text{Möb}^+(\mathbb{H})\xi$ consists of the composites $m = \xi^{-1} \circ n \circ \xi$ with $n(z) = (az + b)/(cz + d)$, where $a, b, c, d \in \mathbb{R}$ and $ad - bc = 1$. Here ξ^{-1}, n and ξ can be represented by the matrices

$$\begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix} \quad , \quad \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}$$

in $GL_2(\mathbb{C})$, respectively, so the composite is represented by the matrix product

$$\begin{aligned} \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix} &= \begin{bmatrix} a - ic & b - id \\ -ia + c & -ib + d \end{bmatrix} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix} \\ &= \begin{bmatrix} a - ic + ib + d & ia + c + b - id \\ -ia + c + b + id & a + ic - ib + d \end{bmatrix} = \begin{bmatrix} 2\alpha & 2\beta \\ 2\bar{\beta} & 2\bar{\alpha} \end{bmatrix} \end{aligned}$$

where

$$2\alpha = (a + d) + i(b - c) \quad \text{and} \quad 2\beta = (b + c) + i(a - d)\,,$$

so that

$$\begin{aligned} a &= \operatorname{Re}\alpha + \operatorname{Im}\beta \\ b &= \operatorname{Im}\alpha + \operatorname{Re}\beta \\ c &= -\operatorname{Im}\alpha + \operatorname{Re}\beta \\ d &= \operatorname{Re}\alpha - \operatorname{Im}\beta\,. \end{aligned}$$

Then $m(z) = (2\alpha z + 2\beta)/(2\bar{\beta}z + 2\bar{\alpha}) = (\alpha z + \beta)/(\bar{\beta}z + \bar{\alpha})$. Comparing determinants, we find $4(ad - bc) = 4(\alpha\bar{\alpha} - \beta\bar{\beta})$, so $|\alpha|^2 - |\beta|^2 = 1$.

Since $\sigma(z) = \bar{z}$ defines an element of $\text{Möb}^-(\mathbb{D})$, each element of $\text{Möb}^-(\mathbb{D})$ can be uniquely written as $m = n \circ \sigma$, where $n \in \text{Möb}^+(\mathbb{D})$. Writing $n(z) = (\alpha z + \beta)/(\bar{\beta}z + \bar{\alpha})$, as above, we get the asserted formula for $m(z)$. □

Proposition 12.2. *Let $m(z) = (\alpha z + \beta)/(\bar{\beta}z + \bar{\alpha})$ in $\text{Möb}^+(\mathbb{D})$ be an FLT preserving $\mathbb{D}$, with $|\alpha|^2 - |\beta|^2 = 1$.*

(a) If $|\operatorname{Re}\alpha| > 1$ then m is of hyperbolic type, and is conjugate in $\text{Möb}^+(\mathbb{D})$ to exactly one FLT preserving $\mathbb{D}$ of the form

$$\ell(z) = \frac{(\cosh t)z + \sinh t}{(\sinh t)z + \cosh t}$$

with $0 < t < \infty$. In this case m has exactly two fixed points in $\bar{\mathbb{C}}$, both of which lie in S^1.

(b) If $|\operatorname{Re}\alpha| = 1$ and $m \neq e$ then m is of parabolic type, and is conjugate in $\text{Möb}^+(\mathbb{D})$ to exactly one of [[ETC]]. In this case m has exactly one fixed point in $\bar{\mathbb{C}}$, which lies in S^1.

(c) If $|\operatorname{Re}\alpha| < 1$ then m is of elliptic type, and is conjugate in $\text{Möb}^+(\mathbb{D})$ to exactly one

$$\ell(z) = e^{i2\theta}z$$

with $0 < \theta < \pi$. In this case m has exactly two fixed points in $\bar{\mathbb{C}}$, one in $\mathbb{D}$ and one outside of S^1.

Proof. The condition $|a + d| > 2$, subject to $ad - bc = 1$, translates to the condition $|\operatorname{Re}\alpha| > 1$, subject to $|\alpha|^2 - |\beta^2| = 1$, and similarly for $|a + d| = 2$ and $|a + d| < 2$.

If $|\operatorname{Re}\alpha| > 1$, we know that the hyperbolic transformation $\xi \circ m \circ \xi^{-1}(z) = (az + b)/(cz + d)$ is conjugate in $\text{Möb}^+(\mathbb{H})$ to a unique FLT $\xi \circ n \circ \xi^{-1}$ (preserving $\mathbb{H}$) of the form $z \mapsto \eta z = (\sqrt{\eta}z + 0)/(0z + 1/\sqrt{\eta})$, with $1 < \eta < \infty$. It follows that m is conjugate in $\xi^{-1}\,\text{Möb}^+(\mathbb{H})\xi = \text{Möb}^+(\mathbb{D})$ to a unique FLT n (preserving $\mathbb{D}$) of the form $z \mapsto \xi^{-1}(\eta \cdot \xi(z))$. Here

$$n(z) = \xi^{-1}(\eta \cdot \xi(z)) = \frac{(\eta + 1)z + i(\eta - 1)}{-i(\eta - 1)z + (\eta + 1)} = \frac{(\cosh t)z + i\sinh t}{(-i\sinh t)z + \cosh t}$$

where $t = \ln\sqrt{\eta} = (\ln\eta)/2$ is uniquely chosen so that $\cosh t = (\sqrt{\eta} + 1/\sqrt{\eta})/2$ and $\sinh t = (\sqrt{\eta} - 1/\sqrt{\eta})/2$. To further simplify this expression we can conjugate again, to $\ell(z) = n(iz)/i$, which equals

$$\ell(z) = \frac{(\cosh t)z + \sinh t}{(\sinh t)z + \cosh t}$$

with $0 < t < \infty$. We know that $\xi \circ m \circ \xi^{-1}$ has exactly two fixed points, both of which lie in $\bar{\mathbb{R}}$, which implies that m also has exactly two fixed points, both of which lie in $\xi^{-1}(\bar{\mathbb{R}}) = S^1$. The axis of m is the $\mathbb{D}$-line with these two endpoints, which m maps to itself as a set.

If $|\operatorname{Re}\alpha| = 1$, the parabolic transformation $\xi \circ m \circ \xi^{-1}(z)$ is conjugate in $\text{Möb}^+(\mathbb{H})$ to exactly one of $z \mapsto z + 1 = (1z + 1)/(0z + 1)$ and $z \mapsto z - 1 = (1z - 1)/(0z + 1)$. It follows that m is conjugate in $\text{Möb}^+(\mathbb{D})$ to exactly one of $z \mapsto \xi^{-1}(\xi(z) + 1)$ and $z \mapsto \xi^{-1}(\xi(z) - 1)$. Here

$$\xi^{-1}(\xi(z) \pm 1) = \frac{(2i \mp 1)z \pm i}{\pm iz + (2i \pm 1)} = \frac{(1 \pm i/2)z \pm 1/2}{\pm z/2 + (1 \mp i/2)}\,.$$

[[Any other standard form?]] We know that $\xi \circ m \circ \xi^{-1}$ has exactly one fixed point, which lies in $\bar{\mathbb{R}}$. This implies that m also has exactly one fixed point, which lies in $\xi^{-1}(\bar{\mathbb{R}}) = S^1$.

If $|\operatorname{Re}\alpha| < 1$, the elliptic transformation $\xi \circ m \circ \xi^{-1}(z)$ is conjugate in $\text{Möb}^+(\mathbb{H})$ to a unique FLT $\xi \circ \ell \circ \xi^{-1}$ of the form $z \mapsto r_\theta(z) = ((\cos\theta)z + \sin\theta)/((-\sin\theta)z + \cos\theta)$, with $0 < \theta < \pi$. Hence m is conjugate in $\text{Möb}^+(\mathbb{D})$ to a unique FLT ℓ of the form $\xi^{-1} \circ r_\theta \circ \xi$. Here

$$\ell(z) = \xi^{-1}\Big(\frac{(\cos\theta)\xi(z) + \sin\theta}{(-\sin\theta)\xi(z) + \cos\theta}\Big) = \frac{e^{i\theta}z + 0}{0z + e^{-i\theta}} = e^{i2\theta}z\,.$$

We know that $\xi \circ m \circ \xi^{-1}$ has exactly two fixed points in $\bar{\mathbb{C}}$, one of which lies in $\mathbb{H}$ and the other its complex conjugate. It follows that m has exactly two fixed points in $\bar{\mathbb{C}}$, one of which lies in $\mathbb{D}$ and the other being its image under inversion in S^1. □

12.2. **Distance in $\mathbb{D}$.** We can translate the distance measure in $\mathbb{H}$, i.e., the metric $d = d_{\mathbb{H}}$, to a distance measure in $\mathbb{D}$, by way of the bijection ξ. For $z, w \in \mathbb{D}$ we define

$$d_{\mathbb{D}}(z, w) = d_{\mathbb{H}}(\xi(z), \xi(z))\,.$$

It is clear that $d_{\mathbb{D}}$ defines a metric on $\mathbb{D}$, so that ξ and ξ^{-1} become isometries, i.e., distance-preserving maps. Furthermore, each Möbius transformation $m \in \text{Möb}(\mathbb{D})$ acts as an isometry on $\mathbb{D}$, so that

$$d_{\mathbb{D}}(z, w) = d_{\mathbb{D}}(m(z), m(w))\,.$$

This follows directly from the corresponding property in $\mathbb{H}$.

Lemma 12.3. *Let $z \ne w \in \mathbb{D}$, and let $p, q \in S^1$ be the endpoints of the $\mathbb{D}$-line through z and w. Then*

$$d_{\mathbb{D}}(z, w) = |\ln(\Big|\frac{w-p}{z-p}\frac{z-q}{w-q}\Big|)|\,.$$

Proof. We may assume that $p * z * w$ and $z * w * q$. Otherwise we interchange p and q, which changes the sign of the logarithm, but does not alter its absolute value. Then $\xi(p)$ and $\xi(q)$ are the endpoints of the $\mathbb{H}$-line through $\xi(z)$ and $\xi(w)$, and $\xi(p) * \xi(z) * \xi(w)$ and $\xi(z) * \xi(w) * \xi(q)$.

Recall that $d_{\mathbb{H}}(\xi(z), \xi(w)) = \ln(|m(\xi(w))|)$, where $m \in \text{Möb}^+(\mathbb{H})$ satisfies $m(\xi(z)) = i$ and $m(\xi(w)) = it$ for some $t > 1$. Then $m(\xi(p)) = 0$ and $m(\xi(q)) = \infty$, so $m(u) = in(u)$, where n is the FLT mapping $\xi(z)$, $\xi(p)$ and $\xi(q)$ to 1, 0 and ∞, respectively. Hence $m(\xi(w)) = in(\xi(w)) = i\ell(w)$, where $\ell = n \circ \xi$ is the FLT mapping z, p and q to 1, 0 and ∞, respectively. Thus

$$m(\xi(w)) = i\frac{w-p}{z-p}\frac{z-q}{w-q}$$

and the formula follows. □

(This argument could also have been formulated in terms of the cross-ratio $[w, z, p, q]$, using its invariance under Möbius transformations.)

We now aim to obtain a formula for $d_{\mathbb{D}}(z, w)$ that does not depend on the endpoints p and q. We begin with the case when $z = 0$. Recall that $\sinh x = (e^x - e^{-x})/2$, $\cosh x = (e^x + e^{-x})/2$ and $\tanh x = \sinh x / \cosh x = (e^x - e^{-x})/(e^x + e^{-x})$.

Lemma 12.4. *Let $w \in \mathbb{D}$, with $r = |w| \in [0, 1)$. Then*

$$d_{\mathbb{D}}(0, w) = \ln\Big(\frac{1+r}{1-r}\Big) \quad \textit{and} \quad r = \tanh\Big(\frac{d_{\mathbb{D}}(0, w)}{2}\Big)\,.$$

Proof. This is the case $z = 0$ and $w = re^{i2\theta}$, for some $\theta \in [0, \pi)$. Multiplication by $e^{i2\theta}$ is an element of $\text{Möb}^+(\mathbb{D})$, corresponding to $\alpha = e^{i\theta}$ and $\beta = 0$, and maps $(0, r)$ to $(0, w)$. By the invariance of $d_{\mathbb{D}}$ under the action of $\text{Möb}(\mathbb{D})$ we deduce that $d_{\mathbb{D}}(0, w) = d_{\mathbb{D}}(0, r)$.

In the special case $z = 0$ and $w = r$, we have $p = -1$ and $q = 1$, so

$$d_{\mathbb{D}}(0, r) = \ln(\Big|\frac{r+1}{0+1} \cdot \frac{0-1}{r-1}\Big|) = \ln\Big(\frac{1+r}{1-r}\Big)$$

as asserted.

The formula $d = \ln((1 + r)/(1 - r))$, with $d = d_{\mathbb{D}}(0, r)$, is equivalent to $(1 + r)/(1 - r) = e^d$, so $(1 + r) = e^d(1 - r)$, $(e^d + 1)r = (e^d - 1)$ and $(e^{d/2} + e^{-d/2})r = (e^{d/2} - e^{-d/2})$, hence $r = \tanh(d/2)$. □

Proposition 12.5. *Let $z, w \in \mathbb{D}$, with*

$$r = \frac{|w-z|}{|1-\bar{z}w|}\,.$$

Then

$$d_{\mathbb{D}}(z,w) = \ln\Big(\frac{1+r}{1-r}\Big) \quad \textit{and} \quad r = \tanh\Big(\frac{d_{\mathbb{D}}(z,w)}{2}\Big)\,.$$

Proof. To find the distance from z to w, we first move z to 0 by an isometry m in $\text{Möb}^+(\mathbb{D})$. If $m(u) = (\alpha u + \beta)/(\bar{\beta} u + \bar{\alpha})$ satisfies $m(z) = 0$, then $\alpha z + \beta = 0$, so

$$m(u) = \frac{\alpha u - \alpha z}{-\bar{\alpha}\bar{z}u + \bar{\alpha}} = \frac{\alpha}{\bar{\alpha}} \cdot \frac{u-z}{1-\bar{z}u}\,.$$

Hence $m(w) = (a/\bar{\alpha})(w-z)/(1-\bar{z}w)$, with $|m(w)| = |w-z|/|1-\bar{z}w| = r$, and $d_{\mathbb{D}}(z,w) = d_{\mathbb{D}}(m(z), m(w)) = d_{\mathbb{D}}(0, m(w)) = \ln((1+r)/(1-r))$, as claimed. □

Proposition 12.6. *Let $z, w \in \mathbb{D}$. Then*

$$\sinh^2\Big(\frac{d_{\mathbb{D}}(z,w)}{2}\Big) = \frac{|w-z|^2}{(1-|z|^2)(1-|w|^2)}$$

and

$$\cosh(d_{\mathbb{D}}(z,w)) = 1 + \frac{2|w-z|^2}{(1-|z|^2)(1-|w|^2)}\,.$$

Proof. Let $d = d_{\mathbb{D}}(z,w)$, so that $r = \tanh(d/2)$. Then

$$\frac{r^2}{1-r^2} = \frac{\tanh^2(d/2)}{1-\tanh^2(d/2)} = \frac{\sinh^2(d/2)}{\cosh^2(d/2) - \sinh^2(d/2)} = \sinh^2(d/2)$$

where we used the identity $\cosh^2 x - \sinh^2 x = 1$. Furthermore,

$$\frac{r^2}{1-r^2} = \frac{\dfrac{|w-z|^2}{|1-\bar{z}w|^2}}{1 - \dfrac{|w-z|^2}{|1-\bar{z}w|^2}} = \frac{|w-z|^2}{|1-\bar{z}w|^2 - |w-z|^2}\,.$$

Here

$$\begin{aligned} |1-\bar{z}w|^2 - |w-z|^2 &= (1-\bar{z}w)(1-z\bar{w}) - (w-z)(\bar{w}-\bar{z}) \\ &= 1 - z\bar{w} - \bar{z}w + |z|^2|w|^2 - |w|^2 + w\bar{z} + z\bar{w} - |z|^2 \\ &= (1-|z|^2)(1-|w|^2) \end{aligned}$$

so

$$\sinh^2(d/2) = \frac{r^2}{1-r^2} = \frac{|w-z|^2}{(1-|z|^2)(1-|w|^2)}\,.$$

Finally we use $\cosh(2x) = \cosh^2 x + \sinh^2 x = 1 + 2\sinh^2 x$ with $x = d/2$ to obtain the convenient formula

$$\cosh(d_{\mathbb{D}}(z,w)) = 1 + \frac{2|w-z|^2}{(1-|z|^2)(1-|w|^2)}\,.$$

□

This achieves our aim of expressing the hyperbolic distance between two points in $\mathbb{D}$, in a formula that does not involve auxiliary quantities (like the endpoints of hyperbolic lines). Note that $\cosh x$ is strictly increasing for $x \geq 0$, so the last formula determines $d_{\mathbb{D}}(z,w) \geq 0$ uniquely.

We can also translate these formulas back to $\mathbb{H}$, using $\xi^{-1} \colon \mathbb{H} \to \mathbb{D}$.

Proposition 12.7. *Let $z, w \in \mathbb{H}$. Then*

$$\cosh(d_{\mathbb{H}}(z,w)) = 1 + \frac{|w-z|^2}{2(\operatorname{Im} z)(\operatorname{Im} w)}\,.$$

Proof. We have $d_{\mathbb{H}}(z,w) = d_{\mathbb{D}}(\xi^{-1}(z), \xi^{-1}(w))$, where

$$\xi^{-1}(w) - \xi^{-1}(z) = \frac{iw+1}{w+i} - \frac{iz+1}{z+i} = \frac{2(z-w)}{(z+i)(w+i)}$$

and

$$1 - |\xi^{-1}(z)|^2 = 1 - \left|\frac{iz+1}{z+i}\right|^2 = \frac{4\operatorname{Im} z}{|z+i|^2}$$

(and likewise for w in place of z). Hence

$$\frac{2|\xi^{-1}(w) - \xi^{-1}(z)|^2}{(1-|\xi^{-1}(z)|^2)(1-|\xi^{-1}(w)|^2)} = \frac{|w-z|^2}{2(\operatorname{Im} z)(\operatorname{Im} w)},$$

which gives the asserted formula. □

13. October 2nd lecture

13.1. **Arc length in the hyperbolic plane.** The length measure of hyperbolic line segments $[z,w]$ in $\mathbb{H}$ can be generalized to a length measure for reasonable parametrized curves $\omega\colon [a,b] \to \mathbb{H}$. Here "reasonable" can be interpreted as "rectifiable", but for simplicity we can limit ourselves to the more restricted class of continuously differentiable (or C^1) curves in $\mathbb{H}$. Here $\omega\colon [a,b] \to \mathbb{H}$ is continuously differentiable if the components $x = \operatorname{Re}\omega\colon [a,b] \to \mathbb{R}$ and $y = \operatorname{Im}\omega\colon [a,b] \to \mathbb{R}$ of the composite map $[a,b] \to \mathbb{H} \subset \mathbb{C}$ are continuously differentiable in the usual sense.

We will assign a hyperbolic length measure $\|\omega'(t)\|_{\mathbb{H}}$ to each tangent vector $\omega'(t)$ of a C^1 curve in $\mathbb{H}$, depending on the starting point $\omega(t)$ of that tangent vector, in such a way that the hyperbolic length of the curve ω is the integral of these hyperbolic lengths of tangent vectors:

$$\operatorname{length}_{\mathbb{H}}(\omega) = \int_a^b \|\omega'(t)\|_{\mathbb{H}}\, dt\,.$$

Since the Möbius transformations $m\colon \mathbb{H} \to \mathbb{H}$ preserve lengths of hyperbolic line segments, we will also want them to preserve lengths of C^1 curves. In order to have

$$\operatorname{length}_{\mathbb{H}}(\omega) = \operatorname{length}_{\mathbb{H}}(m \circ \omega)$$

we will need to have

$$\|\omega'(t)\|_{\mathbb{H}} = \|(m\circ\omega)'(t)\|_{\mathbb{H}}\,,$$

where the length of $v = \omega'(t)$ at the left hand side is measured at $z = \omega(t)$, while the length of $(m\circ\omega)'(t)$ at the right hand side is measured at $m(z) = (m\circ\omega)(t)$. By the chain rule, $(m\circ\omega)'(t) = m'(z)\cdot\omega'(t)$.

For instance, if $m(z) = z + b$, we have $m'(z) = 1$ for all z, so we require that the $\mathbb{H}$-length of $v = \omega'(t)$ at $z = \omega(t)$ is equal to the $\mathbb{H}$-length of $1 \cdot v = v$ at $m(z) = z + b$. In other words, the hyperbolic length measure at z does not depend on the real part of z.

On the other hand, if $m(z) = \eta z$, with $\eta > 0$, we have $m'(z) = \eta$ for all z, so we require that the $\mathbb{H}$-length of $v = \omega'(t)$ at $z = \omega(t)$ is equal to the $\mathbb{H}$-length of $\eta \cdot v = \eta v$ at $m(z) = \eta z$. Hence the $\mathbb{H}$-length of a vector w at ηz is equal to $1/\eta$ times the $\mathbb{H}$-length of the same vector at z. This implies that the hyperbolic length measure at z is inversely proportional to the imaginary part of z.

This suggests the formula

$$\|\omega'(t)\|_{\mathbb{H}} = \frac{|\omega'(t)|}{\operatorname{Im}\omega(t)}$$

for the hyperbolic length measure of the tangent vector $\omega'(t)$ of the curve ω at the point $\omega(t)$. Here $|\omega'(t)|$ refers to the usual Euclidean length of that tangent vector. If we write $\omega(t) = (x(t), y(t))$, then

$$\|\omega'(t)\|_{\mathbb{H}} = \frac{\sqrt{x'(t)^2 + y'(t)^2}}{y(t)}\,.$$

Definition 13.1. The hyperbolic *length* of a C^1 curve $\omega\colon [a,b] \to \mathbb{H}$, with components $\omega(t) = (x(t), y(t))$, is defined to be

$$\mathrm{length}_{\mathbb{H}}(\omega) = \int_a^b \frac{|\omega'(t)|}{\operatorname{Im}\omega(t)}\,dt = \int_a^b \frac{\sqrt{x'(t)^2 + y'(t)^2}}{y(t)}\,dt\,.$$

Lemma 13.2. *The hyperbolic length of a C^1 curve in $\mathbb{H}$ is invariant under Möbius transformations, i.e.,* $\mathrm{length}_{\mathbb{H}}(\omega) = \mathrm{length}_{\mathbb{H}}(m \circ \omega)$ *for* $m \in \mathrm{Möb}(\mathbb{H})$.

Proof. If m is holomorphic, so $m(z) = (az+b)/(cz+d)$ for some $a, b, c, d \in \mathbb{R}$ with $ad - bc = 1$, then

$$(m \circ \omega)'(t) = m'(z) \cdot \omega'(t) = \frac{\omega'(t)}{(cz+d)^2}$$

and

$$\operatorname{Im}(m \circ \omega)(t) = \frac{\operatorname{Im} z}{|cz+d|^2}$$

for $z = \omega(t)$, by earlier calculations, so

$$\frac{|\omega'(t)|}{\operatorname{Im}\omega(t)} = \frac{|(m \circ \omega)'(t)|}{\operatorname{Im}(m \circ \omega)(t)}\,.$$

This implies

$$\mathrm{length}_{\mathbb{H}}(\omega) = \int_a^b \frac{|\omega'(t)|}{\operatorname{Im}\omega(t)}\,dt = \int_a^b \frac{|(m \circ \omega)'(t)|}{\operatorname{Im}(m \circ \omega)(t)}\,dt = \mathrm{length}_{\mathbb{H}}(m \circ \omega)\,.$$

(The limits of integration, a and b, are not related to the FLT coefficients of the same names.)

If $m = \rho$ is given by reflection in the imaginary axis, then $(\rho \circ \omega)'(t)$ is given by the same reflection from $\omega'(t)$, hence $|(\rho \circ \omega)'(t)| = |\omega'(t)|$. This reflection preserves imaginary parts, so $\operatorname{Im}(\rho \circ \omega)(t) = \operatorname{Im}\omega(t)$. Hence the last steps of the calculation for holomorphic m also applies for $m = \rho$, so that $\mathrm{length}_{\mathbb{H}}(\omega) = \mathrm{length}_{\mathbb{H}}(\rho \circ \omega)$.

The general case of anti-holomorphic m now follows from these two cases. □

In the following statement, $[z, w] \subset \overleftrightarrow{zw} \subset \mathbb{H}$ refers to a hyperbolic line segment, whereas $[a, b] \subset \mathbb{R}$ refers to a Euclidean line segment, i.e., a closed interval.

Lemma 13.3. *The hyperbolic length of a hyperbolic line segment $[z, w]$ in $\mathbb{H}$, viewed as a C^1 curve by any continuously differentiable bijection $\omega\colon [a,b] \to [z,w]$, is equal to the hyperbolic distance $d(z,w) = d_{\mathbb{H}}(z,w)$ from z to w.*

Proof. By the invariance of hyperbolic length of curves under Möbius transformations, we may assume that $[z, w]$ is the segment $[i, ib]$ on the imaginary axis, so that $d(z,w) = d(i, ib) = \ln(b)$. We can parametrize that segment by $\omega(t) = it$, for $t \in [1, b]$, with $|\omega'(t)| = |i| = 1$ and $\operatorname{Im}\omega(t) = t$, so that

$$\mathrm{length}_{\mathbb{H}}(\omega) = \int_1^b \frac{1}{t}\,dt = \Big[\ln(t)\Big]_1^b = \ln(b) - \ln(1) = \ln(b)\,.$$

Hence the two length measures agree. It is a consequence of the change-of-variable formulas for integrals that any other (C^1 bijective) choice of parametrization ω gives the same result. □

It follows that the hyperbolic length measure of piecewise $\mathbb{H}$-linear curves in $\mathbb{H}$ equals the sum of the distances along the linear segments. [[Relate length of a C^1 curve to the supremum of the lengths of the piecewise $\mathbb{H}$-linear approximations to the curve.]]

We can translate the length measure for C^1 curves in $\mathbb{H}$ to an equivalent length measure for C^1 curves in $\mathbb{D}$, using the chosen identification $\xi\colon \mathbb{D} \to \mathbb{H}$, with $\xi(z) = (z+i)/(iz+1)$. We need to assign a hyperbolic length measure $\|\omega'(t)\|_{\mathbb{D}}$ to each tangent vector of a C^1 curve $\omega\colon [a,b] \to \mathbb{D}$ in $\mathbb{D}$, again depending on both $z = \omega(t)$ and $v = \omega'(t)$, in such a way that

$$\mathrm{length}_{\mathbb{D}}(\omega) = \int_a^b \|\omega'(t)\|_{\mathbb{D}}\,dt$$

satisfies $\operatorname{length}_{\mathbb{D}}(\omega) = \operatorname{length}_{\mathbb{H}}(\xi \circ \omega)$. To achieve this we must have

$$\|\omega'(t)\|_{\mathbb{D}} = \|(\xi \circ \omega)'(t)\|_{\mathbb{H}} = \frac{|(\xi \circ \omega)'(t)|}{\operatorname{Im}(\xi \circ \omega)(t)} = \frac{|\xi'(z)||\omega'(t)|}{\operatorname{Im} \xi(z)}$$

where $z = \omega(t)$. Here

$$|\xi'(z)| = \frac{2}{|iz+1|^2}$$

and

$$\xi(z) = \frac{2 \operatorname{Re} z + i(1 - |z|^2)}{|iz+1|^2},$$

so the factors $|iz+1|^2$ cancel, and

$$\frac{|\xi'(z)||\omega'(t)|}{\operatorname{Im} \xi(z)} = \frac{2|\omega'(t)|}{1 - |z|^2} = \frac{2|\omega'(t)|}{1 - |\omega(t)|^2}.$$

This proves the following statement.

Proposition 13.4. *The hyperbolic length of a C^1 curve $\omega\colon [a,b] \to \mathbb{D}$, with components $\omega(t) = (x(t), y(t))$, is given by*

$$\operatorname{length}_{\mathbb{D}}(\omega) = \int_a^b \frac{2|\omega'(t)|}{1 - |\omega(t)|^2}\, dt = \int_a^b \frac{2\sqrt{x'(t)^2 + y'(t)^2}}{1 - x(t)^2 - y(t)^2}\, dt.$$

It is equal to the hyperbolic length in $\mathbb{H}$ of the composite $\xi \circ \omega\colon [a,b] \to \mathbb{H}$, and agrees with the distance measure $d_{\mathbb{D}}(z,w)$ when ω is a bijective parametrization of a segment $[z,w]$ in $\mathbb{D}$. □

Example 13.5. Let $0 < r < 1$, and consider the closed curve $\omega\colon [0, 2\pi] \to \mathbb{D}$ given by $\omega(t) = re^{it} = r(\cos t + i \sin t)$, with components $x(t) = r\cos t$ and $y(t) = r \sin t$. Let $\rho = \ln((1+r)/(1-r))$ be the hyperbolic distance from 0 to $\omega(0) = r$. We have $d_{\mathbb{D}}(0, \omega(t)) = \rho$ for all t, since rotations around 0 are FLTs and hence hyperbolic isometries, so the curve parametrized by ω is both the Euclidean circle with center 0 and radius r, and the hyperbolic circle with center 0 and radius ρ.

Note that $\omega'(t) = ire^{it} = ir(\cos t + i \sin t)$, so $|\omega(t)| = r$ and $|\omega'(t)| = r$ for all t. The hyperbolic length of ω is

$$\operatorname{length}_{\mathbb{D}}(\omega) = \int_0^{2\pi} \frac{2r}{1 - r^2}\, dt = \frac{4\pi r}{1 - r^2}.$$

To express this in terms of the hyperbolic radius ρ, recall that $r = \tanh(\rho/2)$, so

$$\frac{4\pi r}{1 - r^2} = \frac{4\pi \tanh(\rho/2)}{1 - \tanh^2(\rho/2)} = 4\pi \sinh(\rho/2)\cosh(\rho/2) = 2\pi \sinh \rho.$$

This uses $1 - \tanh^2 x = 1/\cosh^2 x$ and $\sinh(2x) = 2 \sinh x \cosh x$. The Taylor expansion

$$\operatorname{length}_{\mathbb{D}}(\omega) = 2\pi \sinh \rho = 2\pi(\rho + \frac{\rho^3}{6} + \frac{\rho^5}{120} + \dots)$$

shows that the circumference of a hyperbolic circle with radius ρ is longer than the circumference of a Euclidean circle with an equally long radius (namely $2\pi\rho$), and that the hyperbolic circumference grows faster than the Euclidean circumference. This is a manifestation of the negative curvature of the hyperbolic plane.

Remark 13.6. On the unit sphere, S^2, with the distances measured along great circles on that surface, the circle ω with center the north pole $N = (0,0,1)$ and spherical radius ρ is equal to the Euclidean circle of radius $r = \sin \rho$ and circumference $2\pi r = 2\pi \sin \rho$. (This is the line of latitude $\pi/2 - \rho$.) Its Taylor expansion

$$\operatorname{length}_{S^2}(\omega) = 2\pi \sin \rho = 2\pi(\rho - \frac{\rho^3}{6} + \frac{\rho^5}{120} - \dots)$$

shows that the circumference of a spherical circle with radius ρ is shorter than the circumference of a Euclidean circle with an equally long radius, and that the spherical circumference grows slower of a function of ρ than the Euclidean circumference. This is a manifestation of the positive curvature of the sphere.

13.2. **Area measure in the hyperbolic plane.** We shall assign an area measure to hyperbolic triangles $\triangle ABC \subset \mathbb{H}$, and more generally to nice regions $\Omega \subset \mathbb{H}$, i.e., domains that are bounded by a finite number of C^1 curves (in $\mathbb{H} \subset \mathbb{C} \cong \mathbb{R}^2$). At a point $z = x + iy \in \mathbb{H}$ we have assigned the hyperbolic length $1/\operatorname{Im} z = 1/y$ to each of the unit vectors $e_1 = (1, 0)$ and $e_2 = (0, 1)$, so we shall assign the hyperbolic area $(1/\operatorname{Im} z)^2 = 1/y^2$ to the unit square tangent to $\mathbb{H}$ at z. The hyperbolic area of a nice region Ω is then the integral of the hyperbolic area of these tangent squares:

$$\operatorname{area}_{\mathbb{H}}(\Omega) = \iint_\Omega \frac{1}{y^2}\, dx\, dy = \iint_\Omega \frac{dx\, dy}{y^2}\,.$$

Lemma 13.7. *The hyperbolic area measure satisfies:*

(a) (additivity)

$$\operatorname{area}_{\mathbb{H}}(\Omega_1 \cup \Omega_2) = \operatorname{area}_{\mathbb{H}}(\Omega_1) + \operatorname{area}_{\mathbb{H}}(\Omega_2) - \operatorname{area}_{\mathbb{H}}(\Omega_1 \cap \Omega_2)$$

for nice regions Ω_1 and $\Omega_2 \subset \mathbb{H}$, and

(b) (Möbius invariance)

$$\operatorname{area}_{\mathbb{H}}(\Omega) = \operatorname{area}_{\mathbb{H}}(m(\Omega))$$

for any nice region $\Omega \subset \mathbb{H}$ and Möbius transformation m preserving $\mathbb{H}$.

Proof. (a) This is clear from the additivity properties of double integrals.

(b) Suppose first that $m(z) = (az + b)/(cz + d)$, with $a, b, c, d \in \mathbb{R}$ and $ad - bc = 1$. The real derivative of m at z is the linear map from $\mathbb{R}^2 \cong \mathbb{C}$ to itself given by complex multiplication by $m'(z) = 1/(cz + d)^2$, which has real determinant $J(m)(z) = |m'(z)|^2 = 1/|cz + d|^4$. This is the determinant of the Jacobian of m at z. Writing $m(z) = w = u + iv$, so that $v = \operatorname{Im} m(z) = \operatorname{Im} z/|cz + d|^2$, we have

$$\begin{aligned}\operatorname{area}_{\mathbb{H}}(m(\Omega)) &= \iint_{m(\Omega)} \frac{du\, dv}{v^2} = \iint_\Omega |J(m)(z)| \cdot \frac{dx\, dy}{(\operatorname{Im} m(z))^2} \\ &= \iint_\Omega \frac{|cz + d|^4}{|cz + d|^4} \cdot \frac{dx\, dy}{(\operatorname{Im} z)^2} = \iint_\Omega \frac{dx\, dy}{y^2} = \operatorname{area}_{\mathbb{H}}(\Omega)\,.\end{aligned}$$

If m is not holomorphic, it can be written as the composite of the reflection $\rho(z) = -\bar{z}$ about the imaginary axis and an FLT as above, so it only remains to check that the hyperbolic area is invariant under ρ. This is clear, since the derivative of ρ is the same reflection, with determinant -1, and $\operatorname{Im} \rho(z) = \operatorname{Im} z$ since the reflection preserves the y-coordinate:

$$\operatorname{area}_{\mathbb{H}}(\rho(\Omega)) = \iint_{\rho(\Omega)} \frac{du\, dv}{v^2} = \iint_\Omega |-1| \cdot \frac{dx\, dy}{y^2} = \operatorname{area}_{\mathbb{H}}(\Omega)\,.$$

□

Definition 13.8. Let A, B and C be three distinct points in $\mathbb{H}$. The hyperbolic *triangle* $\triangle ABC$ is the region bounded by the segments $[A, B]$, $[B, C]$ and $[A, C]$. It lies on the same side of $\overleftrightarrow{AB}$ as C, on the same side of $\overleftrightarrow{BC}$ as A, and on the same side of $\overleftrightarrow{AC}$ as B.

If A and B are distinct points in $\mathbb{H}$, and $C \in \bar{\mathbb{R}}$, the (singly) *ideal triangle* $\triangle ABC$ is the region in $\mathbb{H}$ bounded by the segment $[A, B]$ and the rays $\overrightarrow{BC}$ and $\overrightarrow{AC}$.

If $A \in \mathbb{H}$ and B and C are distinct points in $\bar{\mathbb{R}}$, the (doubly) ideal triangle $\triangle ABC$ is the region in $\mathbb{H}$ bounded by the ray $\overrightarrow{AB}$, the line (B, C) and the ray $\overrightarrow{AC}$.

If A, B and C are distinct points in $\bar{\mathbb{R}}$, the (triply) ideal triangle $\triangle ABC$ is the region in $\mathbb{H}$ bounded by the lines (A, B), (B, C) and (A, C).

Theorem 13.9 (Lambert)**.** *Let $\triangle ABC$ be a hyperbolic triangle in $\mathbb{H}$, possibly with some ideal vertices. Let $\alpha = \angle BAC$, $\beta = \angle ABC$ and $\gamma = \angle ACB$ be the angles in the triangle. (The angle at an ideal vertex is* 0*.) Then the area of $\triangle ABC$ satisfies*

$$\operatorname{area}_{\mathbb{H}}(\triangle ABC) = \pi - (\alpha + \beta + \gamma)\,.$$

Proof. We begin with the singly ideal case, in which $C \in \bar{\mathbb{R}}$ and $\gamma = 0$. We can first normalize by a Möbius transform preserving $\mathbb{H}$ to get in the situation where $C = \infty$ and $\overleftrightarrow{AB} = (-1, 1)$ is the semicircle with center 0 and radius 1, so that $A = e^{i(\pi-\alpha)}$ and $B = e^{i\beta}$, where $\alpha = \angle BAC$ and $\beta = \angle ABC$ are the angles in the ideal triangle. (The Möbius transform does not change the area or the angles in the triangle.) Then

$$\triangle ABC = \{(x, y) \mid \cos(\pi - \alpha) \le x \le \cos\beta, \sqrt{1 - x^2} \le y\}$$

and

$$\begin{aligned}
\operatorname{area}_{\mathbb{H}}(\triangle ABC) &= \iint_{\triangle ABC} \frac{dx\,dy}{y^2} = \int_{\cos(\pi-\alpha)}^{\cos\beta} \Big(\int_{\sqrt{1-x^2}}^{\infty} \frac{1}{y^2}\,dy\Big)\,dx \\
&= \int_{\cos(\pi-\alpha)}^{\cos\beta} \Big[-\frac{1}{y}\Big]_{\sqrt{1-x^2}}^{\infty} dx = \int_{\cos(\pi-\alpha)}^{\cos\beta} \frac{dx}{\sqrt{1-x^2}} \\
&= \int_{\pi-\alpha}^{\beta} \frac{-\sin t}{\sin t}\,dt = \Big[-t\Big]_{\pi-\alpha}^{\beta} = -\beta + (\pi - \alpha) = \pi - (\alpha + \beta)\,,
\end{aligned}$$

where we substituted $x = \cos t$, with $dx = -\sin t\,dt$ and $\sqrt{1 - x^2} = \sin t$. This completes the proof in the singly ideal case.

The doubly and triply ideal cases are obtained as limits of this case. In particular, any two triply ideal triangles are congruent, and each of them has area equal to π.

In the finite (non-ideal) case, with A, B and C in $\mathbb{H}$, let $D \in \bar{\mathbb{R}}$ be the endpoint of the ray $\overrightarrow{BC}$, and consider the ideal triangle $\triangle ABD$ as the union of $\triangle ABC$ and the ideal triangle $\triangle ACD$. Let $\alpha = \angle BAC$, $\alpha' = \angle CAD$, $\beta = \angle ABC$, $\gamma = \angle ACB$, and $\gamma' = \angle ACD$. Then $\gamma + \gamma' = \pi$, $\operatorname{area}(\triangle ABD) = \pi - (\alpha + \alpha' + \beta)$ and $\operatorname{area}(\triangle ACD) = \pi - (\alpha' + \gamma') = -\alpha' + \gamma$, so $\operatorname{area}(\triangle ABC) = \operatorname{area}(\triangle ABD) - \operatorname{area}(\triangle ACD) = \pi - (\alpha + \beta + \gamma)$, as asserted. □

13.3. **Area measure in $\mathbb{D}$.** Using $\xi\colon \mathbb{D} \to \mathbb{H}$, we translate the hyperbolic area measure in $\mathbb{H}$ to a hyperbolic area measure in $\mathbb{D}$. For a nice region $\Omega \subset \mathbb{D}$, i.e., one bounded by a finite number of C^1 curves, we define

$$\operatorname{area}_{\mathbb{D}}(\Omega) = \operatorname{area}_{\mathbb{H}}(\xi(\Omega))\,.$$

Proposition 13.10. *The hyperbolic area of a nice region $\Omega \subset \mathbb{D}$ is given by*

$$\operatorname{area}_{\mathbb{D}}(\Omega) = \iint_{\Omega} \frac{4\,dx\,dy}{(1 - |z|^2)^2} = \iint_{\Omega} \frac{4\,dx\,dy}{(1 - x^2 - y^2)^2}\,.$$

Proof. We use the change-of-variables formula for $\xi\colon \mathbb{D} \to \mathbb{H}$ mapping $z = x + iy \in \mathbb{D}$ to $\xi(z) = w = u+iv \in \mathbb{H}$, in the case $\xi(z) = (z+i)/(iz+1)$. Here $\xi'(z) = 2/(iz+1)^2$, so the Jacobian determinant is $J(\xi)(z) = |\xi'(z)|^2 = 4/|iz + 1|^4$. Furthermore $v = \operatorname{Im}\xi(z) = (1 - |z|^2)/|iz + 1|^2$, so

$$\begin{aligned}
\operatorname{area}_{\mathbb{D}}(\Omega) = \operatorname{area}_{\mathbb{H}}(\xi(\Omega)) &= \iint_{\xi(\Omega)} \frac{1}{v^2}\,du\,dv = \iint_{\Omega} \frac{1}{(\operatorname{Im}\xi(z))^2} |J(\xi)(z)|\,dx\,dy \\
&= \iint_{\Omega} \frac{|iz + 1|^4}{(1 - |z|^2)^2} \frac{4\,dx\,dy}{|iz + 1|^4} = \iint_{\Omega} \frac{4\,dx\,dy}{(1 - |z|^2)^2}\,.
\end{aligned}$$

□

In polar coordinates, with $z = Re^{i\theta}$ and $(x, y) = (R\cos\theta, R\sin\theta)$, we have $dx\,dy = R\,dR\,d\theta$, so

$$\operatorname{area}_{\mathbb{D}}(\Omega) = \iint \frac{4R\,dR\,d\theta}{(1 - R^2)^2}\,,$$

where the area of integration on the right hand side is the set of (R, θ) with $R \geq 0$ and $0 \leq \theta \leq 2\pi$ for which $Re^{i\theta} \in \Omega$.

Example 13.11. Let $0 < r < 1$ and consider the hyperbolic disc $\Omega = \{z \in \mathbb{D} || z| < r\}$ of hyperbolic radius $\rho = \ln((1 + r)/(1 - r))$, so that $r = \tanh(\rho/2)$. Its hyperbolic area is

$$\begin{aligned}\text{area}_{\mathbb{D}}(\Omega) &= \iint_{x^2+y^2<r^2} \frac{4\,dx\,dy}{(1-x^2-y^2)^2} = \int_0^{2\pi} \Big(\int_0^r \frac{4R\,dR}{(1-R^2)^2} \Big)\, d\theta \\ &= 2\pi \Big[\frac{2}{1-R^2} \Big]_0^r = 4\pi \frac{r^2}{1-r^2} = 4\pi \sinh^2(\rho/2)\,.\end{aligned}$$

This uses $\tanh^2 x/(1 - \tanh^2 x) = \sinh^2 x$, like earlier. The Taylor expansion of this hyperbolic area

$$\text{area}_{\mathbb{D}}(\Omega) = 4\pi \sinh^2(\rho/2) = \pi(\rho^2 + \frac{\rho^4}{12} + \frac{\rho^6}{360} + \dots)$$

shows that the area of a hyperbolic disc with radius ρ is greater than the area of a Euclidean disc with an equally long radius (namely $\pi\rho^2$), and that the hyperbolic area grows faster than the Euclidean area. Again, this is a consequence of the negative curvature of the hyperbolic plane.

Remark 13.12. On the unit sphere, S^2, with distances measured along great circles on that surface, the disc Ω of radius ρ with center at the north pole is bounded by the circle ω at latitude $\pi/2 - \rho$, as discussed in an earlier remark. The surface area of Ω is

$$\text{area}_{S^2}(\Omega) = \int_0^{\rho} 2\pi \sin R\,dR = 2\pi \Big[-\cos R \Big]_0^{\rho} = 2\pi(1 - \cos\rho) = 4\pi \sin^2(\rho/2)\,.$$

The Taylor expansion of this spherical area

$$\text{area}_{S^2}(\Omega) = 4\pi \sin^2(\rho/2) = \pi(\rho^2 - \frac{\rho^4}{12} + \frac{\rho^6}{360} + \dots)$$

shows that the area of a spherical disc with radius ρ is smaller than the area of a Euclidean disc with the same radius (i.e., $\pi\rho^2$), and that the spherical area grows more slowly than the Euclidean area. Once more, this expresses the positive curvature of the sphere and its elliptic geometry.

14. October 7th lecture

14.1. **Trigonometry in the hyperbolic plane.** Consider a hyperbolic triangle $\triangle ABC$. Let $\alpha = \angle BAC$, $\beta = \angle CBA$ and $\gamma = \angle ACB$ be the hyperbolic angles at the three vertices, and let $a = d(B, C)$, $b = d(C, A)$ and $c = d(A, B)$ by the hyperbolic lengths of the respective opposite sides.

By the side-angle-side (SAS) criterion for congruence, i.e., Hilbert's axiom (C6), knowledge of the angle α and the lengths b and c, which amounts to knowledge of the congruence classes of $\angle BAC$, $[C, A]$ and $[A, B]$, determines the congruence classes of $[B, C]$, $\angle CBA$ and $\angle ACB$, i.e., the length a and the angles β and γ.

We now determine the trigonometric identities that allow us to compute a, β and γ in terms of α, b and c. We begin with a formula that determines a.

Theorem 14.1 (The first hyperbolic law of cosines)**.**

$$\cosh a = \cosh b \cosh c - \sinh b \sinh c \cos\alpha\,.$$

Proof. We work in the Poincaré disc model $\mathbb{D}$. Applying a FLT preserving $\mathbb{D}$ that maps the ray $\overrightarrow{AB}$ to the ray $[0, 1)$, and continuing with complex conjugation if needed, we may assume that $A = 0$, $B = r$ and $C = se^{i\alpha}$ for some $r, s \in (0, 1)$. Then $c = d(0, r) = \ln((1 + r)/(1 - r))$ and $b = d(0, se^{i\alpha}) = d(0, s) = \ln((1 + s)/(1 - s))$. Our task is to determine $a = d(r, se^{i\alpha})$. We use the formula

$$\cosh a = 1 + \frac{2|se^{i\alpha} - r|^2}{(1-r^2)(1-s^2)}\,.$$

Here $|se^{i\alpha} - r|^2 = (s\cos\alpha - r)^2 + (s\sin\alpha)^2 = r^2 + s^2 - 2rs\cos\alpha$, and a little calculation gives the expression

$$\cosh a = \frac{1+r^2}{1-r^2}\,\frac{1+s^2}{1-s^2} - \frac{2r}{1-r^2}\,\frac{2s}{1-s^2}\,\cos\alpha\,.$$

Furthermore, $r = \tanh(c/2)$, so

$$\frac{1+r^2}{1-r^2} = \frac{1+\tanh^2(c/2)}{1-\tanh^2(c/2)} = \cosh c$$

and

$$\frac{2r}{1-r^2} = \frac{2\tanh(c/2)}{1-\tanh^2(c/2)} = \sinh c\,,$$

and likewise for s and b. Hence

$$\cosh a = \cosh c\,\cosh b - \sinh c\,\sinh b\,\cos\alpha\,,$$

as asserted. □

Remark 14.2. In terms of Taylor expansions, the first cosine law becomes

$$(1 + \frac{a^2}{2} + \dots) = (1 + \frac{b^2}{2} + \dots)(1 + \frac{c^2}{2} + \dots) - (b + \dots)(c + \dots)\cos\alpha$$

which we can rewrite as

$$1 + \frac{a^2}{2} \equiv 1 + \frac{b^2}{2} + \frac{c^2}{2} - bc\cos\alpha$$

modulo terms of order 4 or higher (in a, b and c). Hence, for small a, b and c the hyperbolic cosine law tends to the Euclidean cosine law

$$a^2 = b^2 + c^2 - 2bc\cos\alpha\,.$$

Corollary 14.3 (The hyperbolic Pythagorean theorem)**.** *If $\alpha = \pi/2$, then*

$$\cosh a = \cosh b\,\cosh c\,.$$

Remark 14.4. For small triangles, this converges to the Pythagorean theorem $a^2 = b^2 + c^2$.

Conversely, knowing the sides a, b and c we can use the first cosine law to determine the angle α. Permuting the vertices of the triangle, we can in principle also use this to determine the angles β and γ. Here is the clean way to state the result.

Theorem 14.5 (The hyperbolic law of sines)**.**

$$\frac{\sin\alpha}{\sinh a} = \frac{\sin\beta}{\sinh b} = \frac{\sin\gamma}{\sinh c}\,.$$

Proof. Squaring

$$\sinh b\,\sinh c\,\cos\alpha = \cosh b\,\cosh c - \cosh a$$

we get

$$(\sinh^2 b\,\sinh^2 c)(1 - \sin^2\alpha) = (\cosh b\,\cosh c - \cosh a)^2$$

and

$$\begin{aligned}\sin^2\alpha\,\sinh^2 b\,\sinh^2 c &= (\cosh^2 b - 1)(\cosh^2 c - 1) - (\cosh b\,\cosh c - \cosh a)^2\\ &= 1 - \cosh^2 a - \cosh^2 b - \cosh^2 c + 2\cosh a\cosh b\cosh c\,.\end{aligned}$$

This expression is symmetric in a, b and c, so we get

$$\sin^2\alpha\,\sinh^2 b\,\sinh^2 c = \sin^2\beta\,\sinh^2 c\,\sinh^2 a = \sin^2\gamma\,\sinh^2 a\,\sinh^2 b\,.$$

Hence

$$\frac{\sin^2\alpha}{\sinh^2 a} = \frac{\sin^2\beta}{\sinh^2 b} = \frac{\sin^2\gamma}{\sinh^2 c}\,.$$

Taking (positive) square roots gives the hyperbolic sine law. □

Remark 14.6. For small triangles $\sinh a \equiv a$, $\sinh b \equiv b$ and $\sinh c \equiv c$ modulo terms of order 3 or higher, so the hyperbolic sine law tends to the Euclidean sine law

$$\frac{\sin\alpha}{a} = \frac{\sin\beta}{b} = \frac{\sin\gamma}{c}\,.$$

These two laws suffice to determine all sides and angles in a hyperbolic triangle when all three sides are known (SSS), or when two sides and one angle are known (SAS or SSA), except for an ambiguity in recovering an angle from its sine in the side-side-angle case. In the Euclidean case two angles determine the third, and knowing three angles only determines a triangle up to similarity. In the hyperbolic case there is instead a second cosine law, dual to the first law of cosines. It lets us determine all sides and angles in a hyperbolic triangle when all three angles are known (AAA), or when two angles and one side are known (ASA or AAS).

Theorem 14.7 (The second hyperbolic law of cosines)**.**

$$\cos\alpha = -\cos\beta\,\cos\gamma + \sin\beta\,\sin\gamma\,\cosh a\,.$$

Sketch proof. Using the first cosine law and a step in the proof of the sine law we can express

$$\cos\alpha + \cos\beta\,\cos\gamma$$

and

$$\sin\beta\,\sin\gamma$$

in terms of $\cosh a$, $\cosh b$ and $\cosh c$. After some calculation, the ratio of the two expressions is seen to be $\cosh a$. [See Proposition 2.9.4 in Jahren's book.] □

Remark 14.8. For small triangles $\cosh a$ is close to 1, so the second cosine law tends to the relation

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

This recovers the sum-of-angles relation

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

from flat Euclidean geometry. In the Euclidean case knowledge of the three angles only determines a triangle up to similarity. In the hyperbolic case there are no similarity transformations, other than the congruences.

15. October 9th lecture

15.1. Topological surfaces.

Definition 15.1. A *topological surface* is a topological space M that is locally homeomorphic to $\mathbb{R}^2$. We shall also assume that each topological surface is Hausdorff and second countable.

In other words, we assume that each point in M has an open neighborhood that is homeomorphic to (an open subset of) $\mathbb{R}^2$, that distinct points in M lie in disjoint neighborhoods, and that there is a countable basis for the topology of M. These conditions ensure that M is a metrizable topological space, and that M can be embedded in some Euclidean space $\mathbb{R}^N$, i.e., that M is homeomorphic to a subset of $\mathbb{R}^N$ in the subspace topology.

Example 15.2. Here are some examples of topological surfaces:

- The unit sphere S^2 consisting of vectors $x \in \mathbb{R}^3$ of length 1.
- The real projective plane $P^2 = \mathbb{R}P^2 = S^2/\!\sim$ where $x \sim -x$.
- The Euclidean plane $\mathbb{R}^2 \cong \mathbb{C}$.
- The torus $T^2 = \mathbb{R}^2/\!\sim$ where $(x, y) \sim (x+1, y)$ and $(x, y) \sim (x, y+1)$.
- The Klein bottle $K^2 = \mathbb{R}^2/\!\sim$ where $(x, y) \sim (x+1, y)$ and $(x, y) \sim (1-x, y+1)$.
- The hyperbolic plane $\mathbb{D} \cong \mathbb{H}$.
- The genus two surface obtained from an octagon $ABCDEFGH$ by identifying AB with DC, BC with ED, EF with HG and FG with AH.

- The zero set $M = \{(x, y, z) \in \Omega \mid f(x, y, z) = 0\}$ of any C^1 function $f \colon \Omega \to \mathbb{R}$ such that $\nabla f = (\partial f/\partial x, \partial f/\partial y, \partial f/\partial z) \neq 0$ on M. Here Ω is an open subset of $\mathbb{R}^3$. For instance, we might have $f(x, y, z) = x^2 + y^2 + z^2 - 1$, with $\nabla f = (2x, 2y, 2z)$, so that $M = S^2$. As another instance, we might have $f(x, y, z) = z - g(x, y)$, for some C^1 function $g \colon U \to \mathbb{R}$ with U open in $\mathbb{R}^2$, in which case $\nabla f = (\partial g/\partial x, \partial g/\partial y, 1)$ and M is the graph of g.
- The zero set $M = \{(z, w) \in \Omega \mid h(z, w) = 0\}$ of any complex analytic function $h \colon \Omega \to \mathbb{C}$ such that $(\partial h/\partial z, \partial h/\partial w) \neq 0$ on M. Here Ω is an open subset of $\mathbb{C}^2$. This defines a complex curve, which is a real surface. For instance, we might have $h(z, w) = w^2 - (z^3 + az + b)$ with $4a^3 + 27b^2 \neq 0$, so that M is a plane elliptic curve.
- Any open subset of the previous examples.
- Any (finite or countable) disjoint union of the previous examples.

Definition 15.3. A *closed surface* is a topological surface that it compact as a topological space.

Remark 15.4. This terminology reflects the fact that there exists a more general notion of surface with boundary, and the term compact surface usually refers to a surface, with or without boundary, that is compact. The term closed surface refers to a surface without boundary that is compact. We concentrate on surfaces without boundary.

Definition 15.5. Given two connected (topological) surfaces M_1 and M_2 we can choose an embedding $h_i \colon D^2 \to M_i$ for each $i = 1, 2$. Then the closed subspace

$$M'_i = M_i \setminus h_i(\operatorname{int} D^2)$$

is locally homeomorphic to $\mathbb{R}^2$, except at the "boundary" $h_i(S^1) \subset M'_i$. The *connected sum* of M_1 and M_2 is the identification space

$$M_1 \# M_2 = M'_1 \cup_{S^1} M'_2$$

of these two pieces, where we use h_i to identify S^1 with $h_i(S^1) \subset M'_i$ for $i = 1, 2$.

Lemma 15.6. *$M_1 \# M_2$ is a connected (topological) surface. If M_1 and M_2 are closed, then so is $M_1 \# M_2$.*

Sketch proof. The main thing to check is that $M_1 \# M_2$ is locally homeomorphic to $\mathbb{R}^2$ at the common image of $h_1(S^1)$ and $h_2(S^1)$. □

[[Discuss non-dependence of $M_1 \# M_2$ on the choices of h_1 and h_2.]]

[[Similar construction for combinatorial surfaces, removing the interior of a triangle on each side, and identifying the two boundary triangles.]]

Lemma 15.7. *There are homeomorphisms*

- $M_1 \# S^2 \cong M_1$,
- $M_1 \# M_2 \cong M_2 \# M_1$ *and*
- $(M_1 \# M_2) \# M_3 \cong M_1 \# (M_2 \# M_3)$.

Hence the set of homeomorphism classes of connected surfaces becomes a commutative monoid with respect to the connected sum pairing, with neutral element given by the class of S^2.

Proof. For the first case, note that we can choose the embedding $h_2 \colon D^2 \to S^2 = M_2$ so that $h_2(D^2)$ is one hemisphere and $M'_2 = S^2 \setminus h_2(\operatorname{int} D^2)$ is the other hemisphere, meeting at the equator. Any choice of homeomorphisms between these hemispheres, restricting to the identity on their common boundary, induces a homeomorphism between M_1 and $M_1 \# S^2$.

The proofs in the other two cases are similar. □

Theorem 15.8 (Classification of closed surfaces)**.** *Each closed, connected (topological) surface M is homeomorphic to a connected sum*

$$M \cong S(g, h) = \underbrace{T^2 \# \dots \# T^2}_{g \text{ copies}} \# \underbrace{P^2 \# \dots \# P^2}_{h \text{ copies}}$$

of $g \geq 0$ copies of the torus $T^2 = S^1 \times S^1$ and $h \geq 0$ copies of the real projective plane $P^2 = \mathbb{R}P^2$. The only relations among these homeomorphism classes are generated by the identity

$$S(1,1) = T^2 \# P^2 \cong P^2 \# P^2 \# P^2 = S(0,3)\,,$$

so that $S(g,h) \cong S(0, 2g+h)$ if $h \geq 1$. Hence each closed, connected surface M is homeomorphic to exactly one of the following standard surfaces:

- *the orientable surface $M_g = S(g,0) = T^2 \# \dots \# T^2$ (g copies) of genus $g \geq 0$, or*
- *the non-orientable surface $N_h = S(0,h) = P^2 \# \dots \# P^2$ (h copies) of demigenus $h \geq 1$.*

Note that $S^2 = M_0 = S(0,0)$ is treated as the connected sum of zero surfaces, and the Klein bottle appears as $K^2 = S(0,2) = P^2 \# P^2$.

We shall prove the existence part of this classification, assuming that the surface can be triangulated in a sense that we will soon make precise. We will sketch the uniqueness part of the classification, using tools (Euler characteristic, orientability) that are shown to be well-defined in the first algebraic topology course (MAT4530).

16. October 14th lecture

16.1. Combinatorial surfaces.

To get a handle on general topological surfaces, it is useful to first equip them with a more combinatorial structure.

Definition 16.1. An *n-simplex* in some Euclidean space $\mathbb{R}^N$ is the convex hull σ of $(n+1)$ points $v_0, v_1, \dots, v_n$ that are not contained in any $(n-1)$-dimensional affine subspace. The convex hull of a non-empty subset of $\{v_0, v_1, \dots, v_n\}$ is called a *face* of σ, and is again a simplex in $\mathbb{R}^N$.

Example 16.2. A 0-simplex is called a vertex, a 1-simplex is called an edge, and a 2-simplex is called a triangle. The proper faces of an edge are its two endpoints. The proper faces of a triangle are its three side edges and its three vertices.

Definition 16.3. A *simplicial complex* is a collection Σ of simplices in some Euclidean space $\mathbb{R}^N$ such that any face of a simplex in Σ is again in Σ, and the intersection of two simplices in Σ is either empty or a face of both. (If Σ is infinite, we also require that the collection is locally finite, in the sense that each point of $\mathbb{R}^N$ has a neighborhood that meets only finitely many simplices in the collection. Two sets meet if their intersection is not empty.)

The union of the simplices in Σ, as a subspace of $\mathbb{R}^N$, is called the *polyhedron* of Σ, and is denoted $|\Sigma|$. A *triangulation* of a topological space X is a choice of a simplicial complex Σ and a homeomorphism $h \colon |\Sigma| \to X$.

Example 16.4. Here are some examples of simplicial complexes.

- A tetrahedron $T \subset \mathbb{R}^3$, which is a 3-simplex, together with all its proper faces, is a simplicial complex Δ^3. Its polyhedron $|\Delta^3|$ is equal to T, which is homeomorphic to D^3.
- The subcollection $\partial\Delta^3$ consisting only of the proper faces of T is also a simplicial complex. Its polyhedron $|\partial\Delta^3|$ is equal to the topological boundary ∂T of T in $\mathbb{R}^3$, which is homeomorphic to S^2.
- A simplicial complex with polyhedron homeomorphic to S^1 consists of k vertices $v_1, \dots, v_k$ and k edges $[v_1, v_2], \dots, [v_{k-1}, v_k]$ and $[v_k, v_1]$, for some $k \geq 3$.

Lemma 16.5. *A simplicial complex consists of finitely many simplices if and only if its polyhedron is compact.*

Definition 16.6. Let v be a vertex in a simplicial complex Σ. The collection of simplices in Σ that contain v is called the (open) *star* of v. The simplicial complex consisting of the simplices that contain v, and all of their faces, is called the *closed star* of v. The complement of the open star in the closed star, consisting of the simplices that do not contain v but are faces of simplices that contain v, is a simplicial complex called the *link* of v in Σ.

Example 16.7. Here are some examples of links.

- The link of a vertex v in Δ^3 is the triangular face of T opposite to v. Its polyhedron is homeomorphic to D^2.
- The link of v in $\partial\Delta^3$ is the boundary of that triangular face of T. Its polyhedron is the union of the three edges of the triangle, and is homeomorphic to S^1.

Definition 16.8. A *combinatorial surface* is a simplicial complex Σ such that the link of each vertex has polyhedron homeomorphic to S^1.

Remark 16.9. A topological surface is the special case $n = 2$ of a topological n-manifold. A combinatorial surface is the special case $n = 2$ of a combinatorial n-manifold.

Lemma 16.10. *The polyhedron of a combinatorial surface is a topological surface.*

Proof. By the cone of a space X we mean the identification space $CX = X \times [0,1]/\sim$, where $(x,0) \sim (y,0)$ for all $x, y \in X$. It contains a homeomorphic copy of X as the subspace $X \times \{1\}$. The cone on S^1 is homeomorphic to D^2, and the complement $CS^1 \setminus S^1$ is homeomorphic to the open disc $\operatorname{int} D^2 = D^2 \setminus S^1$ in $\mathbb{R}^2$.

The polyhedron of the closed star of a vertex v is homeomorphic to the cone on the polyhedron of its link, so in a combinatorial manifold these are homeomorphic to D^2 and the open stars are homeomorphic to $\operatorname{int} D^2$. These open stars cover the polyhedron, which is therefore locally homeomorphic to $\mathbb{R}^2$.

Any polyhedron is a subspace of some $\mathbb{R}^N$, hence metrizable, and therefore Hausdorff and second countable. □

The converse is a much harder theorem of Tibor Radó from 1925.

Theorem 16.11 (Radó)**.** *Each topological surface admits a triangulation, i.e., is homeomorphic to the polyhedron of a combinatorial surface.*

Remark 16.12. We will not prove this result. Carsten Thomassen (Amer. Math. Monthly, 1992) gave a short proof using the Jordan–Schönflies theorem. Allen Hatcher (arXiv, 2013) wrote up a proof using the Kirby torus trick, which proves existence and uniqueness of smooth structures on topological surfaces, and which implies existence of triangulations. Without these results the arguments that follow will only provide a topological classification of triangulable surfaces, i.e., of surfaces that can be given a combinatorial structure.

Radó also proves that any two triangulations of the same topological surface are equivalent, in the sense that they admit a common subdivision, so there is a one-to-one correspondence between equivalence classes of topological structures and combinatorial structures on any surface. A corresponding result also holds for 3-dimensional manifolds (Moise (1952), Bing (1954, 1959)), but is false in higher dimensions. Kirby and Siebenmann (1969) give examples of 5-manifolds with two inequivalent combinatorial structures, and of topological 6-manifolds that cannot be triangulated as combinatorial manifolds. [[What is the situation for 4-manifolds?]]

17. October 16th lecture

17.1. Gluings.

Definition 17.1. A *gluing pattern* consists of a finite set of triangles

$$\Delta_1, \Delta_2, \dots, \Delta_n ,$$

a pairing of the edges of these triangles such that each edge appears in exactly one of the pairs, and a choice of one of the two possible (affine) linear identifications between the edges in each pair.

The associated *gluing space* is the identification space

$$M = (\Delta_1 \sqcup \Delta_2 \sqcup \cdots \sqcup \Delta_n)/\sim$$

obtained from the disjoint union of these triangles, where $\sim$ is the equivalence relation generated by the chosen linear identifications between the edges in each pair.

For instance, if an edge $[A, B] \subset \Delta_1$ is paired with an edge $[C, D] \subset \Delta_2$, with the choice of linear bijection taking A to D and B to C, then $(1-t)A + tB$ is identified under $\sim$ with $(1-t)D + tC$, for each $t \in [0, 1]$.

Lemma 17.2. *The polyhedron $|\Sigma|$ of a finite combinatorial surface Σ can be obtained as the gluing space M of a gluing pattern on its set of triangles.*

Proof. Let $\Delta_1, \Delta_2, \dots, \Delta_n$ be the set of triangles in Σ. The gluing pattern is defined so that two edges in $\Delta_1, \Delta_2, \dots, \Delta_n$ are paired if and only if they represent the same edge in Σ. Each edge E of Σ corresponds to a (link) vertex in the link of either one of its endpoints, and since the link has polyhedron homeomorphic to S^1, that (link) vertex is the meeting point of precisely two (link) edges. These correspond to precisely two triangles Δ_i and Δ_j in Σ, both containing E as an edge. Hence the gluing pattern pairs each edge in $\Delta_1, \Delta_2, \dots, \Delta_n$ with precisely one other edge in the same collection.

We get a well-defined map $M \to |\Sigma|$ from the gluing space of the gluing pattern to the polyhedron of Σ. It is clearly bijective away from the vertices, and surjective on vertices. To check that the map is injective on vertices, use that the link of any vertex v is connected, to see that any two representatives of that vertex in the disjoint union $\Delta_1 \sqcup \Delta_2 \sqcup \dots \sqcup \Delta_n$ are in fact equivalent under $\sim$. □

Notice that since the $3n$ edges of the n triangles can be grouped into disjoint pairs, n must be an even integer.

Lemma 17.3. *If the gluing space M associated to a gluing pattern with n triangles is connected, then it is homeomorphic to an identification space $F/\sim$, where F is an $(n+2)$-gon, homeomorphic to D^2, and $\sim$ is an equivalence relation generated by pairwise (affine) linear identifications of the $(n+2)$ edges of F.*

Proof. We may choose the ordering of $\Delta_1, \Delta_2, \dots, \Delta_n$ so that each Δ_k has at least one edge E_k in common with the union $\Delta_1 \cup \dots \cup \Delta_{k-1}$, for $2 \le k \le n$. (This uses that M is connected.) Let $F_1 = \Delta_1$ and inductively define

$$F_k = F_{k-1} \cup_{E_k} \Delta_k \,.$$

Then each F_k is a $(k+2)$-gon, since one edge in F_{k-1} is replaced by two new edges in F_k. In particular, $F = F_n$ is an $(n+2)$-gon, homeomorphic to D^2, and the map from F to the gluing space M only identifies the remaining pairs of edges that did not occur in the list $E_2, \dots, E_n$. □

Definition 17.4. Given an $(n+2)$-gon $F \cong D^2$, equipped with pairwise identifications of its $(n+2)$ edges, we label the edges with symbols a or a^{-1}, in such a way that two edges that get identified have the same labels (both a, or both a^{-1}) if the identification preserves the direction of travel around the boundary of F, and have inverse labels (one a and one a^{-1}) if the identification reverses the direction of travel around that boundary.

Let $W = a \dots a^{\pm 1} \dots$ be the word obtained by concatenating the labels around the boundary of F, in the direction corresponding to a counterclockwise lap around the boundary of D^2. Let D^2/W denote the identification space $F/\sim$ given by the equivalence relation encoded by the word W. Let W^{-1} denote the corresponding word read clockwise, with each label replaced with its inverse, and in the reversed ordering. A word W of length $(n+2)$, with $(n+2)/2$ pairs of distinct letters, or their inverses, will be called *admissible*. We also permit the empty word, $W = \{\}$, and define $D^2/\{\} = S^2$.

Example 17.5. Here are some examples of admissible words W and the associated gluing spaces $F/\sim = D^2/W$:

- $D^2/aa^{-1} \cong S^2$.
- $D^2/aa \cong P^2$.
- $D^2/aba^{-1}b^{-1} \cong T^2$.
- $D^2/aba^{-1}b \cong K^2$.

If W_1 and W_2 are words, we write W_1W_2 for their concatenation.

Lemma 17.6. *(a) If $W = W_1W_2$ is admissible, then so is W_2W_1 and $D^2/W_1W_2 \cong D^2/W_2W_1$.*

(b) If W_1 and W_2 are admissible, then so is W_1W_2 and $D^2/W_1W_2 \cong D^2/W_1\#D^2/W_2$.

(c) If W_1xW_2x is admissible then so is $W_1W_2^{-1}yy$ (with x and y not occurring in W_1 and W_2).

Proof. (a) This corresponds to starting to read the labels around the boundary of F at a different point.

(b) Let z be a line segment in F from the beginning of W_1 to its end, or equivalently, from the end of W_2 to its beginning. Since these words are admissible, the endpoints of z are identified in $F/{\sim} = D^2/W_1W_2$, with image $z/{\sim} = Z$, say. Write $F = F_1 \cup_z F_2$, with each of F_1 and F_2 a polygon. If we make the identifications along the boundary of F_1 that are specified by W_1 we get a space homeomorphic to $D^2/W_1 \setminus \operatorname{int}\Delta^2$, where we have removed an open triangle with boundary given by the image Z of z (broken into three edges). Likewise for F_2 with the identifications specified by W_2. Hence

$$F/{\sim} \cong F_1/{\sim} \cup_Z F_2/{\sim} \cong (D^2/W_1 \setminus \operatorname{int}\Delta^2) \cup_Z (D^2/W_2 \setminus \operatorname{int}\Delta^2) \cong D^2/W_1\#D^2/W_2\,.$$

(c) Let y be a line segment in F from the end of the first edge labeled x to the end of the second edge labeled x. Cutting F apart along y we get two polygons F_1 and F_2, the first with boundary W_1xy and the second with boundary W_2xy^{-1}. Gluing F_1 and F_2 together along x (with F_2 turned upside-down), we get a new polygon with boundary $W_1W_2^{-1}yy$. Hence there is a homeomorphism $D^2/W_1xW_2x \cong D^2/W_1W_2^{-1}yy$. □

18. October 21st lecture

18.1. Proof of the classification theorem.

Proof of relations. The relation $T^2\#P^2 \cong P^2\#P^2\#P^2$ among the topological types (= homeomorphism classes) of closed connected surfaces is a consequence of the following two applications of the lemma above:

- $D^2/abab^{-1} \cong D^2/bbyy \cong D^2/bb\#D^2/yy$, so $K^2 \cong P^2\#P^2$.
- $D^2/aba^{-1}b^{-1}xx \cong D^2/acbab^{-1}c \cong D^2/cbab^{-1}ca \cong D^2/cbc^{-1}byy$, so $T^2\#P^2 \cong K^2\#P^2$.

The general relation $S(g,h) \cong S(0,2g+h)$ for $h \geq 1$ follows from this by an easy induction on $g \geq 0$. □

Proof of existence. We now prove that for any connected closed surface M there exists a standard surface $S(g,h)$ such that $M \cong S(g,h)$, i.e., such that M is a connected sum of $g \geq 0$ copies of T^2 and $h \geq 0$ copies of P^2.

By Radó's theorem we may assume that M has been triangulated, hence is homeomorphic to the gluing space associated to a gluing pattern. That gluing space may be written as $F/{\sim} = D^2/W$ for some admissible word W.

Step 1: If any letter x occurs twice in W, so that $W = W_1xW_2xW_3$, we can use case (a) in the lemma above to write D^2/W as $D^2/W_3W_1xW_2x$, and then use case (c) to write $D^2/W_3W_1xW_2x$ as $D^2/W'yy$ with $W' = W_3W_1W_2^{-1}$. By case (b) we have $D^2/W'yy \cong D^2/W'\#D^2/yy$, where $D^2/yy \cong P^2$, so we have recognized M as $D^2/W'\#P^2$ where W' is a shorter admissible word than the initial word W.

In the same way, if any letter x^{-1} occurs twice in W, then we can also write M as $D^2/W'\#P^2$ where W' is shorter than W. We can continue this way until no letter x occurs twice in the word W', nor does any inverse letter x^{-1} occur twice. If Step 1 is taken $h \geq 0$ times to get to this point, we have shown that $M \cong M'\#S(0,h)$, and $M' \cong D^2/W'$ for an admissible word W' such that each letter in W' only occurs together with its inverse.

Step 2: If a letter x is followed by its inverse x^{-1} in W', in the cyclic sense, so that $W' = W_1xx^{-1}W_2$ or $W' = x^{-1}W''x$, then $D^2/W' \cong D^2/W''xx^{-1}$ by case (a), with $W'' = W_2W_1$, and $D^2/W''xx^{-1} \cong D^2/W''\#D^2/xx^{-1}$ by case (b). Here $D^2/xx^{-1} \cong S^2$, so $D^2/W''\#D^2/xx^{-1} \cong$

$D^2/W''\#S^2 \cong D^2/W''$. A similar argument applies if x and x^{-1} are interchanged. Hence in each of these cases we may replace W' with a shorter admissible word (where each letter only occurs together with its inverse), without changing the topological type of D^2/W'.

Step 3: If no letter x is adjacent to its inverse x^{-1} in W', still in the cyclic sense, there must be two letters x and y occurring in the order $\dots x \dots y \dots x^{-1} \dots y^{-1} \dots$ in W', at least up to a cyclic reordering. To see this, assume that x and x^{-1} are as close as possible in the cyclic ordering, and let y (or y^{-1}) be one of the letters between x and x^{-1}. Then y^{-1} (or y) cannot also be between x and x^{-1}, since that would mean that y and y^{-1} were closer together than x and x^{-1}, contradicting the choice of x. If the letters occur in the order $\dots x \dots y^{-1} \dots x^{-1} \dots y \dots$ we cyclically permute the y to the front, and interchange the roles of x and y. This way we may assume that the letters occur in the order $\dots x \dots y \dots x^{-1} \dots y^{-1} \dots$. Cycling x to the front of W', we may assume that $W' = xW_1yW_2x^{-1}W_3y^{-1}W_4$.

Now cut the polygon F along a line segment a going from the beginning of x to the beginning of x^{-1}. The resulting two pieces are polygons with boundaries $xW_1yW_2x^{-1}a^{-1}$ and $aW_3y^{-1}W_4$. Glue these together along y and y^{-1} to get a new polygon F' with boundary $xW_1W_4aW_3W_2x^{-1}a^{-1}$. Thereafter cut F' along a line segment b going from the end of a to the end of a^{-1}. The resulting two pieces are polygons with boundaries $W_3W_2x^{-1}b^{-1}$ and $ba^{-1}xW_1W_4a$. Glue these together along x and x^{-1} to get a final polygon F'' with boundary $W_3W_2W_1W_4aba^{-1}b^{-1}$. Hence there are homeomorphisms $D^2/W' \cong D^2/W''aba^{-1}b^{-1} \cong D^2/W''\#D^2/aba^{-1}b^{-1}$, with $W'' = W_3W_2W_1W_4$ admissible. Here $D^2/aba^{-1}b^{-1} \cong T^2$, so $D^2/W' \cong D^2/W''\#T^2$ where W'' is an admissible word (where each letter only occurs together with its inverse), shorter than W'.

We repeat Steps 2 and 3 until $W'' = \{\}$ is the empty word. If Step 3 was taken $g \geq 0$ times, we have identified $M' = D^2/W'$ with $S^2\#T^2\#\dots\#T^2 \cong S(g,0)$, i.e., the connected sum of g copies of T^2. To conclude, there exist non-negative integers $g, h \geq 0$ such that $M \cong M'\#S(0,h) \cong S(g,0)\#S(0,h) \cong S(g,h)$ is homeomorphic to one of the standard surfaces obtained from finitely many copies of T^2 and P^2 by the formation of connected sums. □

19. October 28th lecture

Sketch proof of uniqueness. We must prove that the surfaces $M_g = S(g,0)$ for $g \geq 0$ and $N_h = S(0,h)$ for $h \geq 1$ are all topologically distinct.

The surfaces M_g are all orientable. An orientation of a surface is a compatible choice of local orientations at each point of the surface, and a local orientation at a point is a choice of which of the two possible directions of travel around that point (say for a simple closed curve contained in a coordinate chart around the point) that counts as the positive direction. Any surface of the form D^2/W where each letter x in the admissible word W only occurs together with its inverse x^{-1} is orientable. This is because the usual orientation of D^2 inherited from $\mathbb{R}^2$, where the counterclockwise direction of travel is declared to be the positive one, descends to a well-defined orientation on D^2/W for these words.

On the other hand, the surfaces N_h with $h \geq 1$ are not orientable. Any surface of the form D^2/W where some letter x occurs twice in the word W is not orientable. This is because either choice of orientation for D^2 gets reversed when we move across the boundary of $F \cong D^2$ along one edge labeled x and return to D^2 across the other edge labeled x. A closed neighborhood of the line segment in F connecting the mid-points of the two edges labeled x is homeomorphic to a square $I \times I$, but in $F/\!\sim$ a pair of opposite sides are identified, after a half-twist, yielding a non-orientable Möbius band within the closed surface.

It follows that we cannot have $M_g \cong N_h$ for some $g \geq 0$ and $h \geq 1$, since two homeomorphic surfaces are either both orientable or both non-orientable.

To show that M_g and $M_{g'}$ cannot be homeomorphic for $g \neq g'$, and that N_h and $N_{h'}$ cannot be homeomorphic for $h \neq h'$, we need another topological invariant. One such invariant is the Euler characteristic. For a compact triangulated surface $M \cong |\Sigma|$ this is the number

$$\chi(M) = v - e + f$$

where v is the number of vertices (or 0-simplices) of the combinatorial surface Σ, e is the number of edges (or 1-simplices) of Σ, and f is the number of triangles (or 2-simplices) of Σ. The fact that remains to be proved is that $\chi(M)$ is a topological invariant, so that if $M \cong M'$ then $\chi(M) = \chi(M')$.

Assuming this, we can finish the proof. The sphere S^2 admits a triangulation (as $\partial\Delta^3$) with $v = 4$, $e = 6$ and $f = 4$, so $\chi(S^2) = 2$. The projective plane P^2 has Euler characteristic $\chi(P^2) = 1$, and the torus T^2 has Euler characteristic $\chi(T^2) = 0$. The latter two facts are perhaps easier to see if we grant that the Euler characteristic can also be computed from a model $F/\!\sim\; \cong D^2/W$, where v is the number of vertices on the boundary of F, after the identifications $\sim$ have been made, e is the number of edges on the boundary of F, also after the identifications $\sim$ have been made, and $f = 1$ (for the single convex 2-cell F). With the model D^2/aa for P^2 we get $v = 1$, $e = 1$ and $f = 1$, so $\chi(P^2) = 1$. With the model $D^2/aba^{-1}b^{-1}$ for T^2 we get $v = 1$, $e = 2$ and $f = 1$, so $\chi(T^2) = 0$.

The connected sum $M_1 \# M_2 = M_1' \cup_{S^1} M_2'$ of two triangulated surfaces can be constructed by removing the interior of a triangle (2-simplex) of each, and identifying the remaining three edges and vertices on one side with those on the other side. Thus $\chi(M_1 \# M_2) = \chi(M_1') + \chi(M_2') + (3 - 3) = \chi(M_1) - 1 + \chi(M_2) - 1 = \chi(M_1) + \chi(M_2) - 2$. This leads to the formulas $\chi(M_g) = 2 - 2g$ for $g \geq 0$ and $\chi(N_h) = 2 - h$ for $h \geq 1$. In particular, $\chi(M_g)$ uniquely determines the number g, and $\chi(N_h)$ uniquely determines the number h. □

20. October 30th lecture

20.1. **Charts and atlases.** A topological surface M is, in particular, a topological space, and this structure carries enough information to make sense of what we mean by saying that a function $f\colon M \to \mathbb{R}$ is continuous. Since M is assumed to be locally homeomorphic to $\mathbb{R}^2$, this condition can be translated into a collection of conditions on functions on open subsets of $\mathbb{R}^2$. To state this in detail, let us spell out in more detail what it means for M to be locally homeomorphic to $\mathbb{R}^2$:

We require that for each point $p \in M$ there exist an open neighborhood $U \subseteq M$ (of $p \in U$), an open subset $V \subseteq \mathbb{R}^2$, and a homeomorphism $\phi\colon U \xrightarrow{\cong} V$. If we view ϕ as a map $U \to V \subseteq \mathbb{R}^2$, it induces a homeomorphism $\phi\colon U \xrightarrow{\cong} \phi(U)$.

Definition 20.1. We call the open subset U a *coordinate patch*, the components of $\phi = (\phi_1, \phi_2)$ are called *coordinates* on U, and the pair (U, ϕ) is a *chart* on M. As p ranges through M, the collection of coordinate patches U forms an open cover of M. The corresponding collection of charts is then called an *atlas*.

Conversely, if we select an open covering $\{U_i\}_i$ of M, where i runs through some set of indices, and each U_i is homeomorphic to an open subset V_i of $\mathbb{R}^2$ by a homeomorphism

$$\phi_i\colon U_i \xrightarrow{\cong} V_i \subset \mathbb{R}^2\,,$$

then each point p lies in at least one U_i, which then serves as the open neighborhood of p required by the condition that M should be locally homeomorphic to $\mathbb{R}^2$. The corresponding atlas can then be denoted $\{(U_i, \phi_i)\}_i$.

A real function $f\colon M \to \mathbb{R}$ is continuous if and only if each restriction $f|U_i\colon U_i \to \mathbb{R}$ is continuous, since $\{U_i\}_i$ is an open cover of M. Since each ϕ_i is a homeomorphism, this is equivalent to the condition that each composite

$$f|U_i \circ \phi_i^{-1}\colon V_i \longrightarrow \mathbb{R}$$

is continuous. Since each V_i is an open subset of $\mathbb{R}^2$, this is just the usual condition that a map from an open subset of $\mathbb{R}^2$ to $\mathbb{R}$ is continuous.

20.2. **Differentiable atlases.** Now suppose that we want to make sense of what it means for a function $f\colon M \to \mathbb{R}$ to be differentiable. Our strategy will be to ask that for each i the composite

$$f|U_i \circ \phi_i^{-1}\colon V_i \longrightarrow \mathbb{R}$$

is differentiable. Since each V_i is an open subset of $\mathbb{R}^2$, this should just be the usual condition that a map from an open subset of $\mathbb{R}^2$ to $\mathbb{R}$ is differentiable, which can be expressed by the existence of a sufficiently good linear approximation at each point. However, in this generality it may happen that a function $f\colon M \to \mathbb{R}$ appears to be differentiable at a point $p \in M$ from the point of view of one chart (U_i, ϕ_i) with $p \in U_i$, but not to be differentiable from the point of view of another chart (U_j, ϕ_j). More precisely, it may happen that

$$f|U_i \circ \phi_i^{-1}\colon V_i \longrightarrow \mathbb{R}$$

is differentiable at $\phi_i(p)$, but also that

$$f|U_j \circ \phi_j^{-1}\colon V_j \longrightarrow \mathbb{R}$$

is not differentiable at $\phi_j(p)$. This ambiguity can happen for all p in $U_{ij} = U_i \cap U_j$. To avoid this ambiguity, we can assume that the coordinate transformation

$$(\phi_i|U_{ij}) \circ (\phi_j|U_{ij})^{-1}\colon \phi_j(U_{ij}) \longrightarrow \phi_i(U_{ij})$$

is differentiable, as a map from the open subset $\phi_j(U_{ij}) \subseteq \phi_j(U_j) = V_j$ of $\mathbb{R}^2$ to the open subset $\phi_i(U_{ij}) \subseteq \phi_i(U_j) = V_i$ of $\mathbb{R}^2$. Then, if $f|U_i \circ \phi_i^{-1}\colon V_i \longrightarrow \mathbb{R}$ is differentiable at $\phi_i(p)$, then so is its restriction $f|U_{ij} \circ (\phi_i|U_{ij})^{-1}\colon \phi_i(U_{ij}) \to \mathbb{R}$. By the chain rule, it follows that the composite

$$f|U_{ij} \circ (\phi_i|U_{ij})^{-1} \circ (\phi_i|U_{ij}) \circ (\phi_j|U_{ij})^{-1} = f|U_{ij} \circ (\phi_j|U_{ij})^{-1}$$

is differentiable at $\phi_j(p)$, as a map $\phi_j(U_{ij}) \to \mathbb{R}$. Hence $f|U_j \circ \phi_j^{-1}\colon V_j \to \mathbb{R}$ is also differentiable at $\phi_j(p)$.

Definition 20.2. If the atlas $\{(U_i, \phi_i)\}_i$ has the property that each *coordinate transformation* $(\phi_i|U_{ij}) \circ (\phi_j|U_{ij})^{-1}$ is differentiable, for any pair of indices i and j, then the condition given for a map $f\colon M \to \mathbb{R}$ to be differentiable at a point $p \in M$ will be independent of the choice of coordinate chart, as long as the charts are chosen from this atlas. Such an atlas will be called a *differentiable atlas*.

The choice of a differentiable atlas is an additional structure, or piece of data, that we can associate with a topological surface M, and which grants us the additional ability to specify which real functions f on M that are differentiable (at the various points of M).

A choice of differentiable atlas $\{(U_i, \phi_i)\}_i$ also lets us make sense of which curves in M are differentiable. To ease the notation, we hereafter use the same symbol for a map and its restrictions to subsets in the domain (= source) or codomain (= target). A continuous map $\omega\colon [a,b] \to M$ is differentiable at a point $t \in [a,b]$ if for any chart (U_i, ϕ_i) with $\omega(t) \in U_i$ the composite map

$$\phi_i \circ \omega\colon \omega^{-1}(U_i) \longrightarrow \mathbb{R}^2$$

is differentiable at t, as a map from an open neighborhood of $t \in [a,b] \subset \mathbb{R}$ to $\mathbb{R}^2$. A different choice of chart (U_j, ϕ_j) in the same atlas, also with $\omega(t) \in U_j$, leads to the same notion of differentiability, since the composite

$$\phi_j \circ \omega = \phi_j \circ \phi_i^{-1} \circ \phi_i \circ \omega\colon \omega^{-1}(U_{ij}) \longrightarrow \mathbb{R}^2$$

will be differentiable at t if $\phi_i \circ \omega$ is differentiable at t and $\phi_j \circ \phi_i^{-1}$ is differentiable at $\phi_i(\omega(t))$, and the last condition always holds, by definition, for a differentiable atlas.

To each differentiable curve $\omega\colon [a,b] \to M$ we will be able to associate a tangent vector $v = \omega'(t)$ to M at $p = \omega(t)$, and the set of all tangent vectors at p will form a tangent plane T_pM, which is a 2-dimensional vector space. To each differentiable function $f\colon M \to \mathbb{R}$ we will be able to associate a differential df, which at each point $p \in M$ gives a linear functional $df_p\colon T_pM \to \mathbb{R}$. These constructions will be essential for doing differential topology, differential geometry and Riemannian geometry on surfaces.

20.3. **Differentiable surfaces.** A topological surface M together with a choice of differentiable atlas $\{(U_i, \phi_i)\}_i$ therefore determines what we will call a differentiable surface. However, there are many different choices of differentiable atlases that lead to the same notions of differentiable functions on M and curves in M. In fact, a second differentiable atlas $\{(U'_k, \phi'_k)\}_k$ will specify the same notion of differentiability for functions $f \colon M \to \mathbb{R}$ if and only if the union of the two collections of charts

$$\{(U_i, \phi_i)\}_i \cup \{(U'_k, \phi'_k)\}_k$$

is again a differentiable atlas. In addition to the conditions that the coordinate transformations $\phi_i \circ \phi_j^{-1}$ and $\phi'_k \circ {\phi'_\ell}^{-1}$ within the first and second atlases, respectively, are differentiable maps, this amounts to the condition that the transformations

$$\phi_i \circ {\phi'_k}^{-1} \colon \phi'_k(U_i \cap U'_k) \longrightarrow \phi_i(U_i \cap U'_k)$$

between the atlases, and their inverses

$$\phi'_k \circ \phi_i^{-1} \colon \phi_i(U_i \cap U'_k) \longrightarrow \phi'_k(U_i \cap U'_k)\,,$$

are differentiable, for all indices i and k.

We say that two differentiable atlases on M are equivalent if their union is again a differentiable atlas. A differentiable surface can then be defined as a topological surface together with a choice of an equivalence class of differentiable atlases.

Each equivalence class of differentiable atlases contains a preferred element, namely the maximal differentiable atlas given by the union of all the differentiable atlases in the equivalence class. This differentiable atlas is maximal in the sense that it is impossible to add any further charts (U, ϕ) to it without breaking the condition than each coordinate transformation is differentiable.

Definition 20.3. A *differentiable surface* is a topological surface M together with a chosen equivalence class of differentiable atlases $\{(U_i, \phi_i)\}_i$ on M. An equivalent definition is that a differentiable surface is a topological surface M together with a choice of a maximal differentiable atlas.

Given two differentiable surfaces, M with the maximal atlas containing a differentiable atlas $\{(U_i, \phi_i)\}_i$, and N with the maximal atlas containing a differentiable atlas $\{(V_k, \psi_k)\}_k$, we can define what we mean by a differentiable map

$$f \colon M \longrightarrow N\,.$$

We say that f is differentiable at a point $p \in M$ if for any chart (U_i, ϕ_i) with $p \in U_i$ and for any chart (V_k, ψ_k) with $f(p) \in V_k$ the composite map

$$\psi_k \circ f \circ \phi_i^{-1} \colon \phi_i(U_i \cap f^{-1}(V_k)) \longrightarrow \mathbb{R}^2$$

is differentiable. The assumption that $\{(U_i, \phi_i)\}_i$ and $\{(V_k, \psi_k)\}_k$ are differentiable atlases ensures that if this holds for some pair of charts (U_i, ϕ_i) with $p \in U_i$ and (V_k, ψ_k) with $f(p) \in V_k$, then it holds for any such pair of charts.

A differentiable map $f \colon M \to N$ with differentiable inverse $g = f^{-1} \colon N \to M$ is called a diffeomorphism. This is the natural notion of isomorphism between differentiable surfaces.

20.4. **Smooth surfaces.** To ensure that the tangent plane T_pM at p of a differentiable surface varies continuously with p, or to speak about continuous fields of tangent vectors (= continuous vector fields), we will need to know what it means for a function $f \colon M \to \mathbb{R}$ to be continuously differentiable, not just differentiable. For each $1 \le r \le \infty$, a C^r atlas is defined to be an atlas $\{(U_i, \phi_i)\}_i$ such that each coordinate transformation

$$\phi_i \circ \phi_j^{-1} \colon \phi_j(U_{ij}) \longrightarrow \phi_i(U_{ij})$$

is r times continuously differentiable as a map between open subsets of $\mathbb{R}^2$. If $r = \infty$, this means that each coordinate transformation is infinitely often (continuously) differentiable. A C^∞ atlas is also called a smooth atlas. In that case we may also ask that each coordinate transformation is real analytic, meaning that the Taylor series expansion at each point converges in a neighborhood of that point. A real analytic atlas is also called a C^ω atlas. A C^r surface, for $r \in \{1, 2, \dots \infty, \omega\}$, is then a topological surface with a choice of a maximal C^r atlas.

Theorem 20.4 (Baer (1928), Epstein)**.** *Each topological surface can be given a C^1 structure. Any two C^1 structures on the same surface are C^1 diffeomorphic.*

Theorem 20.5 (Whitney (1936))**.** *Each C^1 surface can be given a C^r structure, for any $r \in \{1, 2, \dots, \infty, \omega\}$. Any two C^r structures on the same C^1 surface are C^r diffeomorphic.*

Remark 20.6. As in the case of triangulations, we will not prove these results. Allen Hatcher's proof (arXiv, 2013) establishes a slightly stronger theorem using only smooth techniques.

The corresponding existence and uniqueness statement holds for 3-dimensional manifolds (Moise, Bing(?)), but is false in most higher dimensions. Milnor (1956) showed that the 7-sphere S^7 admits smooth structures that are not diffeomorphic to the standard structure. Donaldson (1983), relying on work of Casson and Freedman, showed that $\mathbb{R}^4$ admits a smooth structure that is not diffeomorphic to the standard structure, and it was later found that there are uncountably many such non-diffeomorphic smooth structures on $\mathbb{R}^4$.

21. November 4th lecture

21.1. **Geometric structures on surfaces.** By the discussion above, it makes little difference whether we consider smooth, combinatorial or topological surfaces. However, none of these structures are geometric, in the sense that they specify distances between points in the surface, or a notion of congruence between line segments or between angles. To study such geometric structures we consider surfaces M that are locally modeled on a standard surface X together with a Lie group G of congruences of X.

Definition 21.1 (Thurston). A 2-*dimensional model geometry* (G, X) is

1. a connected and simply connected smooth surface X, and
2. a Lie group G acting smoothly and transitively on X, with compact stabilizers.

To avoid redundancy, we assume that G is a subgroup of the group of diffeomorphisms of X, that no larger group acts with compact stabilizers, and that there exists at least one closed surface M modeled on (G, X).

The condition that X is simply connected means that each closed curve in X can be continuously deformed (homotoped) to a point. Any connected surface admits a “universal covering space” with this property. A Lie group is a smooth manifold G with a group structure, such that the group multiplication $m\colon G \times G \to G$ taking (g, h) to gh, and the group inverse $i\colon G \to G$ taking g to g^{-1}, are both smooth maps. To say that G acts smoothly on X means that there is a smooth action map $a\colon G \times X \to X$ taking (g, x) to $g \cdot x$. This action is transitive if for each pair of points $x, y \in X$ there exists a $g \in G$ with $g \cdot x = y$.

The stabilizer $G_x = \{h \in G \mid h \cdot x = x\}$ of a point $x \in X$ is the subgroup of elements mapping that point to itself. It gets a topology as a subspace of G. If $g \cdot x = y$ then $gG_xg^{-1} = G_y$, so for a transitive action one stabilizer group is compact if and only if every stabilizer group is compact. This implies that X admits a Riemannian metric, hence also a metric $d\colon X \times X \to \mathbb{R}$, such that G acts by isometries. In other words, for each $g \in G$ the map $\gamma\colon X \to X$ given by $x \mapsto g \cdot x$ is an isometry, with $d(x, y) = d(\gamma(x), \gamma(y))$.

Definition 21.2. A surface M is *modeled* on (G, X) if it has an atlas $\{(U_i, \phi_i)\}_i$, consisting of open subsets U_i covering M and homeomorphisms $\phi_i\colon U_i \xrightarrow{\cong} V_i$ to open subsets V_i in X, such that the coordinate transformations are locally given by the action of G. In other words, for each point $x \in \phi_i(U_{ij})$, with $U_{ij} = U_i \cap U_j$, there is an element $g \in G$ such that

$$\phi_i \circ \phi_j^{-1} = \gamma$$

in a neighborhood of x, where $\gamma(y) = g \cdot y$ for all $y \in X$.

Theorem 21.3. *The only* 2*-dimensional model geometries are*

1. *the spherical geometry* $(O(3), S^2)$*, where* S^2 *has a Riemannian metric of constant curvature* $+1$*;*
2. *the Euclidean geometry* $(O(2) \ltimes \mathbb{R}^2, \mathbb{R}^2)$*, where* $\mathbb{R}^2$ *has a Riemannian metric of constant curvature* 0*; and*
3. *the hyperbolic geometry* $(\mathrm{Möb}(\mathbb{H}), \mathbb{H})$*, where* $\mathbb{H}$ *has a Riemannian metric of constant curvature* -1*.*

Example 21.4. A surface M modeled on spherical geometry comes with a Riemannian metric that is locally isometric to S^2. At any point $p \in M$ each coordinate patch $U_i \subseteq M$ containing p is identified with an open subset $V_i = \phi_i(U_i) \subseteq S^2$ in such a way that U_i inherits a well-defined metric from V_i (making ϕ_i an isometry). The locally shortest paths (geodesics) between points in U_i correspond to arcs of great circles on S^2. The only closed, connected examples of such surfaces are S^2 and P^2.

Example 21.5. A surface M modeled on Euclidean geometry comes with a Riemannian metric that is locally isometric to $\mathbb{R}^2$. At any point $p \in M$ each coordinate patch $U_i \subseteq M$ containing p is identified with an open subset $V_i = \phi_i(U_i) \subseteq \mathbb{R}^2$ in such a way that U_i inherits a well-defined metric from V_i (making ϕ_i an isometry). The locally shortest paths (geodesics) between points in U_i correspond to line segments in $\mathbb{R}^2$. The only closed, connected examples of such surfaces are T^2 and K^2, but each of these admits many non-isometric Euclidean structures.

For example, each Euclidean parallelogram $ABCD \subset \mathbb{C}$, with $A = 0$, $B = 1$, $C = 1 + \tau$ and $D = \tau$, with $\mathrm{Im}\,\tau > 0$, specifies a Euclidean structure on the torus T^2, obtained by identifying AB with DC, and BC with AD, by way of parallel translations.

Example 21.6. A surface M modeled on hyperbolic geometry comes with a Riemannian metric that is locally isometric to $\mathbb{H}$. At any point $p \in M$ each coordinate patch $U_i \subseteq M$ containing p is identified with an open subset $V_i = \phi_i(U_i) \subseteq \mathbb{H}$ in such a way that U_i inherits a well-defined metric from V_i (making ϕ_i an isometry). The locally shortest paths (geodesics) between points in U_i correspond to hyperbolic line segments in $\mathbb{H}$. The closed, connected examples of such surfaces are M_g and N_h for $g \geq 2$ and $h \geq 3$, and these admit many non-isometric hyperbolic structures.

For example, each hyperbolic octagon $ABCDEFGH$, with sides AB and DC, BC and ED, EF and GH, and FG and AH of pairwise equal length, and internal angle sum equal to 2π, specifies a hyperbolic structure on the surface $M_2 = T^2 \# T^2$ of genus two.

[[Discuss how for any connected surface M modeled on (G, X) the universal covering space $\widetilde{M}$ is diffeomorphic to X, so that $M \cong X/\Gamma$ where $\Gamma \subset G$ is a discrete subgroup isomorphic to the fundamental group $\pi_1(M)$ of M.]]

21.2. **Thurston's Geometrization Conjecture.** The situation in dimension three is also well understood.

Theorem 21.7 (Thurston)**.** *There are precisely eight* 3*-dimensional model geometries* (G, X)*, namely those modeled on* S^3 *(spherical),* $\mathbb{R}^3$ *(Euclidean),* $\mathbb{H}^3$ *(hyperbolic),* $S^2 \times \mathbb{R}$ *and* $\mathbb{H}^2 \times \mathbb{R}$*, nil geometry, the geometry of* $\widetilde{SL}(2, \mathbb{R})$ *and solv geometry.*

A 3-manifold is prime if it is not the connected sum of two manifolds different from S^3. Each 3-manifold can be written as the connected sum of prime 3-manifolds.

Theorem 21.8 (The Geometrization Theorem, Perelman)**.** *Every oriented prime closed* 3*-manifold* M *can be cut along finitely many disjoint, embedded tori* $T_1, \dots, T_n \subset M$*, such that each component of* $M \setminus (T_1 \cup \cdots \cup T_n)$ *has a geometric structure with finite volume.*

22. November 6th lecture

22.1. **Tangent planes and differentials of maps.** We can give an intrinsic construction of the tangent plane T_pM at a point p of a smooth surface M, and the derivative $df_p\colon T_pM \to T_qN$ at p of a smooth map $f\colon M \to N$, with $q = f(p)$.

Definition 22.1. Let M be a surface, with smooth structure given by a smooth atlas $\{(U_i, \phi_i)\}_i$. Let p be a point in M, and let $\Omega_p(M)$ be the set of smooth curves

$$\omega\colon J \to M$$

with J an open interval in $\mathbb{R}$ containing 0, and $\omega(0) = p$. Say that two curves ω_1 and ω_2 in $\Omega_p(M)$ are equivalent, denoted $\omega_1 \sim \omega_2$, if for any chart (U_i, ϕ_i) with $p \in U_i$ the relation

$$(\phi_i \circ \omega_1)'(0) = (\phi_i \circ \omega_2)'(0)$$

holds. Here $\phi_i \circ \omega_1$ and $\phi_i \circ \omega_2$ are smooth maps from neighborhoods of 0 in $\mathbb{R}$ to $\mathbb{R}^2$, and the relation asks that they have the same derivative at 0. (If the relation holds for one chart (U_i, ϕ_i) with $p \in U_i$ then it holds for any other chart (U_j, ϕ_j) in the smooth atlas with $p \in U_j$. This follows from the chain rule, since the coordinate transformations $\phi_j \circ \phi_i^{-1}$ are smooth.)

Let

$$T_pM = \Omega_p(M)/\sim$$

be the set of equivalence classes of smooth curves through p. The equivalence class of a curve ω is called the tangent vector of ω at p, and is denoted

$$\omega'(0) = [\omega] \in T_pM\,.$$

The set T_pM of tangent vectors is the *tangent plane* of M at p. The rule mapping $\omega'(0)$ to $(\phi_i \circ \omega)'(0)$ is a bijection $T_pM \to \mathbb{R}^2$, and determines a unique vector space structure on T_pM making this bijection a linear isomorphism. (A different choice of chart (U_j, ϕ_j) in the smooth atlas with $p \in U_j$ gives a different bijection, but the same vector space structure.)

Definition 22.2. Let $f\colon M \to N$ be a smooth map, where the surface N is equipped with the smooth atlas $\{(V_k, \psi_k)\}_k$. If $\omega\colon J \to M$ is a smooth curve in M through p, then $f \circ \omega\colon J \to N$ is a smooth curve in N through $q = f(p)$. The equivalence class $(f \circ \omega)'(0) = [f \circ \omega] \in T_qN$ only depends on the equivalence class $(\omega)'(0) = [\omega] \in T_pM$, hence the rule

$$df_p\colon \omega'(0) \mapsto (f \circ \omega)'(0)$$

defines a map $df_p\colon T_pM \to T_qN$, called the *differential* of f at p. It is a linear homomorphism, because for $q \in V_k$ the rule mapping $(\phi_i \circ \omega)'(0)$ to $(\psi_k \circ f \circ \omega)'(0)$ is the linear homomorphism $\mathbb{R}^2 \to \mathbb{R}^2$ given by the differential of the smooth map $\psi_k \circ f \circ \phi_i^{-1}$ at $\phi_i(p)$.

22.2. **Riemannian surfaces.** In order to specify the length of a smooth curve $\beta\colon [a, b] \to M$ in a smooth surface M, it suffices to specify the length of each of its tangent vectors, i.e., to give a norm on each tangent plane T_pM. We shall assume that this norm comes from an inner product on T_pM.

Definition 22.3. A *Riemannian metric* on a smooth surface M is a choice of inner product on each tangent plane T_pM, i.e., a bilinear, symmetric and positive definite paring

$$\langle -, - \rangle_p\colon T_pM \times T_pM \longrightarrow \mathbb{R}$$

for each $p \in M$. The inner product on T_pM is assumed to vary smoothly with p (in a sense that will be specified below). A *Riemannian surface* is a smooth surface equipped with a Riemannian metric.

We write $\| - \|_p$ for the associated norm on T_pM, given by $\|v\|_p^2 = \langle v, v \rangle_p$ for each $v \in T_pM$. The angle θ between two nonzero tangent vectors $v, w \in T_pM$ is determined by the relation

$$\cos\theta = \frac{\langle v, w \rangle_p}{\|v\|_p \|w\|_p}\,.$$

We often omit the subscript p when the point is clear from the context.

Definition 22.4. Let $\beta\colon [a,b] \to M$ be a smooth curve. For each $t \in [a,b]$ the *derivative* $\beta'(t) \in T_{\beta(t)}M$ is the tangent vector $\omega'(0)$ of the curve $\omega\colon J \to M$ defined by $\omega(u) = \beta(t+u)$ for u near 0. (We assume that β extends to an open interval containing $[a,b]$, and the definition of $\beta'(t)$ does not depend on the choice of extension.)

Definition 22.5. The *length* of a smooth curve $\beta\colon [a,b] \to M$ in a Riemannian surface M is defined to be

$$\operatorname{length}(\beta) = \int_a^b \|\beta'(u)\|_{\beta(u)}\, du\,.$$

More generally, for each $t \in [a,b]$ the arc length of $\beta|[a,t]$ equals

$$s(t) = \int_a^t \|\beta'(u)\|_{\beta(u)}\, du\,.$$

In particular,

$$s'(t) = \|\beta'(t)\|_{\beta(t)}\,.$$

Definition 22.6. A *regular curve* $\beta\colon [a,b] \to M$ is a smooth curve such that $\beta'(t) \neq 0$ for all $t \in [a,b]$. Let $\ell = \operatorname{length}(\beta)$. Then $s\colon t \mapsto s(t)$ is a diffeomorphism $[a,b] \to [0,\ell]$. Let $\alpha\colon [0,\ell] \to M$ be the smooth curve $\alpha = \beta \circ s^{-1}$, so that

$$\alpha(s(t)) = \beta(t)\,.$$

Then $\alpha'(s(t))s'(t) = \beta'(t)$, so $\|\alpha'(s)\|_{\alpha(s)} = 1$ for all $s \in [0,\ell]$. The curve α thus traverses the same image as β, but at *unit speed.* We call it the reparametrization of β by arc length.

23. November 11th lecture

23.1. **Regular surfaces in $\mathbb{R}^3$.**

Many (and possibly all) Riemannian surfaces can locally be realized as topological subspaces of $\mathbb{R}^3$, so that each tangent plane appears as a linear subspace of $\mathbb{R}^3$, and the Riemannian metric is obtained by restriction from the Euclidean dot product on $\mathbb{R}^3$. We now concentrate on such concrete realizations of abstract surfaces as subspaces of $\mathbb{R}^3$.

Definition 23.1. Let S be a topological subspace of $\mathbb{R}^3$, and suppose that S is a topological surface, with a chosen atlas $\{(U_i, \phi_i)\}_i$. The inverse of each homeomorphism

$$\phi_i\colon U_i \to \phi_i(U_i) = V_i \subseteq \mathbb{R}^2$$

can be viewed as an embedding

$$x_i = \phi_i^{-1}\colon V_i \to x_i(V_i) = U_i \subseteq S \subset \mathbb{R}^3\,.$$

We call each map $x_i\colon V_i \to S$ a *local parametrization* of S. The images $x_i(V_i)$ are open subsets of S, and their union covers S.

Definition 23.2. A *regular surface* in $\mathbb{R}^3$ is a subspace $S \subset \mathbb{R}^3$ together with a collection $\{(x_i, V_i)\}_i$ of local parametrizations $x_i\colon V_i \to S$, where each V_i is an open subset of $\mathbb{R}^2$, each composite $x_i\colon V_i \to S \subset \mathbb{R}^3$ is a smooth map whose Jacobian has rank 2 at each point, each x_i corestricts to a homeomorphism $x_i\colon V_i \to x_i(V_i)$, and the images $\{x_i(V_i)\}_i$ form an open cover of S. (As for smooth surfaces, one should really require that the collection of local parametrizations is maximal.)

If we write $x_i\colon V_i \to \mathbb{R}^3$ as $x_i(u,v) = (x(u,v), y(u,v), z(u,v))$, the transposed Jacobian of x_i is the matrix

$$\begin{bmatrix} \partial x/\partial u & \partial y/\partial u & \partial z/\partial u \\ \partial x/\partial v & \partial y/\partial v & \partial z/\partial v \end{bmatrix}$$

of partial derivatives. The differential of x_i is the linear map $\mathbb{R}^2 \to \mathbb{R}^3$ given by multiplication by the Jacobian matrix. It has rank 2 if and only if the row vectors

$$(x_i)_u = (\partial x/\partial u, \partial y/\partial u, \partial z/\partial u) \qquad \text{and} \qquad (x_i)_v = (\partial x/\partial v, \partial y/\partial v, \partial z/\partial v)$$

are linearly independent.

Example 23.3. The *graph*

$$S = \{(u, v, h(u, v)) \mid (u, v) \in V\}$$

of a smooth function $h\colon V \to \mathbb{R}$ is a regular surface in $\mathbb{R}^3$. It has a local parametrization

$$x(u, v) = (u, v, h(u, v))$$

with transposed Jacobian matrix

$$\begin{bmatrix} 1 & 0 & h_u \\ 0 & 1 & h_v \end{bmatrix} .$$

where $h_u = \partial h/\partial u$ and $h_v = \partial h/\partial v$.

Example 23.4. Let $S^2 = \{(x, y, z) \mid x^2 + y^2 + z^2 = 1\}$ be the unit sphere in $\mathbb{R}^3$. It is a regular surface, with local parametrizations of the form

$$x_1(u, v) = (u, v, \sqrt{1 - u^2 - v^2}) \qquad x_2(u, v) = (u, v, -\sqrt{1 - u^2 - v^2})$$

for (u, v) in the open unit disc $V \subset \mathbb{R}^2$, together with four more variants obtained by cyclically permuting the coordinates.

Example 23.5. A helicoid is the regular surface S in $\mathbb{R}^3$ with parametrization

$$x(u, v) = (u \cos v, u \sin v, v)$$

for $u \in (0, \infty)$ and $v \in \mathbb{R}$. The vectors

$$x_u(u, v) = (\cos v, \sin v, 0) \qquad \text{and} \qquad x_v(u, v) = (-u \sin v, u \cos v, 1)$$

are linearly independent.

Lemma 23.6. *A regular surface in $\mathbb{R}^3$ is a smooth surface.*

Proof. Given a collection of local parametrizations $\{(x_i, V_i)\}_i$ we define an atlas $\{(U_i, \phi_i)\}_i$ by setting $U_i = x_i(V_i)$ and $\phi_i = x_i^{-1}$. It remains to verify that the coordinate transformations

$$\phi_i \circ \phi_j^{-1} = x_i^{-1} \circ x_j$$

from $x_j^{-1}(U_{ij}) \subseteq V_j \subseteq \mathbb{R}^2$ to $x_i^{-1}(U_{ij}) \subseteq V_i \subseteq \mathbb{R}^2$ are smooth maps.

To do this, we use the inverse function theorem to locally extend $x_i^{-1}\colon U_i \to V_i$ to a smooth map $\pi \circ X_i^{-1}$ defined in an open neighborhood in $\mathbb{R}^3$, so that $x_i^{-1} \circ x_j = \pi \circ X_i^{-1} \circ x_j$ is a composite of smooth maps. Let $p = x_i(u, v) \in U_{ij}$, with $(u, v) \in V_i \subset \mathbb{R}^2$, and choose a vector $N \in \mathbb{R}^3$ so that $(x_i)_u(u, v)$, $(x_i)_v(u, v)$ and N are linearly independent. Define $X_i\colon V_i \times \mathbb{R} \to \mathbb{R}^3$ by

$$X_i(u, v, w) = x_i(u, v) + wN .$$

The transposed Jacobian of X_i at $(u, v, 0)$ has rows $(x_i)_u(u, v)$, $(x_i)_v(u, v)$ and N, hence is invertible, so X_i restricts to a diffeomorphism from a neighborhood of $(u, v, 0)$ in $V_i \times \mathbb{R}$ to a neighborhood of p in $\mathbb{R}^3$. Its inverse X_i^{-1} is a diffeomorphism from a neighborhood of p in $\mathbb{R}^3$ to a neighborhood of $(u, v, 0)$ in $V_i \times \mathbb{R}$. Let $\pi\colon V_i \times \mathbb{R} \to V_i$ be the linear projection on the first coordinate(s). Then $\pi \circ X_i^{-1}$ is a smooth map to V_i defined near p in $\mathbb{R}^3$, and it agrees with x_i^{-1} on $x_i(V_i) \subset S$. □

Lemma 23.7. *At each point $p \in S$ the (abstract) tangent plane T_pS is naturally identified with the (concrete) linear subspace of $\mathbb{R}^3$ consisting of tangent vectors at p of curves in S viewed as curves in $\mathbb{R}^3$.*

Proof. The inclusion $\iota\colon S \to \mathbb{R}^3$ takes each smooth curve $\omega\colon J \to S$ with $\omega(0) = p$ to a smooth curve $\iota \circ \omega\colon J \to \mathbb{R}^3$. The derivative $(\iota \circ \omega)'(0) \in \mathbb{R}^3$ only depends on the equivalence class $\omega'(0) = [\omega]$ of ω in T_pS, and the rule $d\iota_p\colon \omega'(0) \mapsto (\iota \circ \omega)'(0)$ defines the stated identification. □

Lemma 23.8. *A regular surface S inherits a Riemannian metric from the dot product in $\mathbb{R}^3$.*

Proof. For each $p \in S$ the inner product

$$\langle -, - \rangle_p \colon T_pS \times T_pS \to \mathbb{R}$$

is defined by means of the Euclidean dot product:

$$\langle v, w \rangle_p = v \cdot w = v_1 w_1 + v_2 w_2 + v_3 w_3$$

where $v, w \in T_pS$ on the left hand side, and $v = (v_1, v_2, v_3), w = (w_1, w_2, w_3) \in \mathbb{R}^3$ in the middle and on the right hand side. □

23.2. **The first fundamental form.** A local parametrization $x \colon V \to S$ of a regular surface in $\mathbb{R}^3$ gives rise to a preferred basis (x_u, x_v) for $T_pS \subset \mathbb{R}^3$ at each point $p \in x(V)$. The inherited Riemannian metric is determined by its values on these basis vectors.

Definition 23.9. Let $x \colon V \to S$ be a local parametrization. The curves

$$t \mapsto \alpha_v(t) = x(t, v) \qquad \text{and} \qquad t \mapsto \beta_u(t) = x(u, t)$$

for fixed $v \in \mathbb{R}$ and $u \in \mathbb{R}$, respectively, are called *coordinate curves* in S. Let

$$x_u(u, v) = \alpha'_v(u) \qquad \text{and} \qquad x_v(u, v) = \beta'_u(v)$$

be the tangent vectors of these curves, at $t = u$ and $t = v$, respectively. These are tangent vectors in T_pS, for $p = x(u, v)$. Viewed as vectors in $\mathbb{R}^3$, these are the partial derivatives of x at (u, v) with respect to u and v, respectively. Hence $x_u(u, v)$ and $x_v(u, v)$ are the rows vectors of the transposed Jacobian of x at (u, v). By the regularity assumption they are linearly independent, and therefore form a basis for the tangent plane $T_pS \subset \mathbb{R}^3$.

Definition 23.10. The inner product

$$\langle -, - \rangle_p \colon T_pS \times T_pS \longrightarrow \mathbb{R}$$

is determined by its values on pairs of vectors taken from the basis $\{x_u(u, v), x_v(u, v)\}$ of T_pS, where $p = x(u, v)$. Let

$$E = x_u \cdot x_u \quad , \quad F = x_u \cdot x_v = x_v \cdot x_u \quad \text{and} \quad G = x_v \cdot x_v$$

as smooth functions $V \to \mathbb{R}$. The inner product of $ax_u + bx_v$ and $cx_u + dx_v$ (at $p \in x(V)$) then equals

$$(ax_u + bx_v) \cdot (cx_u + dx_v) = Eac + F(ad + bc) + Gbd$$

(evaluated at $x^{-1}(p) \in V$). This expression, in terms of the coordinates (a, b) and (c, d) of the two vectors with respect to the ordered basis $(x_u(u, v), x_v(u, v))$, equals the matrix product

$$\begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} E & F \\ F & G \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} .$$

The bilinear form represented in these coordinates by the symmetric, positive definite matrix

$$I = \begin{bmatrix} E & F \\ F & G \end{bmatrix}$$

is called the *first fundamental form* of the parametrization $x \colon V \to S$. Note that $EG - F^2 > 0$ by the Cauchy–Schwarz inequality, since x_u and x_v are linearly independent.

If $\beta \colon [a, b] \to S$ is a smooth curve, and $\beta(t) = p \in x(V)$, then $\beta = x(u, v)$ for smooth functions u and v defined near t. By the chain rule, $\beta'(t) = x_u u'(t) + x_v v'(t)$, so

$$s'(t)^2 = \|\beta'(t)\|_p^2 = Eu'(t)^2 + 2Fu'(t)v'(t) + Gv'(t)^2 .$$

We can write this as

$$(\frac{ds}{dt})^2 = E(\frac{du}{dt})^2 + 2F\frac{du}{dt}\frac{dv}{dt} + G(\frac{dv}{dt})^2$$

or as

$$ds^2 = E\,du^2 + 2F\,dudv + G\,dv^2 ,$$

in terms of symmetric 2-forms. This expression is also often called the first fundamental form of the parametrization $x \colon V \to S$.

Example 23.11. The tangent plane T_pS at $p = x(u, v)$ of the graph $S \subset \mathbb{R}^3$ of a smooth map $h\colon V \to \mathbb{R}$ is spanned by the vectors

$$x_u(u, v) = (1, 0, h_u(u, v)) \qquad \text{and} \quad x_v(u, v) = (1, 0, h_v(u, v)) .$$

The first fundamental form is given by

$$E = 1 + h_u^2 \quad , \quad F = h_u h_v \quad \text{and} \quad G = 1 + h_v^2 .$$

Example 23.12. The first fundamental form of $S = \mathbb{R}^2 \subset \mathbb{R}^3$, viewed as the graph of the zero function $\mathbb{R}^2 \to \mathbb{R}$, is

$$ds^2 = du^2 + dv^2$$

with $E = G = 1$ and $F = 0$.

Example 23.13. ((Do the case $S = S^2$?))

Remark 23.14. The condition that the inner products $\langle -, - \rangle_p$ of a Riemannian metric vary smoothly with p can be made precise as follows. For each local parametrization (x, V) the inner products

$$E = \langle x_u, x_u \rangle \quad , \quad F = \langle x_u, x_v \rangle = \langle x_v, x_u \rangle \quad \text{and} \quad G = \langle x_v, x_v \rangle$$

are required to be smooth as functions $V \to \mathbb{R}$. Then, if $X : p \to X_p \in T_pS$ and $Y : p \mapsto Y_p \in T_pS$ are smooth vector fields on S, then $\langle X, Y \rangle$ is a smooth function on S. This condition is clearly satisfied for our regular surfaces.

Example 23.15. The upper half-plane model $S = \mathbb{H}$ for the hyperbolic plane cannot be fully realized as a regular surface in $\mathbb{R}^3$, but it admits a Riemannian metric $\langle -, - \rangle_p$ for $p = (u, v) \in \mathbb{H}$ such that

$$s'(t)^2 = \|\beta'(t)\|_p^2 = \frac{u'(t)^2 + v'(t)^2}{v(t)^2}$$

for any smooth curve $\beta = (u, v)$ with $v > 0$. Once we know that $\mathbb{H}$ can be locally realized as a regular surface, its first fundamental form will be

$$ds^2 = \frac{du^2 + dv^2}{v^2}$$

with $E(u, v) = G(u, v) = 1/v^2$ and $F(u, v) = 0$.

Example 23.16. The unit disc model $S = \mathbb{D}$ for the hyperbolic plane also admits a Riemannian metric $\langle -, - \rangle_p$ for $p = (u, v) \in \mathbb{D}$ such that

$$s'(t)^2 = \|\beta'(t)\|_p^2 = \frac{4(u'(t)^2 + v'(t)^2)}{(1 - u(t)^2 - v(t)^2)^2}$$

for any smooth curve $\beta = (u, v)$ with $u^2 + v^2 < 1$. When realized locally as a regular surface, its first fundamental form will be

$$ds^2 = \frac{4(du^2 + dv^2)}{(1 - u^2 - v^2)^2}$$

with $E(u, v) = G(u, v) = 4/(1 - u^2 - v^2)^2$ and $F(u, v) = 0$.

23.3. **Intrinsic and extrinsic properties.**

Definition 23.17. A map $f\colon M \to N$ of Riemannian surfaces is an *isometry* if it is a diffeomorphism and

$$\langle v, w \rangle_p = \langle df_p(v), df_p(w) \rangle_q$$

for each $p \in M$ and $v, w \in T_pM$, with $q = f(p)$.

It is equivalent to require that $\|v\|_p = \|df_p(v)\|_q$ for each $p \in M$ and $v \in T_pM$, with $q = f(p)$. The self-isometries $\gamma\colon M \to M$ of a Riemannian surface form a group, called the isometry group of M.

Example 23.18. The isometries $\gamma\colon \mathbb{R}^2 \to \mathbb{R}^2$ are the Euclidean motions $E(2) = O(2) \ltimes \mathbb{R}^2$ of $\mathbb{R}^3$, of the form $\gamma(v) = Av + b$ with $A \in O(2)$ and $b \in \mathbb{R}^2$.

Example 23.19. The isometries $\gamma\colon S^2 \to S^2$ are the restrictions of the Euclidean motions of $\mathbb{R}^3$ that preserve S^2, i.e., the rotations and reflections of the form $\gamma(v) = Av$ with $A \in O(3)$.

Example 23.20. The isometries $\gamma\colon \mathbb{H} \to \mathbb{H}$ are the Möbius transformations preserving $\mathbb{H}$, i.e., the group Möb($\mathbb{H}$). Similarly, the isometries $\gamma\colon \mathbb{D} \to \mathbb{D}$ are the Möbius transformations preserving $\mathbb{D}$, i.e., the group Möb($\mathbb{D}$).

Definition 23.21. Aspects of the geometry of a Riemannian surface that are preserved under all isometries are said to be *intrinsic*, whereas those that depend on a particular presentation of the surface, e.g. as a regular surface in $\mathbb{R}^3$, are called *extrinsic.*

Example 23.22. Arc length is an intrinsic property of smooth curves in a Riemannian surface. If $\beta\colon [a, b] \to M$ is such a curve, and $f\colon M \to N$ is an isometry, then

$$\text{length}_M(\beta) = \text{length}_N(f \circ \beta)\,.$$

If β is parametrized by arc length, then so is $f \circ \beta$.

Lemma 23.23. *Let S and S' be regular surfaces in $\mathbb{R}^3$ (or Riemannian surfaces), and suppose that there is an isometry $f\colon U \to U'$ from an open subset $U \subseteq S$ to an open subset $U' \subseteq S'$. If $x\colon V \to S$ is a local parametrization of S, with $x(V) \subseteq U$, then $x' = f \circ x\colon V \to S'$ is a local parametrization of S' with $x'(V) \subset U'$. The first fundamental form $Edu^2 + 2Fdudv + Gdv^2$ for S is then equal to the first fundamental form $E'du^2 + 2F'dudv + G'dv^2$ for S'.*

Proof. $E = x_u \cdot x_u = df(x_u) \cdot df(x_u) = x'_u \cdot x'_u = E'$, and likewise $F = F'$ and $G = G'$. □

Lemma 23.24. *Let $x\colon V \to S$ and $x'\colon V \to S'$ be local parametrizations of two regular surfaces in $\mathbb{R}^3$ (or Riemannian surfaces), such that $E = E'$, $F = F'$ and $G = G'$ are equal as functions on V. Then $f = x' \circ x^{-1}$ is an isometry from $x(V)$ to $x'(V)$.*

Proof. We have $df(x_u) = x'_u$ and $df(x_v) = x'_v$, so

$$x_u \cdot x_u = E = E' = x'_u \cdot x'_u = df_p(x_u) \cdot df_p(x_u)$$

and similarly we get $x_u \cdot x_v = df_p(x_u) \cdot df_p(x_v)$ and $x_v \cdot x_v = df_p(x_v) \cdot df_p(x_v)$. It follows by bilinearity that $df_p\colon T_pS \to T_{f(p)}S'$ preserves the inner product inherited from the dot product, for each $p \in x(V)$. □

Example 23.25. Let $\alpha\colon [0, \ell] \to \mathbb{R}^2$ be an embedded curve, parametrized by arc length, and let $S \subset \mathbb{R}^3$ be parametrized by

$$x(u, v) = \alpha(u) + ve_3$$

where $e_3 = (0, 0, 1)$, for $(u, v) \in V = [0, \ell] \times \mathbb{R}$. Let $S' \subset \mathbb{R}^3$ be parametrized by

$$x'(u, v) = (u, 0, v)\,.$$

Then the diffeomorphism $f\colon S \to S'$ given by $f(x(u, v)) = x'(u, v)$ is a (Riemannian) isometry. To see this, note that $x_u = \alpha'$ and $x_v = e_3$ give an orthonormal basis for T_pS at each point $p = x(u, v)$, and $df(x_u) = x'_u = e_1 = (1, 0, 0)$ and $df(x_v) = x'_v = e_3$ give an orthonormal basis for T_qS' at $q = x'(u, v)$. Hence $df_p\colon T_pS \to T_qS'$ preserves the inner product.

24. November 13th lecture

24.1. **Orientations and normal vectors.** For smooth surfaces we can specify orientations in terms of oriented atlases. For regular surfaces in $\mathbb{R}^3$ they can be specified in terms of choices of normal vectors.

Definition 24.1. An *orientation* of a 2-dimensional vector space P is a choice of an equivalence class of ordered bases (b_1, b_2) for P, where two ordered bases are equivalent if the change-of-basis matrix relating them has positive determinant. An ordered basis in the chosen equivalence class is then called *positively oriented.* There is precisely one other equivalence class of ordered bases, and these are called *negatively oriented.*

If (b_1, b_2) is positively oriented, then (b_2, b_1) is negatively oriented, since the change-of-basis matrix

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

has negative determinant. A choice of orientation of the plane P determines a preferred direction of travel for (simple) loops around its origin, namely that of the loop from b_1 to b_2 to $-b_1$ to $-b_2$ and back to b_1, for any positively oriented basis (b_1, b_2). The negatively oriented basis (b_2, b_1) then specifies the loop from b_2 to b_1 to $-b_2$ to $-b_1$ and back to b_2, which corresponds to the opposite direction of travel.

Definition 24.2. For each 2-dimensional subspace $P \subset \mathbb{R}^3$ there are precisely two vectors $\pm N$ of length 1 that are orthogonal to P. We call these the two *unit normal vectors* to P.

A choice N of one of the two unit normal vectors determines an orientation of the plane P, by the rule that an ordered basis (b_1, b_2) is positively oriented if (b_1, b_2, N) forms a right-handed system, i.e., if the matrix with rows b_1, b_2 and N has positive determinant. This is equivalent to asking that

$$N = \frac{b_1 \times b_2}{\|b_1 \times b_2\|}$$

is the unit vector in the direction of the cross product $b_1 \times b_2$.

Lemma 24.3. *A regular surface S in $\mathbb{R}^3$ is orientable if and only if it admits a smooth unit normal vector field, i.e., if and only if we can choose a unit normal vector $N_p \in S^2 \subset \mathbb{R}^3$ to each tangent plane $T_pS \subset \mathbb{R}^3$ so that the Gauss map $N \colon S \to S^2$ sending p to N_p is smooth.*

Proof. It suffices to assume that the unit normal field N is continuous. Then the condition on an ordered basis (b_1, b_2) for T_pS that (b_1, b_2, N) is right-handed suffices to determine an orientation of T_pS, and this varies continuously with $p \in S$. Conversely, a continuous choice of orientations determines a continuous choice of unit normal vectors $p \mapsto N_p$. To see that this continuous choice is in fact smooth, it suffices to consider $p \in x(V)$ for a local parametrization $x \colon V \to S$. We may assume that V is connected. Then $(x_u(u, v), x_v(u, v))$ forms an ordered basis for $T_{x(u,v)}S \subset \mathbb{R}^3$, for all $(u, v) \in V$. It is either positively oriented for all $(u, v) \in V$, or negatively oriented for all $(u, v) \in V$, since V is connected. In the latter case, replace x by the local parametrization y given by $y(u, v) = x(v, u)$, and rename y as x. It then follows that

$$N = \frac{x_u \times x_v}{\|x_u \times x_v\|}$$

at all points p of $x(V)$, and the right hand side varies smoothly with p. Hence the continuous vector field $p \mapsto N_p$ is in fact smooth. □

24.2. **Curvature.** The curvature of a regular surface S in $\mathbb{R}^3$ is a measure of how fast the tangent planes $T_pS \subset \mathbb{R}^3$ moves as p varies on the surface. Equivalently, it measures how much a unit normal vector $N_p \in S^2$ moves as p varies.

The determinant of a linear map $f \colon P \to P$ is the determinant of the matrix representing f with respect to any choice of ordered basis (b_1, b_2) for both copies of P. As long as we use the same ordered basis in the source (domain) and target (codomain), the value of the determinant does not change. Similar remarks apply for the trace of f.

Definition 24.4. The derivative of the Gauss map $N \colon S \to S^2$ at $p \in S^2$ maps the tangent plane T_pS of S at p to the tangent plane $T_{N_p}S^2$ of S^2 at N_p. Since N_p is a unit normal to T_pS, the latter tangent plane is also equal to T_pS. Hence

$$dN_p \colon T_pS \longrightarrow T_{N_p}S^2 = T_pS$$

is a linear self-map of the plane T_pS. The *Gaussian curvature* of S at p is defined to be the determinant of this linear self-map:

$$K(p) = \det(dN_p)\,.$$

This defines a smooth map $K\colon S \to \mathbb{R}$.

This definition does not depend on a choice $x\colon V \to S$ of local parametrization of S near p. It appears to depend on a choice of unit normal vector field N near p, or equivalently an orientation of S near p, but the opposite choice of unit normal vector field (namely, $p \mapsto -N_p$) has differential $d(-N)p = -dN_p$, and $\det(-dN_p) = \det(dN_p)$ since T_pS is 2-dimensional.

Example 24.5. The Euclidean plane $\mathbb{R}^2 \subset \mathbb{R}^3$ has constant unit normal vector $N_p = (0, 0, 1)$ at all $p \in \mathbb{R}^2$, so $dN_p = 0$ is the zero map from $T_p\mathbb{R}^2 = \mathbb{R}^2$ to itself. Its determinant is 0, so $K(p) = 0$ for all $p \in \mathbb{R}^2$. The Euclidean plane has zero curvature everywhere.

Example 24.6. The sphere S of radius $R > 0$, centered at the origin, has unit normal vector $N_p = p/R$ at $p \in S$, so $N\colon S \to S^2$ is the restriction of the linear map $p \mapsto p/R$. Its derivative $dN_p\colon T_pS \to T_{p/R}S^2 = T_pS$ is also the linear map $p \mapsto p/R$, with determinant $1/R^2$. Hence $K(p) = 1/R^2$ for all $p \in S$. The sphere of radius R has constant curvature $1/R^2$. In particular, the unit sphere S^2 has curvature $+1$.

25. November 18th lecture

25.1. The second fundamental form.

Definition 25.1. Let $N\colon S \to S^2$ be a smooth unit normal vector field, and let $x\colon V \to S$ be a local parametrization. We write

$$N_u = (N \circ x)_u = \partial(N \circ x)/\partial u$$
$$N_v = (N \circ x)_v = \partial(N \circ x)/\partial v$$

for the partial derivatives of $(u, v) \mapsto N_p$, with $p = x(u, v)$, viewed as a smooth map $\iota \circ N \circ x\colon V \to S^2 \subset \mathbb{R}^3$. By definition of the differential of N at p, $dN_p(x_u(u, v)) = N_u(u, v)$ and $dN_p(x_v(u, v)) = N_v(u, v)$. More briefly,

$$N_u = dN(x_u) \qquad \text{and} \qquad N_v = dN(x_v)\,.$$

Furthermore, we write

$$x_{uu} \quad , \quad x_{uv} = x_{vu} \quad \text{and} \quad x_{vv}$$

for the second order partial derivatives of x viewed as a smooth map $\iota \circ x\colon V \to S \subset \mathbb{R}^3$. The relation $x_{uv} = x_{vu}$ holds because x is smooth. (It suffices that x is two times continuously differentiable.)

Lemma 25.2. *$N_u \cdot x_u = -N \cdot x_{uu}$, $N_u \cdot x_v = -N \cdot x_{uv} = N_v \cdot x_u$ and $N_v \cdot x_v = -N \cdot x_{vu}$.*

Proof. Since N_p is normal to the plane T_pS with basis $x_u(u, v)$ and $x_v(u, v)$, for $p = x(u, v)$, we have $N_p \cdot x_u(u, v) = 0$ and $N_p \cdot x_v(u, v) = 0$ for all $(u, v) \in V$. Differentiating each of these with respect to u, and with respect to v, gives the stated relations. □

Definition 25.3. Define functions e, f and $g\colon V \to \mathbb{R}$ by

$$e = N \cdot x_{uu}$$
$$f = N \cdot x_{uv} = N \cdot x_{vu}$$
$$g = N \cdot x_{vv}\,.$$

The *second fundamental form* of the parametrization $x\colon V \to S$ is the symmetric bilinear form given by the symmetric matrix

$$II = \begin{bmatrix} e & f \\ f & g \end{bmatrix}\,.$$

Example 25.4. The graph $S \subset \mathbb{R}^3$ of $h\colon V \to \mathbb{R}$ has normal vectors

$$x_u \times x_v = (-h_u, -h_v, 1)$$

and unit normal vectors

$$N = \frac{(-h_u, -h_v, 1)}{\sqrt{1 + h_u^2 + h_v^2}} .$$

The second order partial derivatives are

$$\begin{aligned} x_{uu} &= (0, 0, h_{uu}) \\ x_{uv} &= (0, 0, h_{uv}) \\ x_{vv} &= (0, 0, h_{vv}) \end{aligned}$$

and the second fundamental form has components

$$\begin{aligned} e &= \frac{h_{uu}}{\sqrt{1 + h_u^2 + h_v^2}} \\ f &= \frac{h_{uv}}{\sqrt{1 + h_u^2 + h_v^2}} \\ g &= \frac{h_{vv}}{\sqrt{1 + h_u^2 + h_v^2}} . \end{aligned}$$

At a critical point of h, where $h_u(u, v) = h_v(u, v) = 0$, we get $N_p = (0, 0, 1)$ and $e = h_{uu}(u, v)$, $f = h_{uv}(u, v)$ and $g = h_{vv}(u, v)$. Hence the second order Taylor series of h at (u, v) is

$$T_2 h(u + \Delta u, v + \Delta v) = h(u, v) + \frac{1}{2}(e\Delta u^2 + 2f\Delta u\Delta v + g\Delta v^2) .$$

The symmetric 2-form

$$e\, du^2 + 2f\, dudv + g\, dv^2$$

is also often called the second fundamental form of the parametrization $x \colon V \to S$.

The curvature of a regular surface S at p can be conveniently expressed in terms of the first and second fundamental form of a parametrization of S near p.

Proposition 25.5.

$$K = \frac{eg - f^2}{EG - F^2} .$$

Proof. The curvature $K(p)$ equals the determinant of the matrix representing $dN_p \colon T_pS \to T_pS$ for any choice of ordered basis for T_pS. We use the basis $(x_u(u, v), x_v(u, v))$, with $p = x(u, v)$. The representing matrix

$$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} ,$$

with determinant $\alpha\delta - \beta\gamma$, is then determined by the relations

$$N_u = dN(x_u) = \alpha x_u + \beta x_v \qquad \text{and} \qquad N_v = dN(x_v) = \gamma x_u + \delta x_v .$$

Taking dot products with x_u, and with x_v, we get the relations

$$\begin{aligned} -e &= -N \cdot x_{uu} = N_u \cdot x_u = \alpha x_u \cdot x_u + \beta x_v \cdot x_u = \alpha E + \beta F \\ -f &= -N \cdot x_{uv} = N_u \cdot x_v = \alpha x_u \cdot x_v + \beta x_v \cdot x_v = \alpha F + \beta G \\ -f &= -N \cdot x_{vu} = N_v \cdot x_u = \gamma x_u \cdot x_u + \delta x_v \cdot x_u = \gamma E + \delta F \\ -g &= -N \cdot x_{vv} = N_v \cdot x_v = \gamma x_u \cdot x_v + \delta x_v \cdot x_v = \gamma F + \delta G . \end{aligned}$$

Hence

$$-\begin{bmatrix} e & f \\ f & g \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \begin{bmatrix} E & F \\ F & G \end{bmatrix}$$

and

$$eg - f^2 = (\alpha\delta - \beta\gamma)(EG - F^2) = K(EG - F^2) .$$

Dividing by $EG - F^2 > 0$ concludes the proof. □

Example 25.6. The graph of $h\colon V \to \mathbb{R}$ has

$$eg - f^2 = \frac{h_{uu}h_{vv} - h_{uv}^2}{1 + h_u^2 + h_v^2}$$

and

$$EG - F^2 = (1 + h_u^2)(1 + h_v^2) - (h_u h_v)^2 = 1 + h_u^2 + h_v^2$$

so

$$K = \frac{h_{uu}h_{vv} - h_{uv}^2}{(1 + h_u^2 + h_v^2)^2}\,.$$

Here the numerator

$$h_{uu}h_{vv} - h_{uv}^2 = \det H(h)$$

is the determinant of the Hessian

$$H(h) = \begin{bmatrix} h_{uu} & h_{uv} \\ h_{vu} & h_{vv} \end{bmatrix}$$

of h. At a critical point, where $h_u = h_v = 0$, the formula simplifies and the curvature is equal to the determinant of the Hessian matrix.

25.2. **Theorema egregium.** Gauss called the following result a remarkable theorem, "Theorema egregium" in Latin. It says that the curvature at a point of a regular surface only depends on the Riemannian metric on the surface in a neighborhood of that point, or equivalently, that it only depends on the components E, F and G of the first fundamental form, and their derivatives, at that point.

Theorem 25.7 (Gauss)**.** *Curvature is intrinsic.*

Proof. The result can be stated and proven for general Riemannian surfaces, but we only prove it for regular surfaces in $\mathbb{R}^3$. As discussed above, if $f\colon U \to U'$ is an isometry between open neighborhoods $U \subseteq S$ and $U' \subseteq S'$ of $p \in U$ and $q = f(p) \in U'$, then we can find local parametrizations $x\colon V \to U \subset S$ and $x' = f \circ x\colon V \to U' \subset S'$ such that the first fundamental forms of x and x' are equal. Hence it suffices to prove that $K(p)$ can be expressed in terms of E, F and G in a neighborhood of $(u, v) \in V$ with $x(u, v) = p$. We will prove that $K(p)$ can be expressed in terms of E, F, G and their partial derivatives, up to second order, at p.

Note that

$$\|x_u \times x_v\|^2 = \|x_u\|^2\|x_v\|^2 - (x_u \cdot x_v)^2 = EG - F^2\,,$$

so $\|x_u \times x_v\| = \sqrt{EG - F^2}$. From

$$K(EG - F^2) = eg - f^2 = (N \cdot x_{uu})(N \cdot x_{vv}) - (N \cdot x_{uv})^2\,,$$

where $N = (x_u \times x_v)/\sqrt{EG - F^2}$, we obtain

$$K(EG - F^2)^2 = ((x_u \times x_v) \cdot x_{uu})((x_u \times x_v) \cdot x_{vv}) - ((x_u \times x_v) \cdot x_{uv})^2\,.$$

Using

$$(a \times b) \cdot c = \det \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

we can rewrite the right hand side as

$$\det \begin{bmatrix} x_u \\ x_v \\ x_{uu} \end{bmatrix} \det \begin{bmatrix} x_u \\ x_v \\ x_{vv} \end{bmatrix} - \det \begin{bmatrix} x_u \\ x_v \\ x_{uv} \end{bmatrix}^2 = \det \begin{bmatrix} x_u \\ x_v \\ x_{uu} \end{bmatrix} \begin{bmatrix} x_u \\ x_v \\ x_{vv} \end{bmatrix}^t - \det \begin{bmatrix} x_u \\ x_v \\ x_{uv} \end{bmatrix} \begin{bmatrix} x_u \\ x_v \\ x_{uv} \end{bmatrix}^t \tag{25.1}$$

$$= \det \begin{bmatrix} E & F & x_u \cdot x_{vv} \\ F & G & x_v \cdot x_{vv} \\ x_{uu} \cdot x_u & x_{uu} \cdot x_v & x_{uu} \cdot x_{vv} \end{bmatrix} - \det \begin{bmatrix} E & F & x_u \cdot x_{uv} \\ F & G & x_v \cdot x_{uv} \\ x_{uv} \cdot x_u & x_{uv} \cdot x_v & x_{uv} \cdot x_{uv} \end{bmatrix}\,.$$

Using the first lemma below, we can rewrite this difference as

$$(25.2)\quad \det\begin{bmatrix} E & F & F_v - \frac{1}{2}G_u \\ F & G & \frac{1}{2}G_v \\ \frac{1}{2}E_u & F_u - \frac{1}{2}E_v & x_{uu}\cdot x_{vv}\end{bmatrix} - \det\begin{bmatrix} E & F & \frac{1}{2}E_v \\ F & G & \frac{1}{2}G_u \\ \frac{1}{2}E_v & \frac{1}{2}G_u & x_{uv}\cdot x_{uv}\end{bmatrix}$$
$$= \det\begin{bmatrix} E & F & F_v - \frac{1}{2}G_u \\ F & G & \frac{1}{2}G_v \\ \frac{1}{2}E_u & F_u - \frac{1}{2}E_v & x_{uu}\cdot x_{vv} - x_{uv}\cdot x_{uv}\end{bmatrix} - \det\begin{bmatrix} E & F & \frac{1}{2}E_v \\ F & G & \frac{1}{2}G_u \\ \frac{1}{2}E_v & \frac{1}{2}G_u & 0\end{bmatrix}.$$

The last identity can be verified by expansion of the determinant along the last column. It remains to see that $x_{uu}\cdot x_{vv} - x_{uv}\cdot x_{uv}$ can be expressed in terms of the derivatives of E, F and G. Using the second lemma below, the required formula is

$$x_{uu}\cdot x_{vv} - x_{uv}\cdot x_{uv} = F_{uv} - \frac{1}{2}E_{vv} - \frac{1}{2}G_{uu}\,.$$

□

Lemma 25.8.

$$\begin{bmatrix} x_u\cdot x_{uu} & x_u\cdot x_{uv} & x_u\cdot x_{vv} \\ x_v\cdot x_{uu} & x_v\cdot x_{uv} & x_v\cdot x_{vv}\end{bmatrix} = \begin{bmatrix} \frac{1}{2}E_u & \frac{1}{2}E_v & F_v - \frac{1}{2}G_u \\ F_u - \frac{1}{2}E_v & \frac{1}{2}G_u & \frac{1}{2}G_v\end{bmatrix}.$$

Proof. Taking partial derivatives of the definitions $E = x_u\cdot x_u$, $F = x_u\cdot x_v$ and $G = x_v\cdot x_v$ with respect to u, and to v, we get

$$\begin{aligned} E_u &= 2x_u\cdot x_{uu} \\ E_v &= 2x_u\cdot x_{uv} \\ F_u &= x_u\cdot x_{uv} + x_v\cdot x_{uu} \\ F_v &= x_u\cdot x_{vv} + x_v\cdot x_{uv} \\ G_u &= 2x_v\cdot x_{uv} \\ G_v &= 2x_v\cdot x_{vv}\,. \end{aligned}$$

The claim follows by solving these linear equations. □

Lemma 25.9. $x_{uu}\cdot x_{vv} - x_{uv}\cdot x_{uv} = -\frac{1}{2}E_{vv} + F_{uv} - \frac{1}{2}G_{uu}$.

Proof. Taking partial derivatives once more, we get

$$\begin{aligned} E_{uu} &= 2x_u\cdot x_{uuu} + 2x_{uu}\cdot x_{uu} \\ E_{uv} &= 2x_u\cdot x_{uuv} + 2x_{uu}\cdot x_{uv} \\ E_{vv} &= 2x_u\cdot x_{uvv} + 2x_{uv}\cdot x_{uv} \\ F_{uu} &= x_u\cdot x_{uuv} + x_v\cdot x_{uuu} + 2x_{uu}\cdot x_{uv} \\ F_{uv} &= x_u\cdot x_{uvv} + x_v\cdot x_{uuv} + x_{uu}\cdot x_{vv} + x_{uv}\cdot x_{uv} \\ F_{vv} &= x_u\cdot x_{vvv} + x_v\cdot x_{uvv} + 2x_{uv}\cdot x_{vv} \\ G_{uu} &= 2x_v\cdot x_{uuv} + 2x_{uv}\cdot x_{uv} \\ G_{uv} &= 2x_v\cdot x_{uvv} + 2x_{uv}\cdot x_{vv} \\ G_{vv} &= 2x_v\cdot x_{vvv} + 2x_{vv}\cdot x_{vv}\,. \end{aligned}$$

Comparing the expressions for E_{vv}, F_{uv} and G_{uu} gives the result. □

26. November 20th lecture

26.1. **Geodesics.** We introduce geodesic curves on a surface as curves that do not change direction, as seen from the surface. We concentrate on curves $\alpha = \beta\circ s^{-1}$ that are parametrized by arc length.

Definition 26.1. Let $S \subset \mathbb{R}^3$ be a regular surface, and let $\alpha\colon [0,\ell] \to S$ be a smooth curve on that surface, parametrized by arc length. For each $s \in [0,\ell]$ let

$$T(s) = \alpha'(s)$$

be the *unit tangent vector* of α at $p = \alpha(s)$. Let the *unit bitangent vector*

$$B(s) \in T_pS$$

be chosen so that $(T(s), B(s))$ is an orthonormal basis for the tangent plane T_pS. Let the *unit normal vector*

$$N(s) = T(s) \times B(s)$$

be given by the cross product. Then $(T(s), B(s), N(s))$ is a positively oriented orthonormal basis for $\mathbb{R}^3$. Let $\alpha''(s)$ denote the second derivative at s of α, viewed as a smooth curve $\iota \circ \alpha\colon [0,\ell] \to \mathbb{R}^3$.

The term "bitangent" may not be standard. Jahren writes $n_\alpha(s)$ for $B(s)$. Orthonormality means that $\|T(s)\| = 1$, $T(s) \cdot B(s) = 0$ and $\|B(s)\| = 0$. If β parametrizes the boundary of a region $R \subset S$ we may assume that $B(s)$ points into R. If S is oriented, we may choose $B(s)$ so that $N(s) = N_p$ is the preferred unit normal vector.

Lemma 26.2. *$T(s) \cdot \alpha''(s) = 0$ for all $s \in [0,\ell]$.*

Proof. By assumption, α is parametrized by arc length, so $\alpha'(s) \cdot \alpha'(s) = \|\alpha'(s)\|^2 = 1$ for all s. Differentiating we get $2\alpha'(s) \cdot \alpha''(s) = 0$. □

Definition 26.3. Let $\alpha\colon [0,\ell] \to S$ be parametrized by arc length. Let

$$k_g(s) = B(s) \cdot \alpha''(s)$$

be the *geodesic curvature* of α at s, and let

$$\nu(s) = N(s) \cdot \alpha''(s)$$

be the component of $\alpha''(s)$ that is normal to T_pS. Then

$$\alpha''(s) = k_g(s)B(s) + \nu(s)N(s)\,.$$

We say that the curve α is a *geodesic* if $k_g(s) = 0$ for all s.

In this case the summand $k_g(s)B(s)$ equals the covariant second derivative $D\alpha''(s)$ of α. A curve parametrized by arc length is a geodesic precisely if its geodesic curvature is zero. A general regular curve $\beta\colon [a,b] \to S$ is a geodesic if its reparametrization by arc length, $\alpha = \beta \circ s^{-1}$, is a geodesic.

Example 26.4. In the Euclidean plane $\mathbb{R}^2$, each straight line segment of length ℓ can be parametrized by $\alpha(s) = p + sv$ with $p \in \mathbb{R}^2$ and $\|v\| = 1$. Then $\alpha'(s) = v$ and $\alpha''(s) = 0$, so $k_g(s) = 0$ for all s. Hence these Euclidean line segments are geodesic curves. More generally, any Euclidean line segment $[p,q]$ contained in a regular surface $S \subset \mathbb{R}^3$ is a geodesic in that surface.

Example 26.5. In the Euclidean plane $\mathbb{R}^2$, a circle of radius R, centered at the origin, can be parametrized at unit speed by $\alpha(s) = (R\cos(s/R), R\sin(s/R))$. Then $\alpha'(s) = (-\sin(s/R), \cos(s/R))$ and $\alpha''(s) = (-1/R)(\cos(s/R), \sin(s/R))$. We can take $B(s) = -(\cos(s/R), \sin(s/R))$, so $k_g(s) = 1/R$ for all s. The geodesic curvature equals the inverse of the radius of the circle.

Example 26.6. In the sphere S of radius R, centered at the origin, the curve

$$\alpha(s) = \cos(s/R)p + \sin(s/R)q$$

is parametrized by arc length if p and q are orthogonal vectors with $\|p\| = \|q\| = R$. Here $\alpha'(s) = -\sin(s/R)(p/R) + \cos(s/R)(q/R)$ and $\alpha''(s) = -\cos(s/R)(p/R^2) - \sin(s/R)(q/R^2)$. Since $\alpha''(s) = (-1/R^2)\alpha(s) = (-1/R)N(s)$, where $N(s) = \alpha(s)/R$ is the outward pointing unit normal at $\alpha(s)$, we get $k_g(s) = 0$ and $\nu(s) = -1/R$ for all s. Hence these segments of great circles, with $\alpha(0) = p$ and $\alpha'(0) = q/R$, are geodesics.

We shall see below that at any point $p \in S$ and for any unit vector $v \in T_pS$ there is a geodesic $\alpha\colon [0,\ell] \to S$ with $\alpha(0) = p$ and $\alpha'(0) = v$, at least for ℓ sufficiently small, and that this geodesic is essentially unique. The examples above will then show that the segments of straight lines and great circles are the only unit speed geodesic curves, in the Euclidean plane and on the sphere of radius R, respectively.

26.2. **The geodesic equations.**

Definition 26.7. Let $x\colon V \to S$ be a local parametrization. At each point $p = x(u,v)$ in $x(V)$ the tangent vectors $x_u(u,v)$ and $x_v(u,v)$, together with the unit normal vector N_p give an ordered basis for $\mathbb{R}^3$. The second derivatives x_{uu}, x_{uv} and x_{vv}, of x viewed as a smooth map $\iota \circ x\colon V \to \mathbb{R}^3$, can be written in terms of this basis, as

$$\begin{aligned} x_{uu} &= \Gamma^1_{11} x_u + \Gamma^2_{11} x_v + eN \\ x_{uv} &= \Gamma^1_{12} x_u + \Gamma^2_{12} x_v + fN \\ x_{vv} &= \Gamma^1_{22} x_u + \Gamma^2_{22} x_v + gN\,. \end{aligned}$$

Here the smooth functions $\Gamma^k_{ij}\colon V \to \mathbb{R}$ are called the *Christoffel symbols* (of the second kind) of the parametrization.

The functions e, f and g are the components of the second fundamental form, as introduced earlier.

Theorem 26.8. *Consider a unit speed curve $\alpha\colon [a,b] \to S$, and assume that its image lies in $x(V)$ for a local parametrization $x\colon V \to S$, so that $\alpha(s) = x(u(s), v(s))$ for smooth functions $(u,v)\colon [a,b] \to V$. Then α is a geodesic if and only if u and v satisfy the system of differential equations*

$$\begin{aligned} u'' + (u')^2\Gamma^1_{11} + 2u'v'\Gamma^1_{12} + (v')^2\Gamma^1_{22} &= 0 \\ v'' + (u')^2\Gamma^2_{11} + 2u'v'\Gamma^2_{12} + (v')^2\Gamma^2_{22} &= 0 \end{aligned}$$

(the geodesic equations).

Proof. Differentiating, we find

$$\alpha'(s) = u'x_u + v'x_v$$

and

$$\alpha''(s) = u''x_u + v''x_v + (u')^2 x_{uu} + 2u'v'x_{uv} + (v')^2 x_{vv}$$

(with u', v', u'' and v'' evaluated at s, and x_u, x_v, x_{uu}, x_{uv} and x_{vv} evaluated at $(u(s), v(s))$).

In terms of the basis (x_u, x_v, N), we find

$$\begin{aligned} \alpha''(s) = (u'' &+ (u')^2\Gamma^1_{11} + 2u'v'\Gamma^1_{12} + (v')^2\Gamma^1_{22})x_u \\ &+ (v'' + (u')^2\Gamma^2_{11} + 2u'v'\Gamma^2_{12} + (v')^2\Gamma^2_{22})x_v \\ &+ (e(u')^2 + 2fu'v' + g(v')^2)N\,. \end{aligned}$$

The curve α is a (unit speed) geodesic if and only if the component of $\alpha''(s)$ in T_pS vanishes for all s, where $p = \alpha(s)$. □

Lemma 26.9.

$$\begin{bmatrix} E & F \\ F & G \end{bmatrix} \begin{bmatrix} \Gamma^1_{11} & \Gamma^1_{12} & \Gamma^1_{22} \\ \Gamma^2_{11} & \Gamma^2_{12} & \Gamma^2_{22} \end{bmatrix} = \begin{bmatrix} \frac12 E_u & \frac12 E_v & F_v - \frac12 G_u \\ F_u - \frac12 E_v & \frac12 G_u & \frac12 G_v \end{bmatrix}.$$

Proof. Taking dot products of x_u and x_v with x_{uu}, x_{uv} and x_{vv}, and recalling that $x_u \cdot N = x_v \cdot N = 0$, we get

$$\begin{aligned} x_u \cdot x_{uu} &= \Gamma^1_{11}E + \Gamma^2_{11}F \\ x_u \cdot x_{uv} &= \Gamma^1_{12}E + \Gamma^2_{12}F \\ x_u \cdot x_{vv} &= \Gamma^1_{22}E + \Gamma^2_{22}F \\ x_v \cdot x_{uu} &= \Gamma^1_{11}F + \Gamma^2_{11}G \\ x_v \cdot x_{uv} &= \Gamma^1_{12}F + \Gamma^2_{12}G \\ x_v \cdot x_{vv} &= \Gamma^1_{22}F + \Gamma^2_{22}G\,. \end{aligned}$$

Here E, F and G are the components of the first fundamental form. Now use a lemma from the proof of theorema egregrium. □

Corollary 26.10. *The Christoffel symbols and geodesic curves are intrinsic.*

Proof. The Γ^k_{ij} can be expressed in terms of the first fundamental form and its derivatives, hence are preserved by isometries. The system of geodesic equations is therefore also preserved by isometries, and so are its solutions. □

Example 26.11. The upper half-plane model $\mathbb{H}$ for the hyperbolic plane is a Riemannian surface with (local) trivialization with coordinates $p = (u, v)$ with $v > 0$ and first fundamental form $E(u, v) = G(u, v) = 1/v^2$, $F(u, v) = 0$. By the lemma above

$$\begin{bmatrix} 1/v^2 & 0 \\ 0 & 1/v^2 \end{bmatrix} \begin{bmatrix} \Gamma^1_{11} & \Gamma^1_{12} & \Gamma^1_{22} \\ \Gamma^2_{11} & \Gamma^2_{12} & \Gamma^2_{22} \end{bmatrix} = \begin{bmatrix} 0 & -1/v^3 & 0 \\ 1/v^3 & 0 & -1/v^3 \end{bmatrix},$$

so $\Gamma^1_{11} = \Gamma^1_{22} = \Gamma^2_{12} = 0$ and $-\Gamma^1_{12} = \Gamma^2_{11} = -\Gamma^2_{22} = 1/v$. The geodesic equations are:

$$\begin{aligned} u'' - \frac{2u'v'}{v} &= 0 \\ v'' + \frac{(u')^2}{v} - \frac{(v')^2}{v} &= 0\,. \end{aligned}$$

One set of solutions to this system of equations consists of vertical curves $\alpha(s) = (u(s), v(s))$, with $u(s) = a$ and $v(s) = be^s$, for constants a and b. These curves are parametrized by arc length, since $\alpha'(s) = (0, be^s)$, so $\|\alpha'(s)\|_p = be^s/be^s = 1$ at $p = \alpha(s)$. Furthermore, $u' = u'' = 0$ and $v = v' = v''$, so both geodesic equations are satisfied. Hence for each point $p = (a, b) \in \mathbb{H}$ there is a geodesic $\alpha\colon [0, \infty) \to \mathbb{H}$ with $\alpha(0) = (a, b)$ and $\alpha'(0) = (0, 1)$. In other words, segments of the vertical $\mathbb{H}$-lines are geodesics.

Each Möbius transformation $\gamma\colon \mathbb{H} \to \mathbb{H}$ is an isometry, both in the metric and the Riemannian sense, hence maps geodesics to geodesics. Since any $\mathbb{H}$-line in $\mathbb{H}$ is the image of a vertical $\mathbb{H}$-line by a Möbius transformation, it follows that all segments of $\mathbb{H}$-lines, vertical or not, are geodesics in the hyperbolic plane.

26.3. **The exponential map.** The general theory of ordinary differential equations leads to the following result.

Proposition 26.12. *Let S be a Riemannian surface. For each point $p \in S$ there is an $\epsilon > 0$ such that for each unit vector $v \in T_pS$ there is a unique unit speed geodesic $\gamma^p_v\colon [-\epsilon, \epsilon] \to S$ such that*

$$\gamma^p_v(0) = p \qquad \text{and} \qquad (\gamma^p_v)'(0) = v\,.$$

Moreover, we can choose the same $\epsilon > 0$ for each point q in a neighborhood U of p, and $\gamma^q_v(s)$ depends smoothly on q, $v \in T_qS$ and s (with $q \in U$, $\|v\|_q = 1$ and $|s| \le \epsilon$).

By the uniqueness, $\gamma^p_v(-s) = \gamma^p_{-v}(s)$ for all $|s| \le \epsilon$.

Definition 26.13. Let

$$D_p(\epsilon) = \{v \in T_pS \mid \|v\|_p \le \epsilon\}$$

be the closed ϵ-disc around the origin in T_pS. Define the *exponential map*

$$\exp_p \colon D_p(\epsilon) \longrightarrow S$$

by

$$\exp_p(sv) = \gamma_v^p(s)$$

for $s \in [0, \epsilon]$ and $v \in T_pS$ with $\|v\|_p = 1$. Here $\exp_p(0) = p$ and $\beta \colon s \mapsto \exp_p(sv)$ for $s \in [-\epsilon, \epsilon]$ is a unit speed geodesic with $\beta(0) = p$ and $\beta'(0) = v$.

Remark 26.14. The name "exponential map" comes from a corresponding construction for the Lie group $GL_n(\mathbb{R})$ of invertible real $n \times n$ matrices, whose tangent space $T_IGL_n(\mathbb{R})$ at the identity matrix I can be identified with the vector space $M_n(\mathbb{R})$. There is a map $\exp_I \colon T_IGL_n(\mathbb{R}) \to GL_n(\mathbb{R})$, also defined in terms of unit speed geodesics, which in this case turns out to be given by

$$\exp_I(A) = \sum_{k \ge 0} \frac{A^k}{k!}$$

for all $A \in M_n(\mathbb{R})$. In particular, for $n = 1$ this is the usual exponential function $a \mapsto \exp(a) = e^a \colon \mathbb{R} \to GL_1(\mathbb{R})$.

Proposition 26.15. *For every $p \in S$ there is an $\epsilon > 0$ such that $\exp_p$ is a diffeomorphism between $D_p(\epsilon)$ and a neighborhood of p in S.*

Proof. The differential

$$d(\exp_p)_0 \colon T_0(T_pS) \to T_pS$$

of $\exp_p$ at $0 \in T_pS$ equals the canonical identification $T_0(T_pS) \cong T_pS$. Hence $\exp_p$ is a local diffeomorphism near 0, by the inverse function theorem. □

27. November 25th lecture

27.1. **Geodesic polar coordinates.** The first fundamental form (E, F, G) for a local parametrization $x \colon V \to x(V) \subset S$ simplifies to $E = 1$, $F = 0$ and $G > 0$ if the coordinate curves $u \mapsto x(u, v)$ have unit speed and meet the coordinate curves $v \mapsto x(u, v)$ at right angles. We now use the exponential map to find such local parametrizations.

Definition 27.1. Let p be a point in a Riemannian surface S, with exponential map

$$\exp_p \colon D_p(\epsilon) \to S\,,$$

and fix an ordered orthonormal basis (b_1, b_2) for the tangent plane T_pS. Define a smooth map $x \colon [-\epsilon, \epsilon] \times \mathbb{R} \longrightarrow S$ by

$$x(r, \theta) = \exp_p(r\cos(\theta)b_1 + r\sin(\theta)b_2)\,.$$

When restricted to a subset $V = (0, \epsilon) \times J$, where $J \subset \mathbb{R}$ is an open interval of length $\le 2\pi$, the map $x \colon V \to x(V) \subset S$ is a local parametrization of S. In this case we call the pair (r, θ) *geodesic polar coordinates* on $x(V)$.

The tangent vectors $x_r(r, \theta)$ and $x_\theta(r, \theta)$ are defined for all $|r| \le \epsilon$ and $\theta \in \mathbb{R}$, but are not linearly independent for $r = 0$. In fact $x_r(0, \theta) = x_\theta(0, \theta) = 0$ at $x(0, \theta) = p$.

Lemma 27.2 (Gauss' lemma)**.** *In geodesic polar coordinates*

$$E = x_r \cdot x_r = 1 \qquad \textit{and} \qquad F = x_r \cdot x_\theta = 0\,.$$

Proof. For each fixed θ, with associated unit vector $v = \cos(\theta)b_1 + \sin(\theta)b_2$, the coordinate curve

$$r \mapsto \alpha(r) = x(r, \theta)$$

(a *geodesic radius*) is a unit speed geodesic

$$\alpha(r) = \exp_p(rv) = \gamma_v^p(r) .$$

From the fact that α is parametrized at unit speed we deduce that

$$E(r, \theta) = x_r(r, \theta) \cdot x_r(r, \theta) = \|\alpha'(r)\|^2 = 1$$

for all r and θ. From the further fact that α is a geodesic we deduce that

$$x_{rr}(r, \theta) = \alpha''(r) = eN_p$$

is orthogonal to T_pS at $p = x(r, \theta)$, so

$$\Gamma_{11}^1 x_r(r, \theta) + \Gamma_{11}^2 x_\theta(r, \theta) = 0$$

for all r and θ. Here $x_r(r, \theta)$ and $x_\theta(r, \theta)$ are linearly independent for $r \neq 0$, so $\Gamma_{11}^1 = \Gamma_{11}^2 = 0$. It follows that $F_r - \frac{1}{2}E_\theta = F\Gamma_{11}^1 + G\Gamma_{11}^2 = 0$, so $F_r = 0$ and F is independent of r. Hence

$$F(r, \theta) = x_r(r, \theta) \cdot x_\theta(r, \theta) = x_r(0, \theta) \cdot x_\theta(0, \theta) = 0$$

by passage to the limit as $r \to 0^+$. □

Definition 27.3. In geodesic polar coordinates we can write the first fundamental form as

$$ds^2 = dr^2 + h^2\, d\theta^2$$

where $h = \|x_\theta\|$ and $G = x_\theta \cdot x_\theta = h^2$.

Example 27.4. For $S = \mathbb{R}^2$, with $p = (0, 0)$, $b_1 = (1, 0)$ and $b_2 = (0, 1)$, the geodesic polar coordinates

$$x(r, \theta) = (r \cos\theta, r \sin\theta)$$

agree with the usual polar coordinates. The first fundamental form equals $ds^2 = dr^2 + r^2\, d\theta^2$.

For $S = S^2$, with $p = (0, 0, 1)$ and b_1 and b_2 as above, the geodesic polar coordinates

$$x(r, \theta) = (\sin r \cos\theta, \sin r \sin\theta, \cos r)$$

agree with spherical coordinates. The first fundamental form equals $ds^2 = dr^2 + \sin^2 r\, d\theta^2$.

For $S = \mathbb{D}$ (the Poincaré disc model), with $p = (0, 0)$ and b_1 and b_2 as above, the geodesic polar coordinates are

$$x(r, \theta) = (\tanh(r/2) \cos\theta, \tanh(r/2) \sin\theta) .$$

The first fundamental form equals $ds^2 = dr^2 + \sinh^2 r\, d\theta^2$.

27.2. **Geodesics as shortest curves.**

Theorem 27.5. *Let S be a Riemannian surface. Each point $p \in S$ has a neighborhood U such that any point $q \in U \setminus \{p\}$ can be connected to p by a unique shortest curve, and this curve is a geodesic.*

Proof. Let $\epsilon > 0$ be such that $\exp_p \colon D_p(\epsilon) \to S$ is a diffeomorphism onto its image $U \subset S$. Then any $q \in U \setminus \{p\}$ is of the form $\exp_p(\rho v)$ for a unique $\rho \in (0, \epsilon)$ and $\|v\| = 1$, and the curve $\gamma_v^p \colon [0, \rho] \to S$ is a geodesic of length ρ from p to q.

Let $\beta \colon [a, b] \to S$ be another curve from p to q. We may assume that $\beta(t) \neq p$ for every $t \in (a, b]$, and that the image of β lies in $U \subset S$. [[See Jahren's book for more details.]] Then we can write $\beta(t) = x(r(t), \theta(t))$ for smooth functions $r \colon [a, b] \to [0, \epsilon)$ and $\theta \colon [a, b] \to \mathbb{R}$, with $r(a) = 0$ and $r(b) = \rho$. Then

$$\beta'(t) = r'(t)x_r + \theta'(t)x_\theta$$

and

$$\|\beta'(t)\| = \sqrt{r'(t)^2 + \theta'(t)^2 h^2} \geq |r'(t)|$$

so

$$\text{length}(\beta) = \int_a^b \|\beta'(t)\|\,dt \geq \int_a^b |r'(t)|\,dt \geq \int_a^b r'(t)\,dt = r(b) - r(a) = \rho\,.$$

We have equality only if $r'(t) \geq 0$ and $\theta'(t) = 0$, meaning that β is a reparametrization of the unit speed geodesic γ_v^p. □

27.3. **Calculations.** Let $S \subset \mathbb{R}^3$ be a regular surface and let $U = x(V) \subset S$ be parametrized by geodesic polar coordinates $(r, \theta) \mapsto x(r, \theta) \in S$. Write the first fundamental form as

$$ds^2 = dr^2 + h^2\,d\theta^2$$

with $h > 0$, so that x_r and x_θ/h form an orthonormal basis for each tangent plane.

Proposition 27.6. *(a) The Gaussian curvature is given by the formula*

$$K = -\frac{h_{rr}}{h}\,.$$

(b) The Christoffel symbols are

$$\begin{bmatrix} \Gamma_{11}^1 & \Gamma_{12}^1 & \Gamma_{22}^1 \\ \Gamma_{11}^2 & \Gamma_{12}^2 & \Gamma_{22}^2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & -hh_r \\ 0 & h_r/h & h_\theta/h \end{bmatrix}\,.$$

(c) Let $s \mapsto \alpha(s) = x(r(s), \theta(s))$ be a curve in U, parametrized by arc length, and let $\phi(s)$ be the angle between the geodesic radius and the curve at $\alpha(s)$. The geodesic curvature of α is then

$$k_g = \phi' + h_r\theta'\,.$$

Proof. (a) From the proof of theorema egregium, with $E = 1$, $F = 0$ and $G = h^2$, we get

$$K(h^2)^2 = \det\begin{bmatrix} 1 & 0 & -hh_r \\ 0 & h^2 & hh_\theta \\ 0 & 0 & -h_r^2 - hh_{rr} \end{bmatrix} - \det\begin{bmatrix} 1 & 0 & 0 \\ 0 & h^2 & hh_r \\ 0 & hh_r & 0 \end{bmatrix} = -h^3h_{rr}\,.$$

(b) From the proof that Christoffel symbols are intrinsic, we get

$$\begin{bmatrix} 1 & 0 \\ 0 & h^2 \end{bmatrix}\begin{bmatrix} \Gamma_{11}^1 & \Gamma_{12}^1 & \Gamma_{22}^1 \\ \Gamma_{11}^2 & \Gamma_{12}^2 & \Gamma_{22}^2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & -hh_r \\ 0 & hh_r & hh_\theta \end{bmatrix}\,.$$

(c) By assumption, α is parametrized by arc length, so $\alpha' = r'x_r + \theta' x_\theta$ has unit length. Hence

$$\alpha' = r'x_r + h\theta'\frac{x_\theta}{h} = \cos(\phi)x_r + \sin(\phi)\frac{x_\theta}{h}$$

for a smooth function $s \mapsto \phi(s)$, giving the angle between the geodesic radius (with tangent vector x_r) and the curve (with tangent vector α'). The unit bitangent vector field $s \mapsto B(s)$ is then

$$B = -\sin(\phi)x_r + \cos(\phi)\frac{x_\theta}{h}\,.$$

Differentiating $r' = \cos\phi$ and $h\theta' = \sin\phi$ gives

$$r'' = -\sin(\phi)\phi' \qquad \text{and} \qquad (h_r r' + h_\theta\theta')\theta' + h\theta'' = \cos(\phi)\phi'\,.$$

In terms of Christoffel symbols, the second derivative is

$$\begin{aligned}
\alpha'' &\equiv (r'' + (\theta')^2(-hh_r))x_r + (\theta'' + 2r'\theta'\frac{h_r}{h} + (\theta')^2\frac{h_\theta}{h})x_\theta \\
&= (r'' - (h\theta')(h_r\theta'))x_r + (h\theta'' + h_r r'\theta' + h_\theta(\theta')^2 + h_r\theta' r')\frac{x_\theta}{h} \\
&= (-\sin(\phi)\phi' - \sin(\phi)h_r\theta')x_r + (\cos(\phi)\phi' + h_r\theta'\cos(\phi))\frac{x_\theta}{h} \\
&= (\phi' + h_r\theta')B
\end{aligned}$$

plus a multiple of the unit normal N, so the geodesic curvature is $k_g = B \cdot \alpha'' = \phi' + h_r\theta'$. □

28. November 27th lecture

28.1. Line and surface integrals.

Definition 28.1. If $\beta\colon [a,b] \to \mathbb{R}^3$ is a regular curve, and $f\colon C \to \mathbb{R}$ is a map defined on the image $C = \beta([a,b])$ of β, then the *line integral* of f along C is defined to be

$$\int_C f\,ds = \int_a^b f(\beta(t))s'(t)\,dt\,.$$

Here $s'(t) = \|\beta'(t)\|$ is the length of the tangent vector $\beta'(t)$, and the line integral is independent of the (simple) parametrization β of C. If C is a union of finitely many smooth curves, the line integral over C is defined to be the sum of the line integrals over the smooth pieces.

Definition 28.2. If $x\colon V \to S$ is a local parametrization and a map $f\colon R \to \mathbb{R}$ is defined on a nice compact region $R \subset x(V)$, the *surface integral* of f over R is defined to be

$$\iint_R f\,dA = \iint_{x^{-1}(R)} f(x(u,v))\sqrt{EG-F^2}\,du\,dv\,.$$

Here $\sqrt{EG-F^2} = \|x_u \times x_v\|$ is the area of the parallelogram spanned by the tangent vectors x_u and x_v. The region is nice if it is bounded by finitely many smooth curves. The surface integral is independent of the choice of local parametrization. If the region $R \subseteq S$ is not contained in a single coordinate patch, subdivide it into smaller pieces and define the surface integral as the sum over the pieces.

28.2. The Gauss–Bonnet theorem.

Let S be a regular surface in $\mathbb{R}^3$, and suppose that $R \subseteq S$ is a nice compact region, with boundary ∂R.

For each smooth piece of ∂R, parametrized by $\alpha_k\colon [0,\ell] \to S$, we choose the unit bitangent vector $B(s) \in T_pS$ at $p = \alpha_k(s) \in \partial R$ that points *into* R. For each non-smooth point p_k of ∂R, let $\eta_k \in [0,2\pi]$ be the *interior angle* between the two smooth curves meeting at p_k. Let $\epsilon_k = \pi - \eta_k \in [-\pi,\pi]$ be the angular *change of direction* at p_k.

Let $K\colon S \to \mathbb{R}$ be the Gaussian curvature, let $k_g\colon \partial R \to \mathbb{R}$ be the geodesic curvature, and let $\chi(R) \in \mathbb{Z}$ be the Euler characteristic of R.

Theorem 28.3 (Gauss–Bonnet)**.**

$$\iint_R K\,dA + \int_{\partial R} k_g\,ds + \sum_i \epsilon_i = 2\pi\chi(R)\,.$$

Remark 28.4. The left hand side is analytical and depends on geometric information, while the right hand side only depends on topological information. The right hand side only takes values that are integer multiples of 2π, which is not evident from the form of the left hand side.

If ∂R is smooth (as a collection of closed curves), the formula simplifies to

$$\iint_R K\,dA + \int_{\partial R} k_g\,ds = 2\pi\chi(R)\,.$$

If ∂R consists of finitely many geodesics, the formula simplifies to

$$\iint_R K\,dA + \sum_i \epsilon_i = 2\pi\chi(R)\,.$$

If S is compact (a closed regular surface) we may take $R = S$ with ∂R empty, and the Gauss–Bonnet formula simplifies to

$$\iint_S K\,dA = 2\pi\chi(S)\,.$$

Proof of the Gauss-Bonnet theorem. The region R can be written as the union of a finite collection $\{T_i\}_i$ of f smoothly embedded triangles, where the intersection of two triangles is either a common face, a common vertex, or empty.

We may assume that each triangle T_i is small enough to be contained in a coordinate patch $U_i = x_i(V_i)$ parametrized by geodesic polar coordinates $x_i \colon V_i \to S$. Here U_i is contained in a larger neighborhood W_i which also contains the center $x(0,\theta)$ of the geodesic polar coordinate system. The local parametrization x_i specifies an orientation in U_i.

We first prove the theorem for $R = T_i$ equal to one of these triangles. The boundary ∂T_i is the union of three smooth curves $\alpha_{i,j} \colon [0,\ell_j] \to S$, for $j = 1,2,3 \pmod 3$, traversed in counterclockwise order and parametrized by arc length. Let $\epsilon_{i,j}$ be the angle between $\alpha'_{i,j}(\ell_j)$ and $\alpha'_{i,j+1}(0)$.

Recall that $k_g = \phi' + h_r\theta'$ in geodesic polar coordinates, where $\alpha(s) = x(r(s),\theta(s))$, and ϕ gives the angle between the geodesic radius and the curve. The two parts of the integral

$$\int_{\partial T_i} k_g\,ds = \int_{\partial T_i} \phi'\,ds + \int_{\partial T_i} h_r\theta'\,ds$$

are computed separately.

First, write ϕ_j for the function ϕ associated to $\alpha_{i,j}$. Then ("Hopf's Umlaufsatz"):

$$\begin{aligned}\int_{\partial T_i} \phi'\,ds &= \int_{\alpha_{i,1}} \phi'_1\,ds + \int_{\alpha_{i,2}} \phi'_2\,ds + \int_{\alpha_{i,3}} \phi'_3\,ds \\ &= (\phi_1(\ell_1) - \phi_1(0)) + (\phi_2(\ell_2) - \phi_2(0)) + (\phi_3(\ell_3) - \phi_3(0)) = 2\pi - (\epsilon_{i,1} + \epsilon_{i_2} + \epsilon_{i,3})\,.\end{aligned}$$

The term 2π comes from the fact that ∂T_i is traversed once in the clockwise direction. The terms $\epsilon_{i,1} + \epsilon_{i,2} + \epsilon_{i,3}$ measure the contributions to this clockwise rotation that are omitted at the three vertices of T_i.

Second, by Green's theorem

$$\int_{\partial T_i} h_r\theta'\,ds = \int_{\partial T_i} h_r\,d\theta = \iint_{T_i} h_{rr}\,dr\,d\theta = \iint_{T_i} \frac{h_{rr}}{h} h\,dr\,d\theta = -\iint_{T_i} K\,dA\,.$$

This uses that $K = -h_{rr}/h$ and $\sqrt{EG - F^2} = h$.

Taken together, we get

$$\iint_{T_i} K\,dA + \int_{\partial T_i} k_g\,ds + \sum_{j=1}^{3} \epsilon_{i,j} = 2\pi\,.$$

This proves the Gauss-Bonnet formula for $R = T_i$, since $\chi(T_i) = 1$ is the Euler characteristic of any triangular region. Summing over all f triangles we get

$$\iint_R K\,dA + \sum_i \int_{\partial T_i} k_g\,ds + \sum_{i,j} \epsilon_{i,j} = 2\pi f\,.$$

Each interior edge in the smooth triangulation of R contributes twice to the sum $\sum_i \int_{\partial T_i} k_g\,ds$, but the two terms cancel, because the unit bitangent vectors B point in opposite directions on the two occasions (into T_i when parametrized as $\alpha_{i,j}$, into $T_{i'}$ when parametrized as $\alpha_{i',j'}$), so that the geodesic curvatures k_g occur with opposite signs. Hence only the boundary edges $\alpha_k = \alpha_{i,j}$ in ∂R contribute, and we can write

$$\sum_i \int_{\partial T_i} k_g\,ds = \sum_k \int_{\alpha_k} k_g\,ds = \int_{\partial R} k_g\,ds\,.$$

Let $\eta_{i,j} = \pi - \epsilon_{i,j}$ be the interior angle at the j-th vertex of T_i. At each interior vertex of R the interior angles add up to 2π. At each boundary vertex the interior angles add up to π if the vertex is a smooth point of ∂R, or to $\eta_k = \pi - \epsilon_k$ if the vertex is a non-smooth point of ∂R. Summing,

$$\sum_{i,j} \eta_{i,j} = 2\pi(v - v_\partial) + \pi v_\partial - \sum_k \epsilon_k$$

where v is the total number of vertices in the triangulation of R and v_∂ is the number of vertices in ∂R. Each of the f triangles contributes three terms to the sum on the left hand side, so

$$\sum_{i,j} \eta_{i,j} = 3\pi f - \sum_{i,j} \epsilon_{i,j}$$

and

$$\sum_{i,j} \epsilon_{i,j} = 3\pi f - 2\pi v + \pi v_\partial + \sum_k \epsilon_k \,.$$

Each triangle has three edges, each interior edge lies in two triangles and each boundary edge lies in one triangle, so

$$3f = 2(e - e_\partial) + e_\partial = 2e - e_\partial \,,$$

where e is the total number of edges in the triangulation and e_∂ is the number of edges in ∂R. Furthermore $v_\partial = e_\partial$, since ∂R is a union of closed loops. Hence

$$3f - 2v + v_\partial = 2e - e_\partial - 2v + v_\partial = 2(e - v) \,.$$

Thus

$$\sum_{i,j} \epsilon_{i,j} = 2\pi(e - v) + \sum_k \epsilon_k$$

and

$$\iint_R K\,dA + \int_{\partial R} k_g\,ds + 2\pi(e - v) + \sum_k \epsilon_k = 2\pi f \,.$$

Subtracting $2\pi(e - v)$ on both sides gives the general Gauss-Bonnet formula, with $2\pi\chi(R) = 2\pi f - 2\pi(e - v) = 2\pi(v - e + f)$. □

28.3. **Applications.**

Example 28.5. Let S be a closed, connected Riemannian surface, so that

$$\iint_S K\,dA = 2\pi\chi(S) \,.$$

If the Gaussian curvature $K > 0$ is everywere positive, then $\iint_S K\,dA > 0$, so $\chi(S) > 0$. This implies that $S \cong M_0 = S^2$ (the sphere) or $S \cong N_1 = P^2$ (the projective plane). The same conclusion follows if $K \geq 0$ is everywhere non-negative, and positive at some point.

If the Gaussian curvature $K < 0$ is everywere negative, then $\iint_S K\,dA < 0$, so $\chi(S) < 0$. This implies that $S \cong M_g$ with $g \geq 2$ or $S \cong N_h$ with $h \geq 3$. The same conclusion follows if $K \leq 0$ is everywhere non-positive, and negative at some point.

If $\chi(S) = 0$, so $M \cong M_1 = T^2$ (the torus) or $M \cong N_2 = K^2$ (the Klein bottle), we must have $\iint_S K\,dA = 0$, so either $K = 0$ everywhere, or K takes both positive and negative values on S.

Example 28.6. If S has constant curvature K, and $R \subseteq S$ is a nice compact region, we have

$$\iint_R K\,dA = K\,\mathrm{area}(R)$$

where $\mathrm{area}(R) = \iint_R dA$ is the area of R. If R is a triangle with geodesic sides and interior angles α, β and γ, then the direction changes are $\pi - \alpha$, $\pi - \beta$ and $\pi - \gamma$, and the Euler characteristic is $\chi(R) = 1$, so

$$K\,\mathrm{area}(R) = \alpha + \beta + \gamma - \pi \,.$$

This specializes to Girard's theorem in spherical geometry when $K = +1$, and to Lambert's theorem in hyperbolic geometry when $K = -1$. It specializes to Euclid's proposition $\alpha + \beta + \gamma = \pi$ when $K = 0$.

[[See Jahren's book for more examples.]]