Part IB — Geometry

Based on lectures by A. G. Kovalev

Notes taken by Dexter Chua

Lent 2016

These notes are not endorsed by the lecturers, and I have modified them (often significantly) after lectures. They are nowhere near accurate representations of what was actually lectured, and in particular, all errors are almost surely mine.

Parts of Analysis II will be found useful for this course.

Groups of rigid motions of Euclidean space. Rotation and reflection groups in two and three dimensions. Lengths of curves. [2]

Spherical geometry: spherical lines, spherical triangles and the Gauss-Bonnet theorem. Stereographic projection and Möbius transformations. [3]

Triangulations of the sphere and the torus, Euler number. [1]

Riemannian metrics on open subsets of the plane. The hyperbolic plane. Poincaré models and their metrics. The isometry group. Hyperbolic triangles and the Gauss-Bonnet theorem. The hyperboloid model. [4]

Embedded surfaces in $\mathbb{R}^3$. The first fundamental form. Length and area. Examples. [1]

Length and energy. Geodesics for general Riemannian metrics as stationary points of the energy. First variation of the energy and geodesics as solutions of the corresponding Euler-Lagrange equations. Geodesic polar coordinates (informal proof of existence). Surfaces of revolution. [2]

The second fundamental form and Gaussian curvature. For metrics of the form $du^2 + G(u, v)dv^2$, expression of the curvature as $\sqrt{G_{uu}}/\sqrt{G}$. Abstract smooth surfaces and isometries. Euler numbers and statement of Gauss-Bonnet theorem, examples and applications. [3]

Contents

0 Introduction

In the very beginning, Euclid came up with the axioms of geometry, one of which is the *parallel postulate.* This says that given any point P and a line ℓ not containing P, there is a line through P that does not intersect ℓ. Unlike the other axioms Euclid had, this was not seen as "obvious". For many years, geometers tried hard to prove this axiom from the others, but failed.

Eventually, people realized that this axiom *cannot* be proved from the others. There exists some other "geometries" in which the other axioms hold, but the parallel postulate fails. This was known as *hyperbolic geometry.* Together with Euclidean geometry and spherical geometry (which is the geometry of the surface of a sphere), these constitute the three classical geometries. We will study these geometries in detail, and see that they actually obey many similar properties, while being strikingly different in other respects.

That is not the end. There is no reason why we have to restrict ourselves to these three types of geometry. In later parts of the course, we will massively generalize the notions we began with and eventually define an *abstract smooth surface.* This covers all three classical geometries, and many more!

1 Euclidean geometry

We are first going to look at Euclidean geometry. Roughly speaking, this is the geometry of the familiar $\mathbb{R}^n$ under the usual inner product. There really isn't much to say, since we are already quite familiar with Euclidean geometry. We will quickly look at isometries of $\mathbb{R}^n$ and curves in $\mathbb{R}^n$. Afterwards, we will try to develop analogous notions in other more complicated geometries.

1.1 Isometries of the Euclidean plane

The purpose of this section is to study maps on $\mathbb{R}^n$ that preserve distances, i.e. *isometries* of $\mathbb{R}^n$. Before we begin, we define the notion of distance on $\mathbb{R}^n$ in the usual way.

Definition ((Standard) inner product)**.** The *(standard) inner product* on $\mathbb{R}^n$ is defined by

$$(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i.$$

Definition (Euclidean Norm)**.** The *Euclidean norm* of $\mathbf{x} \in \mathbb{R}^n$ is

$$\|\mathbf{x}\| = \sqrt{(\mathbf{x}, \mathbf{x})}.$$

This defines a metric on $\mathbb{R}^n$ by

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|.$$

Note that the inner product and the norm both depend on our choice of origin, but the distance does not. In general, we don't like having a choice of origin — choosing the origin is just to provide a (very) convenient way labelling points. The origin should not be a special point (in theory). In fancy language, we say we view $\mathbb{R}^n$ as an *affine space* instead of a *vector space*.

Definition (Isometry)**.** A map $f : \mathbb{R}^n \to \mathbb{R}^n$ is an *isometry* of $\mathbb{R}^n$ if

$$d(f(\mathbf{x}), f(\mathbf{y})) = d(\mathbf{x}, \mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Note that f is not required to be linear. This is since we are viewing $\mathbb{R}^n$ as an affine space, and linearity only makes sense if we have a specified point as the origin. Nevertheless, we will still view the linear isometries as "special" isometries, since they are more convenient to work with, despite not being special fundamentally.

Our current objective is to classify *all* isometries of $\mathbb{R}^n$. We start with the linear isometries. Recall the following definition:

Definition (Orthogonal matrix)**.** An $n \times n$ matrix A is *orthogonal* if $AA^T = A^T A = I$. The group of all orthogonal matrices is the orthogonal group $\mathrm{O}(n)$.

In general, for any matrix A and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we get

$$(A\mathbf{x}, A\mathbf{y}) = (A\mathbf{x})^T (A\mathbf{y}) = \mathbf{x}^T A^T A \mathbf{y} = (\mathbf{x}, A^T A \mathbf{y}).$$

So A is orthogonal if and only if $(A\mathbf{x}, A\mathbf{y}) = (\mathbf{x}, \mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Recall also that the inner product can be expressed in terms of the norm by

$$(\mathbf{x}, \mathbf{y}) = \frac{1}{2}(\|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{x}\|^2 - \|\mathbf{y}\|^2).$$

So if A preserves norm, then it preserves the inner product, and the converse is obviously true. So A is orthogonal if and only if $\|A\mathbf{x}\| = \|\mathbf{x}\|$ for all $\mathbf{x} \in \mathbb{R}^n$. Hence matrices are orthogonal if and only if they are isometries.

More generally, let

$$f(\mathbf{x}) = A\mathbf{x} + \mathbf{b}.$$

Then

$$d(f(\mathbf{x}), f(\mathbf{y})) = \|A(\mathbf{x} - \mathbf{y})\|.$$

So any f of this form is an isometry if and only if A is orthogonal. This is not too surprising. What might not be expected is that all isometries are of this form.

Theorem. Every isometry of $f : \mathbb{R}^n \to \mathbb{R}^n$ is of the form

$$f(\mathbf{x}) = A\mathbf{x} + \mathbf{b}.$$

for A orthogonal and $\mathbf{b} \in \mathbb{R}^n$.

Proof. Let f be an isometry. Let $\mathbf{e}_1, \cdots, \mathbf{e}_n$ be the standard basis of $\mathbb{R}^n$. Let

$$\mathbf{b} = f(\mathbf{0}), \quad \mathbf{a}_i = f(\mathbf{e}_i) - \mathbf{b}.$$

The idea is to construct our matrix A out of these $\mathbf{a}_i$. For A to be orthogonal, $\{\mathbf{a}_i\}$ must be an orthonormal basis.

Indeed, we can compute

$$\|\mathbf{a}_i\| = \|\mathbf{f}(\mathbf{e}_i) - f(\mathbf{0})\| = d(f(\mathbf{e}_i), f(\mathbf{0})) = d(\mathbf{e}_i, \mathbf{0}) = \|\mathbf{e}_i\| = 1.$$

For $i \neq j$, we have

$$\begin{aligned}(\mathbf{a}_i, \mathbf{a}_j) &= -(\mathbf{a}_i, -\mathbf{a}_j)\\ &= -\frac{1}{2}(\|\mathbf{a}_i - \mathbf{a}_j\|^2 - \|\mathbf{a}_i\|^2 - \|\mathbf{a}_j\|^2)\\ &= -\frac{1}{2}(\|f(\mathbf{e}_i) - f(\mathbf{e}_j)\|^2 - 2)\\ &= -\frac{1}{2}(\|\mathbf{e}_i - \mathbf{e}_j\|^2 - 2)\\ &= 0\end{aligned}$$

So $\mathbf{a}_i$ and $\mathbf{a}_j$ are orthogonal. In other words, $\{\mathbf{a}_i\}$ forms an orthonormal set. It is an easy result that any orthogonal set must be linearly independent. Since we have found n orthonormal vectors, they form an orthonormal basis.

Hence, the matrix A with columns given by the column vectors $\mathbf{a}_i$ is an orthogonal matrix. We define a new isometry

$$g(\mathbf{x}) = A\mathbf{x} + \mathbf{b}.$$

We want to show $f = g$. By construction, we know $g(\mathbf{x}) = f(\mathbf{x})$ is true for $\mathbf{x} = \mathbf{0}, \mathbf{e}_1, \cdots, \mathbf{e}_n$.

We observe that g is invertible. In particular,

$$g^{-1}(\mathbf{x}) = A^{-1}(\mathbf{x} - \mathbf{b}) = A^T\mathbf{x} - A^T\mathbf{b}.$$

Moreover, it is an isometry, since A^T is orthogonal (or we can appeal to the more general fact that inverses of isometries are isometries).

We define

$$h = g^{-1} \circ f.$$

Since it is a composition of isometries, it is also an isometry. Moreover, it fixes $\mathbf{x} = \mathbf{0}, \mathbf{e}_1, \cdots, \mathbf{e}_n$.

It currently suffices to prove that h is the identity.

Let $\mathbf{x} \in \mathbb{R}^n$, and expand it in the basis as

$$\mathbf{x} = \sum_{i=1}^n x_i \mathbf{e}_i.$$

Let

$$\mathbf{y} = h(\mathbf{x}) = \sum_{i=1}^n y_i \mathbf{e}_i.$$

We can compute

$$\begin{aligned} d(\mathbf{x}, \mathbf{e}_i)^2 &= (\mathbf{x} - \mathbf{e}_i, \mathbf{x} - \mathbf{e}_i) = \|\mathbf{x}\|^2 + 1 - 2x_i \\ d(\mathbf{x}, \mathbf{0})^2 &= \|\mathbf{x}\|^2. \end{aligned}$$

Similarly, we have

$$\begin{aligned} d(\mathbf{y}, \mathbf{e}_i)^2 &= (\mathbf{y} - \mathbf{e}_i, \mathbf{y} - \mathbf{e}_i) = \|\mathbf{y}\|^2 + 1 - 2y_i \\ d(\mathbf{y}, \mathbf{0})^2 &= \|\mathbf{y}\|^2. \end{aligned}$$

Since h is an isometry and fixes $\mathbf{0}, \mathbf{e}_1, \cdots, \mathbf{e}_n$, and by definition $h(\mathbf{x}) = \mathbf{y}$, we must have

$$d(\mathbf{x}, \mathbf{0}) = d(\mathbf{y}, \mathbf{0}), \quad d(\mathbf{x}, \mathbf{e}_i) = d(\mathbf{y}, \mathbf{e}_i).$$

The first equality gives $\|\mathbf{x}\|^2 = \|\mathbf{y}\|^2$, and the others then imply $x_i = y_i$ for all i. In other words, $\mathbf{x} = \mathbf{y} = h(\mathbf{x})$. So h is the identity. □

We now collect all our isometries into a group, for convenience.

Definition (Isometry group)**.** The *isometry group* $\operatorname{Isom}(\mathbb{R}^n)$ is the group of all isometries of $\mathbb{R}^n$, which is a group by composition.

Example (Reflections in an affine hyperplane)**.** Let $H \subseteq \mathbb{R}^n$ be an affine hyperplane given by

$$H = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{u} \cdot \mathbf{x} = c\},$$

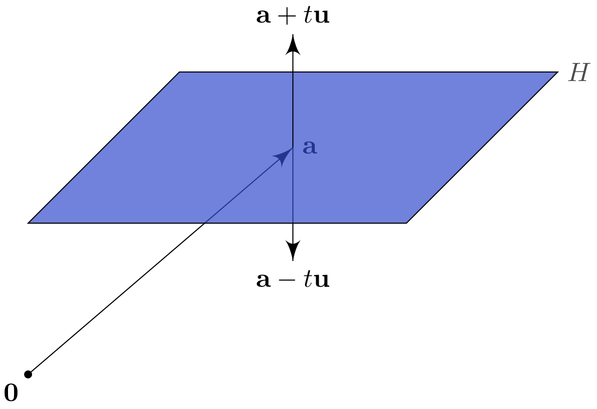
where $\|\mathbf{u}\| = 1$ and $c \in \mathbb{R}$. This is just a natural generalization of a 2-dimensional plane in $\mathbb{R}^3$. Note that unlike a vector subspace, it does not have to contain the origin (since the origin is not a special point).

Reflection in H, written R_H, is the map

$$\begin{aligned} R_H : \mathbb{R}^n &\to \mathbb{R}^n \\ \mathbf{x} &\mapsto \mathbf{x} - 2(\mathbf{x}\cdot\mathbf{u} - c)\mathbf{u} \end{aligned}$$

It is an exercise in the example sheet to show that this is indeed an isometry.

We now check this is indeed what we think a reflection should be. Note that every point in $\mathbb{R}^n$ can be written as $\mathbf{a} + t\mathbf{u}$, where $\mathbf{a} \in H$. Then the reflection should send this point to $\mathbf{a} - t\mathbf{u}$.



This is a routine check:

$$R_H(\mathbf{a} + t\mathbf{u}) = (\mathbf{a} + t\mathbf{u}) - 2t\mathbf{u} = \mathbf{a} - t\mathbf{u}.$$

In particular, we know R_H fixes exactly the points of H.

The converse is also true — any isometry $S \in \operatorname{Isom}(\mathbb{R}^n)$ that fixes the points in some affine hyperplane H is either the identity or R_H.

To show this, we first want to translate the plane such that it becomes a vector subspace. Then we can use our linear algebra magic. For any $\mathbf{a} \in \mathbb{R}^n$, we can define the translation by $\mathbf{a}$ as

$$T_{\mathbf{a}}(\mathbf{x}) = \mathbf{x} + \mathbf{a}.$$

This is clearly an isometry.

We pick an arbitrary $\mathbf{a} \in H$, and let $R = T_{-\mathbf{a}} S T_{\mathbf{a}} \in \operatorname{Isom}(\mathbb{R}^n)$. Then R fixes exactly $H' = T_{-\mathbf{a}} H$. Since $\mathbf{0} \in H'$, H' is a vector subspace. In particular, if $H = \{\mathbf{x} : \mathbf{x}\cdot\mathbf{u} = c\}$, then by putting $c = \mathbf{a}\cdot\mathbf{u}$, we find

$$H' = \{\mathbf{x} : \mathbf{x}\cdot\mathbf{u} = 0\}.$$

To understand R, we already know it fixes everything in H'. So we want to see what it does to $\mathbf{u}$. Note that since R is an isometry and fixes the origin, it is in fact an orthogonal map. Hence for any $\mathbf{x} \in H'$, we get

$$(R\mathbf{u}, \mathbf{x}) = (R\mathbf{u}, R\mathbf{x}) = (\mathbf{u}, \mathbf{x}) = 0.$$

So $R\mathbf{u}$ is also perpendicular to H'. Hence $R\mathbf{u} = \lambda\mathbf{u}$ for some λ. Since R is an isometry, we have $\|R\mathbf{u}\|^2 = 1$. Hence $|\lambda|^2 = 1$, and thus $\lambda = \pm 1$. So either

$\lambda = 1$, and $R = \mathrm{id}$; or $\lambda = -1$, and $R = R_{H'}$, as we already know for orthogonal matrices.

It thus follow that $S = \mathrm{id}_{\mathbb{R}^n}$, or S is the reflection in H.

Thus we find that each reflection R_H is the (unique) isometry fixing H but not $\mathrm{id}_{\mathbb{R}^n}$.

It is an exercise in the example sheet to show that every isometry of $\mathbb{R}^n$ is a composition of at most $n+1$ reflections. If the isometry fixes 0, then n reflections will suffice.

Consider the subgroup of $\mathrm{Isom}(\mathbb{R}^n)$ that fixes $\mathbf{0}$. By our general expression for the general isometry, we know this is the set $\{f(\mathbf{x}) = A\mathbf{x} : AA^T = I\} \cong \mathrm{O}(n)$, the orthogonal group.

For each $A \in \mathrm{O}(n)$, we must have $\det(A)^2 = 1$. So $\det A = \pm 1$. We use this to define a further subgroup, the special orthogonal group.

Definition (Special orthogonal group)**.** The *special orthogonal group* is the group

$$\mathrm{SO}(n) = \{A \in \mathrm{O}(n) : \det A = 1\}.$$

We can look at these explicitly for low dimensions.

Example. Consider

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{O}(2)$$

Orthogonality then requires

$$a^2 + c^2 = b^2 + d^2 = 1, \quad ab + cd = 0.$$

Now we pick $0 \leq \theta, \varphi \leq 2\pi$ such that

$$\begin{aligned} a &= \cos\theta & b &= -\sin\varphi \\ c &= \sin\theta & d &= \cos\varphi. \end{aligned}$$

Then $ab + cd = 0$ gives $\tan\theta = \tan\varphi$ (if $\cos\theta$ and $\cos\varphi$ are zero, we formally say these are both infinity). So either $\theta = \varphi$ or $\theta = \varphi \pm \pi$. Thus we have

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \text{ or } A = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$$

respectively. In the first case, this is a rotation through θ about the origin. This is determinant 1, and hence $A \in \mathrm{SO}(2)$.

In the second case, this is a reflection in the line ℓ at angle $\frac{\theta}{2}$ to the x-axis. Then $\det A = -1$ and $A \not\in \mathrm{SO}(2)$.

So in two dimensions, the orthogonal matrices are either reflections or rotations — those in SO(2) are rotations, and the others are reflections.

Before we can move on to three dimensions, we need to have the notion of orientation. We might intuitively know what an orientation is, but it is rather difficult to define orientation formally. The best we can do is to tell whether two given bases of a vector space have "the same orientation". Thus, it would make sense to define an orientation as an equivalence class of bases of "the same orientation". We formally define it as follows:

Definition (Orientation)**.** An *orientation* of a vector space is an equivalence class of bases — let $\mathbf{v}_1, \cdots, \mathbf{v}_n$ and $\mathbf{v}'_1, \cdots, \mathbf{v}'_n$ be two bases and A be the change of basis matrix. We say the two bases are equivalent iff $\det A > 0$. This is an equivalence relation on the bases, and the equivalence classes are the orientations.

Definition (Orientation-preserving isometry)**.** An isometry $f(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ is *orientation-preserving* if $\det A = 1$. Otherwise, if $\det A = -1$, we say it is *orientation-reversing.*

Example. We now want to look at $\mathrm{O}(3)$. First focus on the case where $A \in \mathrm{SO}(3)$, i.e. $\det A = 1$. Then we can compute

$$\det(A - I) = \det(A^T - I) = \det(A)\det(A^T - I) = \det(I - A) = -\det(A - I).$$

So $\det(A - I) = 0$, i.e. $+1$ is an eigenvalue in $\mathbb{R}$. So there is some $\mathbf{v}_1 \in \mathbb{R}^3$ such that $A\mathbf{v}_1 = \mathbf{v}_1$.

We set $W = \langle \mathbf{v}_1 \rangle^\perp$. Let $\mathbf{w} \in W$. Then we can compute

$$(A\mathbf{w}, \mathbf{v}_1) = (A\mathbf{w}, A\mathbf{v}_1) = (\mathbf{w}, \mathbf{v}_1) = 0.$$

So $A\mathbf{w} \in W$. In other words, W is fixed by A, and $A|_W : W \to W$ is well-defined. Moreover, it is still orthogonal and has determinant 1. So it is a rotation of the two-dimensional vector space W.

We choose $\{\mathbf{v}_2, \mathbf{v}_3\}$ an orthonormal basis of W. Then under the bases $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$, A is represented by

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}.$$

This is the most general orientation-preserving isometry of $\mathbb{R}^3$ that fixes the origin.

How about the orientation-reversing ones? Suppose $\det A = -1$. Then $\det(-A) = 1$. So in some orthonormal basis, we can express A as

$$-A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}.$$

So A takes the form

$$A = \begin{pmatrix} -1 & 0 & 0 \\ 0 & \cos\varphi & -\sin\varphi \\ 0 & \sin\varphi & \cos\varphi \end{pmatrix},$$

where $\varphi = \theta + \pi$. This is a rotated reflection, i.e. we first do a reflection, then rotation. In the special case where $\varphi = 0$, this is a pure reflection.

That's all we're going to talk about isometries.

1.2 Curves in $\mathbb{R}^n$

Next we will briefly look at curves in $\mathbb{R}^n$. This will be very brief, since curves in $\mathbb{R}^n$ aren't too interesting. The most interesting fact might be that the shortest curve between two points is a straight line, but we aren't even proving that, because it is left for the example sheet.

Definition (Curve)**.** A *curve* Γ in $\mathbb{R}^n$ is a continuous map $\Gamma : [a, b] \to \mathbb{R}^n$.

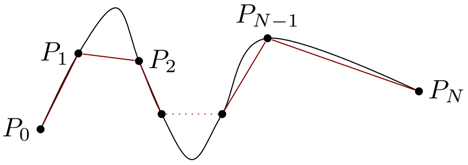
Here we can think of the curve as the trajectory of a particle moving through time. Our main objective of this section is to define the length of a curve. We might want to define the length as

$$\int_a^b \|\Gamma'(t)\| \, \mathrm{d}t,$$

as is familiar from, say, IA Vector Calculus. However, we can't do this, since our definition of a curve does not require Γ to be continuously differentiable. It is merely required to be continuous. Hence we have to define the length in a more roundabout way.

Similar to the definition of the Riemann integral, we consider a dissection $\mathcal{D} = a = t_0 < t_1 < \cdots < t_N = b$ of $[a, b]$, and set $P_i = \Gamma(t_i)$. We define

$$S_{\mathcal{D}} = \sum_i \|\overrightarrow{P_i P_{i+1}}\|.$$



Notice that if we add more points to the dissection, then $S_{\mathcal{D}}$ will necessarily increase, by the triangle inequality. So it makes sense to define the length as the following supremum:

Definition (Length of curve)**.** The length of a curve $\Gamma : [a, b] \to \mathbb{R}^n$ is

$$\ell = \sup_{\mathcal{D}} S_{\mathcal{D}},$$

if the supremum exists.

Alternatively, if we let

$$\operatorname{mesh}(\mathcal{D}) = \max_i (t_i - t_{i-1}),$$

then if ℓ exists, then we have

$$\ell = \lim_{\operatorname{mesh}(\mathcal{D}) \to 0} s_{\mathcal{D}}.$$

Note also that by definition, we can write

$$\ell = \inf\{\tilde{\ell} : \tilde{\ell} \geq S_{\mathcal{D}} \text{ for all } \mathcal{D}\}.$$

The definition by itself isn't too helpful, since there is no nice and easy way to check if the supremum exists. However, differentiability allows us to compute this easily in the expected way.

Proposition. If Γ is continuously differentiable (i.e. C^1), then the length of Γ is given by

$$\operatorname{length}(\Gamma) = \int_a^b \|\Gamma'(t)\| \, \mathrm{d}t.$$

The proof is just a careful check that the definition of the integral coincides with the definition of length.

Proof. To simplify notation, we assume $n = 3$. However, the proof works for all possible dimensions. We write

$$\Gamma(t) = (f_1(t), f_2(t), f_3(t)).$$

For every $s \neq t \in [a, b]$, the mean value theorem tells us

$$\frac{f_i(t) - f_i(s)}{t - s} = f_i'(\xi_i)$$

for some $\xi_i \in (s, t)$, for all $i = 1, 2, 3$.

Now note that f_i' are continuous on a closed, bounded interval, and hence uniformly continuous. For all $\varepsilon \in 0$, there is some $\delta > 0$ such that $|t - s| < \delta$ implies

$$|f_i'(\xi_i) - f'(\xi)| < \frac{\varepsilon}{3}$$

for all $\xi \in (s, t)$. Thus, for any $\xi \in (s, t)$, we have

$$\left\|\frac{\Gamma(t) - \Gamma(s)}{t - s} - \Gamma'(\xi)\right\| = \left\|\begin{pmatrix} f_1'(\xi_1) \\ f_2'(\xi_2) \\ f_3'(\xi_3)\end{pmatrix} - \begin{pmatrix} f_1'(\xi) \\ f_2'(\xi) \\ f_3'(\xi)\end{pmatrix}\right\| \leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

In other words,

$$\|\Gamma(t) - \Gamma(s) - (t - s)\Gamma'(\xi)\| \leq \varepsilon(t - s).$$

We relabel $t = t_i$, $s = t_{i-1}$ and $\xi = \frac{s+t}{2}$.

Using the triangle inequality, we have

$$\begin{aligned}(t_i - t_{i-1})\left\|\Gamma'\left(\frac{t_i + t_{i-1}}{2}\right)\right\| - \varepsilon(t_i - t_{i-1}) &< \|\Gamma(t_i) - \Gamma(t_{i-1})\| \\ &< (t_i - t_{i-1})\left\|\Gamma'\left(\frac{t_i + t_{i-1}}{2}\right)\right\| + \varepsilon(t_i - t_{i-1}).\end{aligned}$$

Summing over all i, we obtain

$$\begin{aligned}\sum_i (t_i - t_{i-1})\left\|\Gamma'\left(\frac{t_i + t_{i-1}}{2}\right)\right\| - \varepsilon(b - a) &< S_{\mathcal{D}} \\ &< \sum_i (t_i - t_{i-1})\left\|\Gamma'\left(\frac{t_i + t_{i-1}}{2}\right)\right\| + \varepsilon(b - a),\end{aligned}$$

which is valid whenever $\operatorname{mesh}(\mathcal{D}) < \delta$.

Since Γ' is continuous, and hence integrable, we know

$$\sum_i (t_i - t_{i-1})\left\|\Gamma'\left(\frac{t_i + t_{i-1}}{2}\right)\right\| \to \int_a^b \|\Gamma'(t)\|\;\mathrm{d}t$$

as $\operatorname{mesh}(\mathcal{D}) \to 0$, and

$$\operatorname{length}(\Gamma) = \lim_{\operatorname{mesh}(\mathcal{D}) \to 0} S_{\mathcal{D}} = \int_a^b \|\Gamma'(t)\|\;\mathrm{d}t. \qquad \square$$

This is all we are going to say about Euclidean space.

2 Spherical geometry

The next thing we are going to study is the geometry on the surface of a sphere. This is a rather sensible thing to study, since it so happens that we all live on something (approximately) spherical. It turns out the geometry of the sphere is very different from that of $\mathbb{R}^2$.

In this section, we will always think of S^2 as a subset of $\mathbb{R}^3$ so that we can reuse what we know about $\mathbb{R}^3$.

Notation. We write $S = S^2 \subseteq \mathbb{R}^3$ for the unit sphere. We write $O = \mathbf{0}$ for the origin, which is the center of the sphere (and not on the sphere).

When we live on the sphere, we can no longer use regular lines in $\mathbb{R}^3$, since these do not lie fully on the sphere. Instead, we have a concept of a *spherical line*, also known as a *great circle*.

Definition (Great circle)**.** A *great circle* (in S^2) is $S^2 \cap$ (a plane through O). We also call these *(spherical) lines*.

We will also call these *geodesics*, which is a much more general term defined on any surface, and happens to be these great circles in S^2.

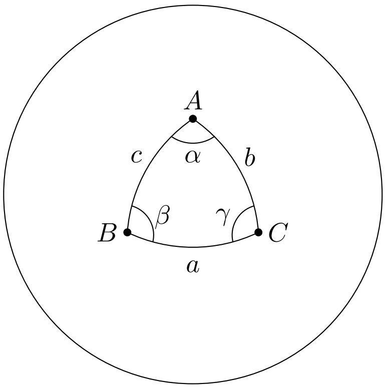
In $\mathbb{R}^3$, we know that any three points that are not colinear determine a unique plane through them. Hence given any two non-antipodal points $P, Q \in S$, there exists a unique spherical line through P and Q.

Definition (Distance on a sphere)**.** Given $P, Q \in S$, the *distance* $d(P, Q)$ is the shorter of the two (spherical) line segments (i.e. arcs) PQ along the respective great circle. When P and Q are antipodal, there are infinitely many line segments between them of the same length, and the distance is π.

Note that by the definition of the radian, $d(P, Q)$ is the angle between $\overrightarrow{OP}$ and $\overrightarrow{OQ}$, which is also $\cos^{-1}(\mathbf{P} \cdot \mathbf{Q})$ (where $\mathbf{P} = \overline{OP}$, $\mathbf{Q} = \overline{OQ}$).

2.1 Triangles on a sphere

One main object of study is spherical triangles – they are defined just like Euclidean triangles, with AB, AC, BC line segments on S of length $< \pi$. The restriction of length is just merely for convenience.



We will take advantage of the fact that the sphere sits in $\mathbb{R}^3$. We set

$$\begin{aligned}\mathbf{n}_1 &= \frac{\mathbf{C}\times\mathbf{B}}{\sin a}\\ \mathbf{n}_2 &= \frac{\mathbf{A}\times\mathbf{C}}{\sin b}\\ \mathbf{n}_3 &= \frac{\mathbf{B}\times\mathbf{A}}{\sin c}.\end{aligned}$$

These are unit normals to the planes OBC, OAC and OAB respectively. They are pointing out of the solid $OABC$.

The angles α, β, γ are the angles between the planes for the respective sides. Then $0 < \alpha, \beta, \gamma < \pi$. Note that the angle between $\mathbf{n}_2$ and $\mathbf{n}_3$ is $\pi + \alpha$ (not α itself — if $\alpha = 0$, then the angle between the two normals is π). So

$$\begin{aligned}\mathbf{n}_2\cdot\mathbf{n}_3 &= -\cos\alpha\\ \mathbf{n}_3\cdot\mathbf{n}_1 &= -\cos\beta\\ \mathbf{n}_1\cdot\mathbf{n}_2 &= -\cos\gamma.\end{aligned}$$

We have the following theorem:

Theorem (Spherical cosine rule)**.**

$$\sin a\sin b\cos\gamma = \cos c - \cos a\cos b.$$

Proof. We use the fact from IA Vectors and Matrices that

$$(\mathbf{C}\times\mathbf{B})\cdot(\mathbf{A}\times\mathbf{C}) = (\mathbf{A}\cdot\mathbf{C})(\mathbf{B}\cdot\mathbf{C}) - (\mathbf{C}\cdot\mathbf{C})(\mathbf{B}\cdot\mathbf{A}),$$

which follows easily from the double-epsilon identity

$$\varepsilon_{ijk}\varepsilon_{imn} = \delta_{jm}\delta_{kn} - \delta_{jn}\delta_{km}.$$

In our case, since $\mathbf{C}\cdot\mathbf{C} = 1$, the right hand side is

$$(\mathbf{A}\cdot\mathbf{C})(\mathbf{B}\cdot\mathbf{C}) - (\mathbf{B}\cdot\mathbf{A}).$$

Thus we have

$$\begin{aligned}-\cos\gamma &= \mathbf{n}_1\cdot\mathbf{n}_2\\ &= \frac{\mathbf{C}\times\mathbf{B}}{\sin a}\cdot\frac{\mathbf{A}\times\mathbf{C}}{\sin b}\\ &= \frac{(\mathbf{A}\cdot\mathbf{C})(\mathbf{B}\cdot\mathbf{C}) - (\mathbf{B}\cdot\mathbf{A})}{\sin a\sin b}\\ &= \frac{\cos b\cos a - \cos c}{\sin a\sin b}.\end{aligned}$$

□

Corollary (Pythagoras theorem)**.** If $\gamma = \frac{\pi}{2}$, then

$$\cos c = \cos a\cos b.$$

Analogously, we have a spherical sine rule.

Theorem (Spherical sine rule).

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

Proof. We use the fact that

$$(\mathbf{A} \times \mathbf{C}) \times (\mathbf{C} \times \mathbf{B}) = (\mathbf{C} \cdot (\mathbf{B} \times \mathbf{A}))\mathbf{C},$$

which we again are not bothered to prove again. The left hand side is

$$-(\mathbf{n}_1 \times \mathbf{n}_2) \sin a \sin b$$

Since the angle between $\mathbf{n}_1$ and $\mathbf{n}_2$ is $\pi + \gamma$, we know $\mathbf{n}_1 \times \mathbf{n}_2 = \mathbf{C} \sin \gamma$. Thus the left hand side is

$$-\mathbf{C} \sin a \sin b \sin \gamma.$$

Thus we know

$$\mathbf{C} \cdot (\mathbf{A} \times \mathbf{B}) = \sin a \sin b \sin \gamma.$$

However, since the scalar triple product is cyclic, we know

$$\mathbf{C} \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}).$$

In other words, we have

$$\sin a \sin b \sin \gamma = \sin b \sin c \sin \alpha.$$

Thus we have

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

Similarly, we know this is equal to $\frac{\sin \beta}{\sin b}$. □

Recall that for small a, b, c, we know

$$\sin a = a + O(a^3).$$

Similarly,

$$\cos a = 1 - \frac{a^2}{2} + O(a^4).$$

As we take the limit $a, b, c \to 0$, the spherical sine and cosine rules become the usual Euclidean versions. For example, the cosine rule becomes

$$ab \cos \gamma = 1 - \frac{c^2}{2} - \left(1 - \frac{a^2}{2}\right)\left(1 - \frac{b^2}{2}\right) + O(\|(a, b, c)\|^3).$$

Rearranging gives

$$c^2 = a^2 + b^2 - 2ab \cos \gamma + O(\|(a, b, c,)\|^3).$$

The sine rule transforms similarly as well. This is what we would expect, since making a, b, c small is equivalent to zooming into the surface of the sphere, and it looks more and more like flat space.

Note that if $\gamma = \pi$, it then follows that C is in the line segment given by AB. So $c = a + b$. Otherwise, we get

$$\cos c > \cos a \cos b - \sin a \sin b = \cos(a + b),$$

since $\cos \gamma < 1$. Since cos is decreasing on $[0, \pi]$, we know

$$c < a + b.$$

Corollary (Triangle inequality)**.** For any $P, Q, R \in S^2$, we have

$$d(P, Q) + d(Q, R) \geq d(P, R),$$

with equality if and only if Q lies is in the line segment PR of shortest length.

Proof. The only case left to check is if $d(P, R) = \pi$, since we do not allow our triangles to have side length π. But in this case they are antipodal points, and any Q lies in a line through PR, and equality holds. □

Thus, we find that (S^2, d) is a metric space.

On $\mathbb{R}^n$, straight lines are curves that minimize distance. Since we are calling spherical lines *lines*, we would expect them to minimize distance as well. This is in fact true.

Proposition. Given a curve Γ on $S^2 \subseteq \mathbb{R}^3$ from P to Q, we have $\ell = \text{length}(\Gamma) \geq d(P, Q)$. Moreover, if $\ell = d(P, Q)$, then the image of Γ is a spherical line segment PQ.

Proof. Let $\Gamma : [0, 1] \to S$ and $\ell = \text{length}(\Gamma)$. Then for any dissection $\mathcal{D}$ of $[0, 1]$, say $0 = t_0 < \cdots < t_N = 1$, write $P_i = \Gamma(t_i)$. We define

$$\tilde{S}_{\mathcal{D}} = \sum_i d(P_{i-1}, P_i) > S_{\mathcal{D}} = \sum_i |\overrightarrow{P_{i-1}P_i}|,$$

where the length in the right hand expression is the distance in Euclidean 3-space.

Now suppose $\ell < d(P, Q)$. Then there is some $\varepsilon > 0$ such that $\ell(1 + \varepsilon) < d(P, Q)$.

Recall from basic trigonometric that if $\theta > 0$, then $\sin\theta < \theta$. Also,

$$\frac{\sin\theta}{\theta} \to 1 \text{ as } \theta \to 0.$$

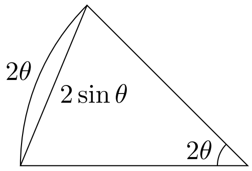
Thus we have

$$\theta \leq (1 + \varepsilon)\sin\theta.$$

for small θ. What we really want is the double of this:

$$2\theta \leq (1 + \varepsilon)2\sin\theta.$$

This is useful since these lengths appear in the following diagram:



This means for P, Q sufficiently close, we have $d(P, Q) \leq (1 + \varepsilon)|\overrightarrow{PQ}|$.

From Analysis II, we know Γ is uniformly continuous on $[0, 1]$. So we can choose $\mathcal{D}$ such that

$$d(P_{i-1}, P_i) \leq (1 + \varepsilon)|\overrightarrow{P_{i-1}P_i}|$$

for all i. So we know that for sufficiently fine $\mathcal{D}$,

$$\tilde{S}_{\mathcal{D}} \leq (1 + \varepsilon)S_{\mathcal{D}} < d(P, Q),$$

since $S_{\mathcal{D}} \to \ell$. However, by the triangle inequality $\tilde{S}_{\mathcal{D}} \geq d(P, Q)$. This is a contradiction. Hence we must have $\ell \geq d(P, Q)$.

Suppose now $\ell = d(P, Q)$ for some $\Gamma : [0, 1] \to S$, $\ell = \text{length}(\Gamma)$. Then for every $t \in [0, 1]$, we have

$$\begin{aligned} d(P, Q) = \ell &= \text{length}\,\Gamma|_{[0,t]} + \text{length}\,\Gamma|_{[t,1]} \\ &\geq d(P, \Gamma(t)) + d(\Gamma(t), Q) \\ &\geq d(P, Q). \end{aligned}$$

Hence we must have equality all along the way, i.e.

$$d(P, Q) = d(P, \Gamma(t)) + d(\Gamma(t), Q)$$

for all $\Gamma(t)$.

However, this is possible only if $\Gamma(t)$ lies on the shorter spherical line segment PQ, as we have previously proved. So done. □

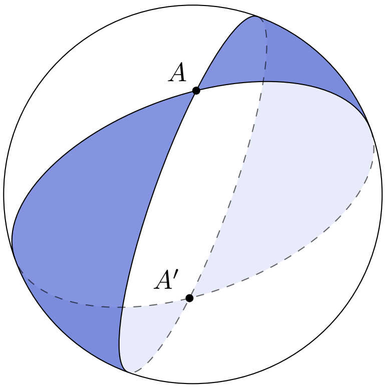
Note that if Γ is a curve of minimal length from P to Q, then Γ is a spherical line segment. Further, from the proof of this proposition, we know $\text{length}\,\Gamma|_{[0,t]} = d(P, \Gamma(t))$ for all t. So the parametrisation of Γ is monotonic. Such a Γ is called a *minimizing geodesic.*

Finally, we get to an important theorem whose prove involves complicated pictures. This is known as the *Gauss-Bonnet theorem.* The Gauss-Bonnet theorem is in fact a much more general theorem. However, here we will specialize in the special case of the sphere. Later, when doing hyperbolic geometry, we will prove the hyperbolic version of the Gauss-Bonnet theorem. Near the end of the course, when we have developed sufficient machinery, we would be able to *state* the Gauss-Bonnet theorem in its full glory. However, we will not be able to prove the general version.

Proposition (Gauss-Bonnet theorem for S^2)**.** If Δ is a spherical triangle with angles α, β, γ, then

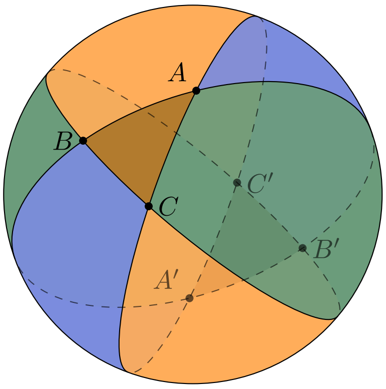
$$\text{area}(\Delta) = (\alpha + \beta + \gamma) - \pi.$$

Proof. We start with the concept of a double lune. A *double lune* with angle $0 < \alpha < \pi$ is two regions S cut out by two planes through a pair of antipodal points, where α is the angle between the two planes.



It is not hard to show that the area of a double lune is 4α, since the area of the sphere is 4π.

Now note that our triangle $\Delta = ABC$ is the intersection of 3 *single* lunes, with each of A, B, C as the pole (in fact we only need two, but it is more convenient to talk about 3).



Therefore Δ together with its antipodal partner Δ' is a subset of each of the 3 double lunes with areas $4\alpha, 4\beta, 4\gamma$. Also, the union of all the double lunes cover the whole sphere, and overlap at exactly Δ and Δ'. Thus

$$4(\alpha+\beta+\gamma) = 4\pi + 2(\mathrm{area}(\Delta) + \mathrm{area}(\Delta')) = 4\pi + 4\,\mathrm{area}(\Delta). \qquad \square$$

This is easily generalized to arbitrary convex n-gons on S^2 (with $n \geq 3$). Suppose M is such a convex n-gon with interior angles $\alpha_1, \cdots, \alpha_n$. Then we have

$$\mathrm{area}(M) = \sum_1^n \alpha_i - (n-2)\pi.$$

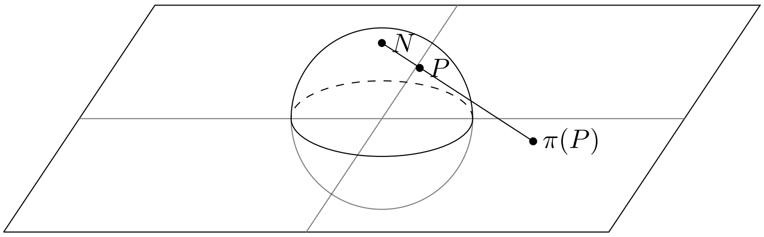
This follows directly from cutting the polygon up into the constituent triangles.

This is very unlike Euclidean space. On $\mathbb{R}^2$, we always have $\alpha + \beta + \gamma = \pi$. Not only is this false on S^2, but by measuring the difference, we can tell the area of the triangle. In fact, we can identify triangles up to congruence just by knowing the three angles.

2.2 Möbius geometry

It turns out it is convenient to identify the sphere S^2 withe the extended complex plane $\mathbb{C}_\infty = \mathbb{C} \cup \{\infty\}$. Then isometries of S^2 will translate to nice class of maps of $\mathbb{C}_\infty$.

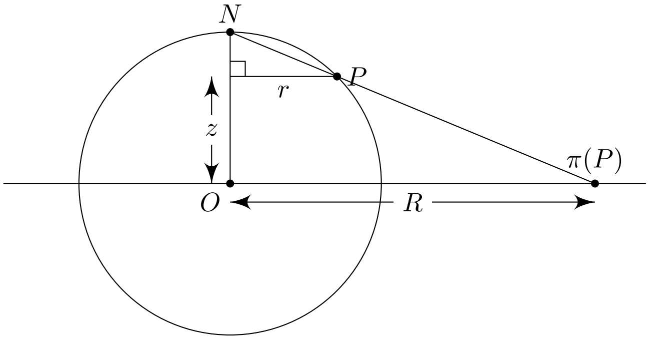
We first find a way to identify S^2 with C_∞. We use coordinates $\zeta \in \mathbb{C}_\infty$. We define the *stereographic projection* $\pi: S^2 \to \mathbb{C}_\infty$ by



$$\pi(P) = (\text{line } PN) \cap \{z = 0\},$$

which is well defined except where $P = N$, in which case we define $\pi(N) = \infty$.

To give an explicit formula for this, consider the cross-section through the plane ONP.



If P has coordinates (x, y), then we see that $\pi(P)$ will be a scalar multiple of $x + iy$. To find this factor, we notice that we have two similar triangles, and hence obtain

$$\frac{r}{R} = \frac{1 - z}{1}.$$

Then we obtain the formula

$$\pi(x, y, z) = \frac{x + iy}{1 - z}.$$

If we do the projection from the South pole instead, we get a related formula.

Lemma. If $\pi' : S^2 \to \mathbb{C}_\infty$ denotes the stereographic projection from the South Pole instead, then

$$\pi'(P) = \frac{1}{\overline{\pi(P)}}.$$

Proof. Let $P(x, y, z)$. Then

$$\pi(x, y, z) = \frac{x + iy}{1 - z}.$$

Then we have

$$\pi'(x, y, z) = \frac{x + iy}{1 + z},$$

since we have just flipped the z axis around. So we have

$$\overline{\pi(P)}\pi'(P) = \frac{x^2 + y^2}{1 - z^2} = 1,$$

noting that we have $x^2 + y^2 + z^2 = 1$ since we are on the unit sphere. □

We can use this to infer that $\pi' \circ \pi^{-1} : C_\infty \to C_\infty$ takes $\zeta \mapsto 1/\bar{\zeta}$, which is the inversion in the unit circle $|\zeta| = 1$.

From IA Groups, we know Möbius transformations act on $\mathbb{C}_\infty$ and form a group G by composition. For each matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}(2, \mathbb{C}),$$

we get a Möbius map $\mathbb{C}_\infty \to \mathbb{C}_\infty$ by

$$\zeta \mapsto \frac{a\zeta + b}{c\zeta + d}.$$

Moreover, composition of Möbius map is the same multiplication of matrices.

This is not exactly a bijective map between G and $\mathrm{GL}(2, \mathbb{C})$. For any $\lambda \in \mathbb{C}^* = \mathbb{C} \setminus \{0\}$, we know λA defines the same map Möbius map as A. Conversely, if A_1 and A_2 gives the same Möbius map, then there is some $\lambda_1 \neq 0$ such that $A_1 = \lambda A_2$.

Hence, we have

$$G \cong \mathrm{PGL}(2, \mathbb{C}) = \mathrm{GL}(2, \mathbb{C})/\mathbb{C}^*,$$

where

$$\mathbb{C}^* \cong \{\lambda I : \lambda \in \mathbb{C}^*\}.$$

Instead of taking the whole $\mathrm{GL}(2, \mathbb{C})$ and quotienting out multiples of the identities, we can instead start with $\mathrm{SL}(2, \mathbb{C})$. Again, $A_1, A_2 \in \mathrm{SL}(2, \mathbb{C})$ define the same map if and only if $A_1 = \lambda A_2$ for some λ. What are the possible values of λ? By definition of the special linear group, we must have

$$1 = \det(\lambda A) = \lambda^2 \det A = \lambda^2.$$

So $\lambda^2 = \pm 1$. So each Möbius map is represented by two matrices, A and $-A$, and we get

$$G \cong \mathrm{PSL}(2, \mathbb{C}) = \mathrm{SL}(2, \mathbb{C})/\{\pm 1\}.$$

Now let's think about the sphere. On S^2, the rotations $\mathrm{SO}(3)$ act as isometries. In fact, the full isometry group of S^2 is $\mathrm{O}(3)$ (the proof is on the first example sheet). What we would like to show that rotations of S^2 correspond to Möbius transformations coming from the subgroup $\mathrm{SU}(2) \leq \mathrm{GL}(2, \mathbb{C})$.

Theorem. Via the stereographic projection, every rotation of S^2 induces a Möbius map defined by a matrix in $\mathrm{SU}(2) \subseteq \mathrm{GL}(2, \mathbb{C})$, where

$$\mathrm{SU}(2) = \left\{ \begin{pmatrix} a & -b \\ \bar{b} & \bar{a} \end{pmatrix} : |a|^2 + |b|^2 = 1 \right\}.$$

Proof.

(i) Consider the $r(\hat{\mathbf{z}}, \theta)$, the rotations about the z axis by θ. These corresponds to the Möbius map $\zeta \mapsto e^{i\theta}\zeta$, which is given by the unitary matrix

$$\begin{pmatrix} e^{i\theta/2} & 0 \\ 0 & e^{-i\theta/2}. \end{pmatrix}$$

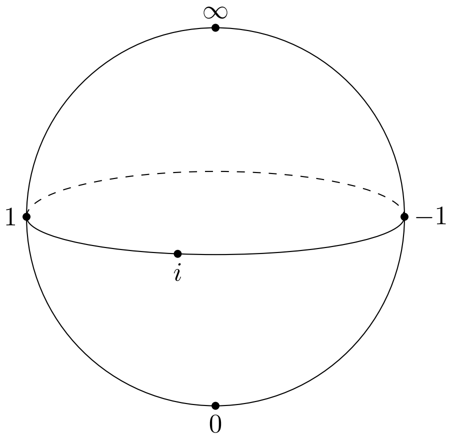
(ii) Consider the rotation $r(\hat{\mathbf{y}}, \frac{\pi}{2})$. This has the matrix

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

This corresponds to the map

$$\zeta = \frac{x + iy}{1 - z} \mapsto \zeta' = \frac{z + iy}{1 + x}$$

We want to show this is a Möbius map. To do so, we guess what the Möbius map should be, and check it works. We can manually compute that $-1 \mapsto \infty$, $1 \mapsto 0$, $i \mapsto i$.



The only Möbius map that does this is

$$\zeta' = \frac{\zeta - 1}{\zeta + 1}.$$

We now check:

$$\begin{aligned}\frac{\zeta - 1}{\zeta + 1} &= \frac{x + iy - 1 + z}{x + iy + 1 - z}\\ &= \frac{x - 1 + z + iy}{x + 1 - (z - iy)}\\ &= \frac{(z + iy)(x - 1 + z + iy)}{(x + 1)(z + iy) - (z^2 + y^2)}\\ &= \frac{(z + iy)(x - 1 + z + iy)}{(x + 1)(z + iy) + (x^2 - 1)}\\ &= \frac{(z + iy)(x - 1 + z + iy)}{(x + 1)(z + iy + x - 1)}\\ &= \frac{z + iy}{x + 1}.\end{aligned}$$

So done. We finally have to write this in the form of an SU(2) matrix:

$$\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

(iii) We claim that SO(3) is generated by $r\left(\hat{\mathbf{y}}, \frac{\pi}{2}\right)$ and $r(\hat{\mathbf{z}}, \theta)$ for $0 \leq \theta < 2\pi$.

To show this, we observe that $r(\hat{\mathbf{x}}, \varphi) = r(\hat{\mathbf{y}}, \frac{\pi}{2})r(\hat{\mathbf{z}}, \varphi)r(\hat{\mathbf{y}}, -\frac{\pi}{2})$. Note that we read the composition from right to left. You can convince yourself this is true by taking a physical sphere and try rotating. To prove it formally, we can just multiply the matrices out.

Next, observe that for $\mathbf{v} \in S^2 \subseteq \mathbb{R}^3$, there are some angles φ, ψ such that $g = r(\hat{\mathbf{z}}, \psi) r(\hat{\mathbf{x}}, \varphi)$ maps $\mathbf{v}$ to $\hat{\mathbf{x}}$. We can do so by first picking $r(\hat{\mathbf{x}}, \varphi)$ to rotate $\mathbf{v}$ into the (x, y)-plane. Then we rotate about the z-axis to send it to $\hat{\mathbf{x}}$.

Then for any θ, we have $r(\mathbf{v}, \theta) = g^{-1} r(\hat{\mathbf{x}}, \theta) g$, and our claim follows by composition.

(iv) Thus, via the stereographic projection, every rotation of S^2 corresponds to products of Möbius transformations of $\mathbb{C}_\infty$ with matrices in SU(2). □

The key of the proof is step (iii). Apart from enabling us to perform the proof, it exemplifies a useful technique in geometry — we know how to rotate arbitrary things in the z axis. When we want to rotate things about the x axis instead, we first rotate the sphere to move the x axis to where the z axis used to be, do those rotations, and then rotate it back. In general, we can use some isometries or rotations to move what we want to do to a convenient location.

Theorem. The group of rotations SO(3) acting on S^2 corresponds precisely with the subgroup $\mathrm{PSU}(2) = \mathrm{SU}(2)/\pm 1$ of Möbius transformations acting on $\mathbb{C}_\infty$.

What this is in some sense a converse of the previous theorem. We are saying that for each Möbius map from SU(2), we can find some rotation of S^2 that induces that Möbius map, and there is exactly one.

Proof. Let $g \in \mathrm{PSU}(2)$ be a Möbius transformation

$$g(z) = \frac{az + b}{\bar{b}z + \bar{a}}.$$

Suppose first that $g(0) = 0$. So $b = 0$. So $a\bar{a} = 1$. Hence $a = e^{i\theta/2}$. Then g corresponds to $r(\hat{\mathbf{z}}, \theta)$, as we have previously seen.

In general, let $g(0) = w \in \mathbb{C}_\infty$. Let $Q \in S^2$ be such that $\pi(Q) = w$. Choose a rotation $A \in \mathrm{SO}(3)$ such that $A(Q) = -\hat{\mathbf{z}}$. Since A is a rotation, let $\alpha \in \mathrm{PSU}(2)$ be the corresponding Möbius transformation. By construction we have $\alpha(w) = 0$. Then the composition $\alpha \circ g$ fixes zero. So it corresponds to some $B = r(z, \theta)$. We then see that g corresponds to $A^{-1}B \in \mathrm{SO}(3)$. So done. □

Again, we solved an easy case, where $g(0) = 0$, and then performed some rotations to transform any other case into this simple case.

We have now produced a 2-to-1 map

$$\mathrm{SU}(2) \to \mathrm{PSU}(2) \cong \mathrm{SO}(3).$$

If we treat these groups as topological spaces, this map does something funny.

Suppose we start with a (non-closed) path from I to $-I$ in SU(2). Applying the map, we get a *closed* loop from I to I in SO(3).

Hence, in SO(3), loops are behave slightly weirdly. If we go around this loop in SO(3), we didn't really get back to the same place. Instead, we have actually moved from I to $-I$ in SU(2). It takes *two* full loops to actually get back to I. In physics, this corresponds to the idea of spinors.

We can also understand this topologically as follows: since SU(2) is defined by two complex points $a, b \in \mathbb{C}$ such that $|a|^2 + |b|^2$, we can view it as as three-sphere S^3 in SO(3).

A nice property of S^3 is it is *simply connected*, as in any loop in S^3 can be shrunk to a point. In other words, given any loop in S^3, we can pull and stretch it (continuously) until becomes the "loop" that just stays at a single point.

On the other hand, SO(3) is not simply connected. We have just constructed a loop in SO(3) by mapping the path from I to $-I$ in SU(2). We *cannot* deform this loop until it just sits at a single point, since if we lift it back up to SU(2), it still has to move from I to $-I$.

The neat thing is that in some sense, $S^3 \cong \mathrm{SU}(2)$ is just "two copies" of SO(3). By duplicating SO(3), we have produced SU(2), a simply connected space. Thus we say SU(2) is a *universal cover* of SO(3).

We've just been waffling about spaces and loops, and throwing around terms we haven't defined properly. These vague notions will be made precise in the IID Algebraic Topology course, and we will then (maybe) see that SU(2) is a *universal cover* of SO(3).

3 Triangulations and the Euler number

We shall now study the idea of triangulations and the Euler number. We aren't going to do much with them in this course — we will not even prove that the Euler number is a well-defined number. However, we need Euler numbers in order to state the full Gauss-Bonnet theorem at the end, and the idea of triangulations is useful in the IID Algebraic Topology course for defining simplicial homology. More importantly, the discussion of these concepts is required by the schedules. Hence we will get *some* exposure to these concepts in this chapter.

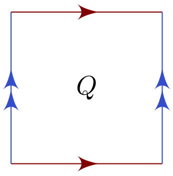
It is convenient to have an example other than the sphere when discussing triangulations and Euler numbers. So we introduce the *torus*

Definition ((Euclidean) torus)**.** The *(Euclidean) torus* is the set $\mathbb{R}^2/\mathbb{Z}^2$ of equivalence classes of $(x, y) \in \mathbb{R}^2$ under the equivalence relation

$$(x_1, y_1) \sim (x_2, y_2) \Leftrightarrow x_1 - x_2, y_1 - y_2 \in \mathbb{Z}.$$

It is easy to see this is indeed an equivalence relation. Thus a point in T represented by (x, y) is a coset $(x, y) + \mathbb{Z}^2$ of the subgroup $\mathbb{Z}^2 \leq \mathbb{R}^2$.

Of course, there are many ways to represent a point in the torus. However, for any closed square $Q \subseteq \mathbb{R}^2$ with side length 1, we can obtain T is by identifying the sides



We can define a distance d for all $P_1, P_2 \in T$ to be

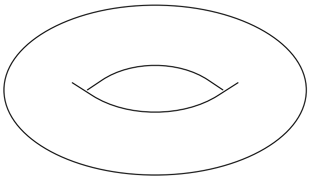
$$d(P_1, P_2) = \min\{\|\mathbf{v}_1 - \mathbf{v}_2\| : \mathbf{v}_i \in \mathbb{R}^2, \mathbf{v}_i + \mathbb{Z}^2 = P_i\}.$$

It is not hard to show this definition makes (T, d) into a metric space. This allows us to talk about things like open sets and continuous functions. We will later show that this is not just a metric space, but a *smooth surface*, after we have defined what that means.

We write $\mathring{Q}$ for the interior of Q. Then the natural map $f : \mathring{Q} \to T$ given by $\mathbf{v} \mapsto \mathbf{v} + \mathbb{Z}^2$ is a bijection onto some open set $U \subseteq T$. In fact, U is just $T \setminus \{\text{two circles meeting in 1 point}\}$, where the two circles are the boundary of the square Q.

Now given any point in the torus represented by $P + \mathbb{Z}^2$, we can find a square Q such that $P \in \mathring{Q}$. Then $f : \mathring{Q} \to T$ restricted to an open disk about P is an isometry (onto the image, a subset of $\mathbb{R}^2$). Thus we say d is a *locally Euclidean metric.*

One can also think of the torus T as the surface of a doughnut, "embedded" in Euclidean space $\mathbb{R}^3$.



Given this, it is natural to define the distance between two points to be the length of the shortest curve between them on the torus. However, this distance function is *not* the same as what we have here. So it is misleading to think of our locally Euclidean torus as a "doughnut".

With one more example in our toolkit, we can start doing what we really want to do. The idea of a triangulation is to cut a space X up into many smaller triangles, since we like triangles. However, we would like to first define what a triangle is.

Definition (Topological triangle)**.** A *topological triangle* on $X = S^2$ or T (or any metric space X) is a subset $R \subseteq X$ equipped with a homeomorphism $R \to \Delta$, where Δ is a closed Euclidean triangle in $\mathbb{R}^2$.

Note that a spherical triangle is in fact a topological triangle, using the radial projection to the plane $\mathbb{R}^2$ from the center of the sphere.

Definition (Topological triangulation)**.** A *topological triangulation* τ of a metric space X is a finite collection of topological triangles of X which cover X such that

(i) For every pair of triangles in τ, either they are disjoint, or they meet in exactly one edge, or meet at exactly one vertex.

(ii) Each edge belongs to exactly two triangles.

These notions are useful only if the space X is "two dimensional" — there is no way we can triangulate, say $\mathbb{R}^3$, or a line. We can generalize triangulation to allow higher dimensional "triangles", namely topological tetrahedrons, and in general, n-simplices, and make an analogous definition of triangulation. However, we will not bother ourselves with this.

Definition (Euler number)**.** The *Euler number* of a triangulation $e = e(X, \tau)$ is

$$e = F - E + V,$$

where F is the number of triangles; E is the number of edges; and V is the number of vertices.

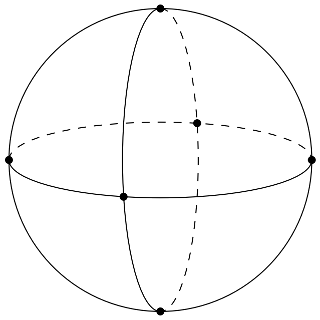
Note that each edge actually belongs to two triangles, but we will only count it once.

There is one important fact about triangulations from algebraic topology, which we will state without proof.

Theorem. The Euler number e is independent of the choice of triangulation.

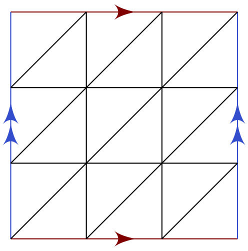
So the Euler number $e = e(X)$ is a property of the space X itself, not a particular triangulation.

Example. Consider the following triangulation of the sphere:



This has 8 faces, 12 edges and 6 vertices. So $e = 2$.

Example. Consider the following triangulation of the torus. Be careful not to double count the edges and vertices at the sides, since the sides are glued together.



This has 18 faces, 27 edges and 9 vertices. So $e = 0$.

In both cases, we did not cut up our space with funny, squiggly lines. Instead, we used "straight" lines. These triangles are known as *geodesic triangles*.

Definition (Geodesic triangle)**.** A *geodesic triangle* is a triangle whose sides are geodesics, i.e. paths of shortest distance between two points.

In particular, we used spherical triangles in S^2 and Euclidean triangles in $\mathring{Q}$. Triangulations made of geodesic triangles are rather nice. They are so nice that we can actually prove something about them!

Proposition. For every geodesic triangulation of S^2 (and respectively T) has $e = 2$ (respectively, $e = 0$).

Of course, we know this is true for *any* triangulation, but it is difficult to prove that without algebraic topology.

Proof. For any triangulation τ, we denote the "faces" of $\Delta_1, \cdots, \Delta_F$, and write $\tau_i = \alpha_i + \beta_i + \gamma_i$ for the sum of the interior angles of the triangles (with $i = 1, \cdots, F$).

Then we have

$$\sum \tau_i = 2\pi V,$$

since the total angle around each vertex is 2π. Also, each triangle has three edges, and each edge is from two triangles. So $3F = 2E$. We write this in a more convenient form:

$$F = 2E - 2F.$$

How we continue depends on whether we are on the sphere or the torus.

- For the sphere, Gauss-Bonnet for the sphere says the area of Δ_i is $\tau_i - \pi$. Since the area of the sphere is 4π, we know

  $$\begin{aligned} 4\pi &= \sum \operatorname{area}(\Delta_i) \\ &= \sum (\tau_i - \pi) \\ &= 2\pi V - F\pi \\ &= 2\pi V - (2E - 2F)\pi \\ &= 2\pi(F - E + V). \end{aligned}$$

  So $F - E + V = 2$.

- For the torus, we have $\tau_i = \pi$ for every face in $\mathring{Q}$. So

  $$2\pi V = \sum \tau_i = \pi F.$$

  So

  $$2V = F = 2E - 2F.$$

  So we get

  $$2(F - V + E) = 0,$$

  as required. □

Note that in the definition of triangulation, we decomposed X into topological triangles. We can also use decompositions by topological polygons, but they are slightly more complicated, since we have to worry about convexity. However, apart from this, everything works out well. In particular, the previous proposition also holds, and we have Euler's formula for S^2: $V - E + F = 2$ for any polygonal decomposition of S^2. This is not hard to prove, and is left as an exercise on the example sheet.

4 Hyperbolic geometry

At the beginning of the course, we studied Euclidean geometry, which was not hard, because we already knew about it. Later on, we studied spherical geometry. That also wasn't too bad, because we can think of S^2 concretely as a subset of $\mathbb{R}^3$.

We are next going to study hyperbolic geometry. Historically, hyperbolic geometry was created when people tried to prove Euclid's parallel postulate (that given a line ℓ and a point $P \not\in \ell$, there exists a unique line ℓ' containing P that does not intersect ℓ). Instead of proving the parallel postulate, they managed to create a new geometry where this is false, and this is hyperbolic geometry.

Unfortunately, hyperbolic geometry is much more complicated, since we cannot directly visualize it as a subset of $\mathbb{R}^3$. Instead, we need to develop the machinery of a *Riemannian metric* in order to properly describe hyperbolic geometry. In a nutshell, this allows us to take a subset of $\mathbb{R}^2$ and measure distances in it in a funny way.

4.1 Review of derivatives and chain rule

We start by reviewing some facts about taking derivatives, and make explicit the notation we will use.

Definition (Smooth function)**.** Let $U \subseteq \mathbb{R}^n$ be open, and $f = (f_1, \cdots, f_m) : U \to \mathbb{R}^m$. We say f is *smooth* (or C^∞) if each f_i has continuous partial derivatives of each order. In particular, a C^∞ map is differentiable, with continuous first-order partial derivatives.

Definition (Derivative)**.** The *derivative* for a function $f : U \to \mathbb{R}^m$ at a point $a \in U$ is a linear map $\mathrm{d}f_a : \mathbb{R}^n \to \mathbb{R}^m$ (also written as $\mathrm{D}f(a)$ or $f'(a)$) such that

$$\lim_{h \to 0} \frac{\|f(a + h) - f(a) - \mathrm{d}f_a \cdot h\|}{\|h\|} \to 0,$$

where $h \in \mathbb{R}^n$.

If $m = 1$, then $\mathrm{d}f_a$ is expressed as $\left(\frac{\partial f}{\partial x_a}(a), \cdots, \frac{\partial f}{\partial x_n}(a)\right)$, and the linear map is given by

$$(h_1, \cdots, h_n) \mapsto \sum_{i=1}^n \frac{\partial f}{\partial x_i}(a) h_i,$$

i.e. the dot product. For a general m, this vector becomes a matrix. The *Jacobian matrix* is

$$J(f)_a = \left(\frac{\partial f_i}{\partial x_j}(a)\right),$$

with the linear map given by matrix multiplication, namely

$$h \mapsto J(f)_a \cdot h.$$

Example. Recall that a holomorphic (analytic) function of complex variables $f : U \subseteq \mathbb{C} \to \mathbb{C}$ has a derivative f', defined by

$$\lim_{|w| \to 0} \frac{|f(z + w) - f(z) - f'(z) w|}{|w|} \to 0$$

We let $f'(z) = a + ib$ and $w = h_1 + ih_2$. Then we have

$$f'(z)w = ah_1 - bh_2 + i(ah_2 + bh_1).$$

We identify $\mathbb{R}^2 = \mathbb{C}$. Then $f : U \subseteq \mathbb{R}^2 \to \mathbb{R}^2$ has a derivative $\mathrm{d}f_z : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}.$$

We're also going to use the chain rule quite a lot. So we shall write it out explicitly.

Proposition (Chain rule)**.** Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^p$. Let $f : U \to \mathbb{R}^m$ and $g : V \to U$ be smooth. Then $f \circ g : V \to \mathbb{R}^m$ is smooth and has a derivative

$$d(f \circ g)_p = (\mathrm{d}f)_{g(p)} \circ (\mathrm{d}g)_p.$$

In terms of the Jacobian matrices, we get

$$J(f \circ g)_p = J(f)_{g(p)} J(g)_p.$$

4.2 Riemannian metrics

Finally, we get to the idea of a Riemannian metric. The basic idea of a Riemannian metric is not too unfamiliar. Presumably, we have all seen maps of the Earth, where we try to draw the spherical Earth on a piece of paper, i.e. a subset of $\mathbb{R}^2$. However, this does not behave like $\mathbb{R}^2$. You cannot measure distances on Earth by placing a ruler on the map, since distances are distorted. Instead, you have to find the coordinates of the points (e.g. the longitude and latitude), and then plug them into some complicated formula. Similarly, straight lines on the map are not really straight (spherical) lines on Earth.

We really should not think of Earth a subset of $\mathbb{R}^2$. All we have done was to "force" Earth to live in $\mathbb{R}^2$ to get a convenient way of depicting the Earth, as well as a convenient system of labelling points (in many map projections, the x and y axes are the longitude and latitude).

This is the idea of a Riemannian metric. To describe some complicated surface, we take a subset U of $\mathbb{R}^2$, and define a new way of measuring distances, angles and areas on U. All these information are packed into an entity known as the *Riemannian metric.*

Definition (Riemannian metric)**.** We use coordinates $(u, v) \in \mathbb{R}^2$. We let $V \subseteq \mathbb{R}^2$ be open. Then a Riemannian metric on V is defined by giving C^∞ functions $E, F, G : V \to \mathbb{R}$ such that

$$\begin{pmatrix} E(P) & F(P) \\ F(P) & G(P) \end{pmatrix}$$

is a positive definite definite matrix for all $P \in V$.

Alternatively, this is a smooth function that gives a 2×2 symmetric positive definite matrix, i.e. inner product $\langle\,\cdot\,,\,\cdot\,\rangle_P$, for each point in V. By definition, if $\mathbf{e}_1, \mathbf{e}_2$ are the standard basis, then

$$\begin{aligned} \langle \mathbf{e}_1, \mathbf{e}_1 \rangle_P &= E(P) \\ \langle \mathbf{e}_1, \mathbf{e}_2 \rangle_P &= F(P) \\ \langle \mathbf{e}_2, \mathbf{e}_2 \rangle_P &= G(P). \end{aligned}$$

Example. We can pick $E = G = 1$ and $F = 0$. Then this is just the standard Euclidean inner product.

As mentioned, we should not imagine V as a subset of $\mathbb{R}^2$. Instead, we should think of it as an abstract two-dimensional surface, with some coordinate system given by a subset of $\mathbb{R}^2$. However, this coordinate system is just a convenient way of labelling points. They do not represent any notion of distance. For example, $(0, 1)$ need not be closer to $(0, 2)$ than to $(7, 0)$. These are just abstract labels.

With this in mind, V does not have any intrinsic notion of distances, angles and areas. However, we do want these notions. We can certainly *write down* things like the difference of two points, or even the compute the derivative of a function. However, these numbers you get are not meaningful, since we can easily use a different coordinate system (e.g. by scaling the axes) and get a different number. They have to be interpreted with the *Riemannian metric*. This tells us how to measure these things, via an inner product "that varies with space". This variation in space is not an oddity arising from us not being able to make up our minds. This is since we have "forced" our space to lie in $\mathbb{R}^2$. Inside V, going from $(0, 1)$ to $(0, 2)$ might be very different from going from $(5, 5)$ to $(6, 5)$, since coordinates don't mean anything. Hence our inner product needs to measure "going from $(0, 1)$ to $(0, 2)$" differently from "going from $(5, 5)$ to $(6, 5)$", and must vary with space.

We'll soon come to defining how this inner product gives rise to the notion of distance and similar stuff. Before that, we want to understand what we can put into the inner product $\langle\,\cdot\,,\,\cdot\,\rangle_P$. Obviously these would be vectors in $\mathbb{R}^2$, but where do these vectors come from? What are they supposed to represent?

The answer is "directions" (more formally, tangent vectors). For example, $\langle \mathbf{e}_1, \mathbf{e}_1\rangle_P$ will tell us how far we actually are going if we move in the direction of $\mathbf{e}_1$ from P. Note that we say "move in the direction of $\mathbf{e}_1$", not "move by $\mathbf{e}_1$". We really should read this as "if we move by $h\mathbf{e}_1$ for some small h, then the distance covered is $h\sqrt{\langle \mathbf{e}_1, \mathbf{e}_1\rangle_P}$". This statement is to be interpreted along the same lines as "if we vary x by some small h, then the value of f will vary by $f'(x)h$". Notice how the inner product allows us to translate a length in $\mathbb{R}^2$ (namely $\|h\mathbf{e}_1\|_{\mathrm{eucl}} = h$) into the actual length in V.

What we needed for this is just the norm induced by the inner product. Since what we have is the whole inner product, we in fact can define more interesting things such as areas and angles. We will formalize these ideas very soon, after getting some more notation out of the way.

Often, instead of specifying the three functions separately, we write the metric as

$$E\ \mathrm{d}u^2 + 2F\ \mathrm{d}u\ \mathrm{d}v + G\ \mathrm{d}v^2.$$

This notation has some mathematical meaning. We can view the coordinates as smooth functions $u : V \to \mathbb{R}$, $v : U \to \mathbb{R}$. Since they are smooth, they have derivatives. They are linear maps

$$\begin{aligned} \mathrm{d}u_P : \mathbb{R}^2 &\to \mathbb{R} & \qquad \mathrm{d}v_P : \mathbb{R}^2 &\to \mathbb{R} \\ (h_1, h_2) &\mapsto h_1 & (h_1, h_2) &\mapsto h_2. \end{aligned}$$

These formula are valid for all $P \in V$. So we just write $\mathrm{d}u$ and $\mathrm{d}v$ instead. Since they are maps $\mathbb{R}^2 \to \mathbb{R}$, we can view them as vectors in the dual space,

$\mathrm{d}u, \mathrm{d}v \in (\mathbb{R}^2)^*$. Moreover, they form a basis for the dual space. In particular, they are the dual basis to the standard basis $\mathbf{e}_1, \mathbf{e}_2$ of $\mathbb{R}^2$.

Then we can consider $\mathrm{d}u^2, \mathrm{d}u\ \mathrm{d}v$ and $\mathrm{d}v^2$ as *bilinear forms* on $\mathbb{R}^2$. For example,

$$\begin{aligned}\mathrm{d}u^2(\mathbf{h},\mathbf{k}) &= \mathrm{d}u(\mathbf{h})\mathrm{d}u(\mathbf{k})\\ \mathrm{d}u\ \mathrm{d}v(\mathbf{h},\mathbf{k}) &= \frac{1}{2}(\mathrm{d}u(\mathbf{h})\mathrm{d}v(\mathbf{k}) + \mathrm{d}u(\mathbf{k})\mathrm{d}v(\mathbf{h}))\\ \mathrm{d}v^2(\mathbf{h},\mathbf{k}) &= \mathrm{d}v(\mathbf{h})\mathrm{d}v(\mathbf{k})\end{aligned}$$

These have matrices

$$\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix},\quad \begin{pmatrix}0 & \frac{1}{2}\\ \frac{1}{2} & 0\end{pmatrix},\quad \begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix}$$

respectively. Then we indeed have

$$E\ \mathrm{d}u^2 + 2F\ \mathrm{d}u\ \mathrm{d}v + G\ \mathrm{d}v^2 = \begin{pmatrix}E & F\\ F & G\end{pmatrix}.$$

We can now start talking about what this is good for. In standard Euclidean space, we have a notion of length and area. A Riemannian metric also gives a notion of length and area.

Definition (Length)**.** The *length* of a smooth curve $\gamma = (\gamma_1, \gamma_2) : [0, 1] \to V$ is defined as

$$\int_0^1 \left(E\dot\gamma_1^2 + 2F\dot\gamma_1\dot\gamma_2 + G\dot\gamma_2{}^2\right)^{\frac{1}{2}}\ \mathrm{d}t,$$

where $E = E(\gamma_1(t), \gamma_2(t))$ etc. We can also write this as

$$\int_0^1 \langle\dot\gamma, \dot\gamma\rangle_{\gamma(t)}^{\frac{1}{2}}\ \mathrm{d}t.$$

Definition (Area)**.** The *area* of a region $W \subseteq V$ is defined as

$$\int_W (EG - F^2)^{\frac{1}{2}}\ \mathrm{d}u\ \mathrm{d}v$$

when this integral exists.

In the area formula, what we are integrating is just the determinant of the metric. This is also known as the Gram determinant.

We define the distance between two points P and Q to be the infimum of the lengths of all curves from P to Q. It is an exercise on the second example sheet to prove that this is indeed a metric.

Example. We will not do this in full detail — the details are to be filled in in the third example sheet.

Let $V = \mathbb{R}^2$, and define the Riemannian metric by

$$\frac{4(\mathrm{d}u^2 + \mathrm{d}v^2)}{(1 + u^2 + v^2)^2}.$$

This looks somewhat arbitrary, but we shall see this actually makes sense by identifying $\mathbb{R}^2$ with the sphere by the stereographic projection $\pi : S^2 \setminus \{N\} \to \mathbb{R}^2$.

For every point $P \in S^2$, the tangent plane to S^2 at P is given by $\{\mathbf{x} \in \mathbb{R}^3 : \mathbf{x} \cdot \overrightarrow{OP} = 0\}$. Note that we translated it so that P is the origin, so that we can view it as a vector space (points on the tangent plane are points "from P"). Now given any two tangent vectors $\mathbf{x}_1, \mathbf{x}_2 \perp \overrightarrow{OP}$, we can take the inner product $\mathbf{x}_1 \cdot \mathbf{x}_2$ in $\mathbb{R}^3$.

We want to say this inner product is "the same as" the inner product provided by the Riemannian metric on $\mathbb{R}^2$. We cannot just require

$$\mathbf{x}_1 \cdot \mathbf{x}_2 = \langle \mathbf{x}_1, \mathbf{x}_2 \rangle_{\pi(P)},$$

since this makes no sense at all. Apart from the obvious problem that $\mathbf{x}_1, \mathbf{x}_2$ have three components but the Riemannian metric takes in vectors of two components, we know that $\mathbf{x}_1$ and $\mathbf{x}_2$ are vectors tangent to $P \in S^2$, but to apply the Riemannian metric, we need the corresponding tangent vector at $\pi(P) \in \mathbb{R}^2$. To do so, we act by $\mathrm{d}\pi_p$. So what we want is

$$\mathbf{x}_1 \cdot \mathbf{x}_2 = \langle \mathrm{d}\pi_P(\mathbf{x}_1), \mathrm{d}\pi_P(\mathbf{x}_2) \rangle_{\pi(P)}.$$

Verification of this equality is left as an exercise on the third example sheet. It is helpful to notice

$$\pi^{-1}(u, v) = \frac{(2u, 2v, u^2 + v^2 - 1)}{1 + u^2 + v^2}.$$

In some sense, we say the surface $S^2 \setminus \{N\}$ is "isometric" to $\mathbb{R}^2$ via the stereographic projection π. We can define the notion of isometry between two open sets with Riemannian metrics in general.

Definition (Isometry)**.** Let $V, \tilde{V} \subseteq \mathbb{R}^2$ be open sets endowed with Riemannian metrics, denoted as $\langle \cdot, \cdot \rangle_P$ and $\langle \cdot, \cdot \rangle_Q^{\sim}$ for $P \in V, Q \in \tilde{V}$ respectively.

A diffeomorphism (i.e. C^∞ map with C^∞ inverse) $\varphi : V \to \tilde{V}$ is an *isometry* if for every $P \in V$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$, we get

$$\langle \mathbf{x}, \mathbf{y} \rangle_P = \langle \mathrm{d}\varphi_P(\mathbf{x}), \mathrm{d}\varphi_P(\mathbf{y}) \rangle_{\varphi(P)}^{\sim}.$$

Again, in the definition, $\mathbf{x}$ and $\mathbf{y}$ represent tangent vectors at $P \in V$, and on the right of the equality, we need to apply $\mathrm{d}\varphi_P$ to get tangent vectors at $\varphi(P) \in \tilde{V}$.

How are we sure this indeed is the right definition? We, at the very least, would expect isometries to preserve lengths. Let's see this is indeed the case. If $\gamma : [0, 1] \to V$ is a C^∞ curve, the composition $\tilde{\gamma} = \varphi \circ \gamma : [0, 1] \to \tilde{V}$ is a path in $\tilde{V}$. We let $P = \gamma(t)$, and hence $\varphi(P) = \tilde{\gamma}(t)$. Then

$$\langle \tilde{\gamma}'(t), \tilde{\gamma}'(t) \rangle_{\tilde{\gamma}(t)}^{\sim} = \langle \mathrm{d}\varphi_P \circ \gamma'(t), \mathrm{d}\varphi_P \circ \gamma'(t) \rangle_{\varphi(P)}^{\sim} = \langle \gamma'(t), \gamma'(t) \rangle_{\gamma(t)=P}.$$

Integrating, we obtain

$$\mathrm{length}(\tilde{\gamma}) = \mathrm{length}(\gamma) = \int_0^1 \langle \gamma'(t), \gamma'(t) \rangle_{\gamma(t)} \;\mathrm{d}t.$$

4.3 Two models for the hyperbolic plane

That's enough preparation. We can start talking about hyperbolic plane. We will in fact provide two models of the hyperbolic plane. Each model has its own strengths, and often proving something is significantly easier in one model than the other.

We start with the disk model.

Definition (Poincaré disk model)**.** The *(Poincaré) disk model* for the hyperbolic plane is given by the unit disk

$$D \subseteq \mathbb{C} \cong \mathbb{R}^2, \quad D = \{\zeta \in \mathbb{C} : |\zeta| < 1\},$$

and a Riemannian metric on this disk given by

$$\frac{4(\mathrm{d}u^2 + \mathrm{d}v^2)}{(1 - u^2 - v^2)^2} = \frac{4|\mathrm{d}\zeta|^2}{(1 - |\zeta|^2)^2}, \tag{$*$}$$

where $\zeta = u + iv$.

Note that this is similar to our previous metric for the sphere, but we have $1 - u^2 - v^2$ instead of $1 + u^2 + v^2$.

To interpret the term $|\mathrm{d}\zeta|^2$, we can either formally set $|\mathrm{d}\zeta|^2 = \mathrm{d}u^2 + \mathrm{d}v^2$, or interpret it as the derivative $\mathrm{d}\zeta = \mathrm{d}u + i\mathrm{d}v : \mathbb{C} \to \mathbb{C}$.

We see that $(*)$ is a scaling of the standard Riemannian metric by a factor depending on the polar radius $r = |\zeta|^2$. The distances are scaled by $\frac{2}{1-r^2}$, while the areas are scaled by $\frac{4}{(1-r^2)^2}$. Note, however, that the angles in the hyperbolic disk are the same as that in $\mathbb{R}^2$. This is in general true for metrics that are just scaled versions of the Euclidean metric (exercise).

Alternatively, we can define the hyperbolic plane with the upper-half plane.

Definition (Upper half-plane)**.** The *upper half-plane* is

$$H = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}.$$

What is the appropriate Riemannian metric to put on the upper half plane? We know D bijects to H via the Möbius transformation

$$\varphi : \zeta \in D \mapsto i\frac{1+\zeta}{1-\zeta} \in H.$$

This bijection is in fact a conformal equivalence, as defined in IB Complex Analysis/Methods. The idea is to pick a metric on H such that this map is an isometry. Then H together with this Riemannian metric will be the upper half-plane model for the hyperbolic plane.

To avoid confusion, we reserve the letter z for points $z \in H$, with $z = x + iy$, while we use ζ for points $\zeta \in D$, and write $\zeta = u + iv$. Then we have

$$z = i\frac{1+\zeta}{1-\zeta}, \quad \zeta = \frac{z-i}{z+i}.$$

Instead of trying to convert the Riemannian metric on D to H, which would be a complete algebraic horror, we first try converting the Euclidean metric. The Euclidean metric on $\mathbb{R}^2 = \mathbb{C}$ is given by

$$\langle w_1, w_2 \rangle = \mathrm{Re}(w_1\overline{w_2}) = \frac{1}{2}(w_1\bar{w}_2 + \bar{w}_1 w_2).$$

So if $\langle\,\cdot\,,\,\cdot\,\rangle_{\mathrm{eucl}}$ is the Euclidean metric at ζ, then at z such that $\zeta = \frac{z-i}{z+i}$, we require (by definition of isometry)

$$\langle w, v\rangle_z = \left\langle \frac{\mathrm{d}\zeta}{\mathrm{d}z} w, \frac{\mathrm{d}\zeta}{\mathrm{d}z} v\right\rangle_{\mathrm{eucl}} = \left|\frac{\mathrm{d}\zeta}{\mathrm{d}z}\right|^2 \operatorname{Re}(w\bar{v}) = \left|\frac{\mathrm{d}\zeta}{\mathrm{d}z}\right|^2 (w_1 v_1 + w_2 v_2),$$

where $w = w_1 + iw_2, v = v_1 + iv_2$.

Hence, on H, we obtain the Riemannian metric

$$\left|\frac{\mathrm{d}\zeta}{\mathrm{d}z}\right|^2 (\mathrm{d}x^2 + \mathrm{d}y^2).$$

We can compute

$$\frac{\mathrm{d}\zeta}{\mathrm{d}z} = \frac{1}{z+i} - \frac{z-i}{(z+i)^2} = \frac{2i}{(z+i)^2}.$$

This is what we get if we started with a Euclidean metric. If we start with the hyperbolic metric on D, we get an additional scaling factor. We can do some computations to get

$$1 - |\zeta|^2 = 1 - \frac{|z-i|^2}{|z+i|^2},$$

and hence

$$\frac{1}{1-|\zeta|^2} = \frac{|z+i|^2}{|z+i|^2 - |z-i|^2} = \frac{|z+i|^2}{4 \operatorname{Im} z}.$$

Putting all these together, metric corresponding to $\frac{4|\mathrm{d}\zeta|^2}{(1-|\zeta|^2)^2}$ on D is

$$4 \cdot \frac{4}{|z+i|^4} \cdot \left(\frac{|z+i|^2}{4 \operatorname{Im} z}\right)^2 \cdot |\mathrm{d}z|^2 = \frac{|\mathrm{d}z|^2}{(\operatorname{Im} z)^2} = \frac{\mathrm{d}x^2 + \mathrm{d}y^2}{y^2}.$$

We now use all these ingredients to define the upper half-plane model.

Definition (Upper half-plane model)**.** The *upper half-plane* model of the hyperbolic plane is the upper half-plane H with the Riemannian metric

$$\frac{\mathrm{d}x^2 + \mathrm{d}y^2}{y^2}.$$

The lengths on H are scaled (from the Euclidean one) by $\frac{1}{y}$, while the areas are scaled by $\frac{1}{y^2}$. Again, the angles are the same.

Note that we did not have to go through so much mess in order to define the sphere. This is since we can easily "embed" the surface of the sphere in $\mathbb{R}^3$. However, there is no easy surface in $\mathbb{R}^3$ that gives us the hyperbolic plane. As we don't have an actual prototype, we need to rely on the more abstract data of a Riemannian metric in order to work with hyperbolic geometry.

We are next going to study the geometry of H, We claim that the following group of Möbius maps are isometries of H:

$$\mathrm{PSL}(2, \mathbb{R}) = \left\{z \mapsto \frac{az+b}{cz+d} : a, b, c, d \in \mathbb{R}, ad - bc = 1\right\}.$$

Note that the coefficients have to be real, not complex.

Proposition. The elements of $\mathrm{PSL}(2,\mathbb{R})$ are isometries of H, and this preserves the lengths of curves.

Proof. It is easy to check that $\mathrm{PSL}(2,\mathbb{R})$ is generated by

(i) Translations $z \mapsto z + a$ for $a \in \mathbb{R}$

(ii) Dilations $z \mapsto az$ for $a > 0$

(iii) The single map $z \mapsto -\frac{1}{z}$.

So it suffices to show each of these preserves the metric $\frac{|\mathrm{d}z|^2}{y^2}$, where $z = x + iy$. The first two are straightforward to see, by plugging it into formula and notice the metric does not change.

We now look at the last one, given by $z \mapsto -\frac{1}{z}$. The derivative at z is

$$f'(z) = \frac{1}{z^2}.$$

So we get

$$\mathrm{d}z \mapsto \mathrm{d}\left(-\frac{1}{z}\right) = \frac{\mathrm{d}z}{z^2}.$$

So

$$\left|\mathrm{d}\left(-\frac{1}{z}\right)\right|^2 = \frac{|\mathrm{d}z|^2}{|z|^4}.$$

We also have

$$\mathrm{Im}\left(-\frac{1}{z}\right) = -\frac{1}{|z|^2}\,\mathrm{Im}\,\bar{z} = \frac{\mathrm{Im}\,z}{|z|^2}.$$

So

$$\frac{|\mathrm{d}(-1/z)|^2}{\mathrm{Im}(-1/z)^2} = \left(\frac{|\mathrm{d}z|^2}{|z^4|}\right) \Big/ \left(\frac{(\mathrm{Im}\,z)^2}{|z|^4}\right) = \frac{|\mathrm{d}z|^2}{(\mathrm{Im}\,z)^2}.$$

So this is an isometry, as required. □

Note that each $z \mapsto az + b$ with $a > 0, b \in \mathbb{R}$ is in $\mathrm{PSL}(2,\mathbb{R})$. Also, we can use maps of this form to send any point to any other point. So $\mathrm{PSL}(2,\mathbb{R})$ acts transitively on H. Moreover, everything in $\mathrm{PSL}(2,\mathbb{R})$ fixes $\mathbb{R} \cup \{\infty\}$.

Recall also that each Möbius transformation preserves circles and lines in the complex plane, as well as angles between circles/lines. In particular, consider the line $L = i\mathbb{R}$, which meets $\mathbb{R}$ perpendicularly, and let $g \in \mathrm{PSL}(2,\mathbb{R})$. Then the image is either a circle centered at a point in $\mathbb{R}$, or a straight line perpendicular to $\mathbb{R}$.

We let $L^+ = L \cap H = \{it : t > 0\}$. Then $g(L^+)$ is either a vertical half-line or a semi-circle that ends in $\mathbb{R}$.

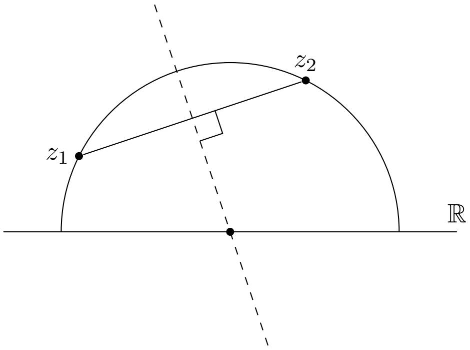
Definition (Hyperbolic lines)**.** *Hyperbolic lines* in H are vertical half-lines or semicircles ending in $\mathbb{R}$.

We will now prove some lemmas to justify why we call these hyperbolic lines.

Lemma. Given any two distinct points $z_1, z_2 \in H$, there exists a unique hyperbolic line through z_1 and z_2.

Proof. This is clear if $\operatorname{Re} z_1 = \operatorname{Re} z_2$ — we just pick the vertical half-line through them, and it is clear this is the only possible choice.

Otherwise, if $\operatorname{Re} z_1 \neq \operatorname{Re} z_2$, then we can find the desired circle as follows:



It is also clear this is the only possible choice. □

Lemma. $\mathrm{PSL}(2,\mathbb{R})$ acts transitively on the set of hyperbolic lines in H.

Proof. It suffices to show that for each hyperbolic line ℓ, there is some $g \in \mathrm{PSL}(2,\mathbb{R})$ such that $g(\ell) = L^+$. This is clear when ℓ is a vertical half-line, since we can just apply a horizontal translation.

If it is a semicircle, suppose it has end-points $s < t \in \mathbb{R}$. Then consider

$$g(z) = \frac{z-t}{z-s}.$$

This has determinant $-s+t > 0$. So $g \in \mathrm{PSL}(2,\mathbb{R})$. Then $g(t) = 0$ and $g(s) = \infty$. Then we must have $g(\ell) = L^+$, since $g(\ell)$ is a hyperbolic line, and the only hyperbolic lines passing through ∞ are the vertical half-lines. So done. □

Moreover, we can achieve $g(s) = 0$ and $g(t) = \infty$ by composing with $-\frac{1}{z}$. Also, for any $P \in \ell$ not on the endpoints, we can construct a g such that $g(P) = i \in L^+$, by composing with $z \mapsto az$. So the isometries act transitively on pairs (ℓ, P), where ℓ is a hyperbolic line and $P \in \ell$.

Definition (Hyperbolic distance)**.** For points $z_1, z_2 \in H$, the *hyperbolic distance* $\rho(z_1, z_2)$ is the length of the segment $[z_1, z_2] \subseteq \ell$ of the hyperbolic line through z_1, z_2 (parametrized monotonically).

Thus $\mathrm{PSL}(2,\mathbb{R})$ preserves hyperbolic distances. Similar to Euclidean space and the sphere, we show these lines minimize distance.

Proposition. If $\gamma : [0,1] \to H$ is a piecewise C^1-smooth curve with $\gamma(0) = z_1, \gamma(1) = z_2$, then $\operatorname{length}(\gamma) \geq \rho(z_1, z_2)$, with equality iff γ is a monotonic parametrisation of $[z_1, z_2] \subseteq \ell$, where ℓ is the hyperbolic line through z_1 and z_2.

Proof. We pick an isometry $g \in \mathrm{PSL}(2,\mathbb{R})$ so that $g(\ell) = L^+$. So without loss of generality, we assume $z_1 = iu$ and $z_2 = iv$, with $u < v \in \mathbb{R}$.

We decompose the path as $\gamma(t) = x(t) + iy(t)$. Then we have

$$\begin{aligned}\text{length}(\gamma) &= \int_0^1 \frac{1}{y}\sqrt{\dot{x}^2 + \dot{y}^2}\;\mathrm{d}t\\ &\geq \int_0^1 \frac{|\dot{y}|}{y}\;\mathrm{d}z\\ &\geq \left|\int_0^1 \frac{\dot{y}}{y}\;\mathrm{d}t\right|\\ &= [\log y(t)]_0^1\\ &= \log\left(\frac{v}{u}\right)\end{aligned}$$

This calculation also tells us that $\rho(z_1, z_2) = \log\left(\frac{v}{u}\right)$. so $\text{length}(\gamma) \geq \rho(z_1, z_2)$ with equality if and only if $x(t) = 0$ (hence $\gamma \subseteq L^+$) and $\dot{y} \geq 0$ (hence monotonic). □

Corollary (Triangle inequality)**.** Given three points $z_1, z_2, z_3 \in H$, we have

$$\rho(z_1, z_3) \leq \rho(z_1, z_2) + \rho(z_2, z_3),$$

with equality if and only if z_2 lies between z_1 and z_2.

Hence, (H, ρ) is a metric space.

4.4 Geometry of the hyperbolic disk

So far, we have worked with the upper half-plane model. This is since the upper half-plane model is more convenient for these calculations. However, sometimes the disk model is more convenient. So we also want to understand that as well.

Recall that $\zeta \in D \mapsto z = i\frac{1+\zeta}{1-\zeta} \in H$ is an isometry, with an (isometric) inverse $z \in H \mapsto \zeta = \frac{z-i}{z+i} \in D$. Moreover, since these are Möbius maps, circle-lines are preserved, and angles between the lines are also preserved.

Hence, immediately from previous work on H, we know

(i) $\text{PSL}(2, \mathbb{R}) \cong \{\text{Möbius transformations sending } D \text{ to itself}\} = G$.

(ii) Hyperbolic lines in D are circle segments meeting $|\zeta| = 1$ orthogonally, including diameters.

(iii) G acts *transitively* on hyperbolic lines in D (and also on pairs consisting of a line and a point on the line).

(iv) The length-minimizing geodesics on D are a segments of hyperbolic lines parametrized monotonically.

We write ρ for the (hyperbolic) distance in D.

Lemma. Let G be the set of isometries of the hyperbolic disk. Then

(i) Rotations $z \mapsto e^{i\theta}z$ (for $\theta \in \mathbb{R}$) are elements of G.

(ii) If $a \in D$, then $g(z) = \frac{z-a}{1-\bar{a}z}$ is in G.

Proof.

(i) This is clearly an isometry, since this is a linear map, preserves $|z|$ and $|\mathrm{d}z|$, and hence also the metric

$$\frac{4|\mathrm{d}z|^2}{(1-|z|^2)^2}.$$

(ii) First, we need to check this indeed maps D to itself. To do this, we first make sure it sends $\{|z|=1\}$ to itself. If $|z|=1$, then

$$|1-\bar{a}z| = |\bar{z}(1-\bar{a}z)| = |\bar{z}-\bar{a}| = |z-a|.$$

So

$$|g(z)| = 1.$$

Finally, it is easy to check $g(a)=0$. By continuity, G must map D to itself. We can then show it is an isometry by plugging it into the formula. □

It is an exercise on the second example sheet to show all $g \in G$ is of the form

$$g(z) = e^{i\theta}\frac{z-a}{1-\bar{a}z}$$

or

$$g(z) = e^{i\theta}\frac{\bar{z}-a}{1-\bar{a}\bar{z}}$$

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

We shall now use the disk model to do something useful. We start by coming up with an explicit formula for distances in the hyperbolic plane.

Proposition. If $0 \leq r < 1$, then

$$\rho(0, re^{i\theta}) = 2\tanh^{-1} r.$$

In general, for $z_1, z_2 \in D$, we have

$$g(z_1, z_2) = 2\tanh^{-1}\left|\frac{z_1-z_2}{1-\bar{z_1}z_2}\right|.$$

Proof. By the lemma above, we can rotate the hyperbolic disk so that $re^{i\theta}$ is rotated to r. So

$$\rho(0, re^{i\theta}) = \rho(0, r).$$

We can evaluate this by performing the integral

$$\rho(0,r) = \int_0^r \frac{2\,\mathrm{d}t}{1-t^2} = 2\tanh^{-1} r.$$

For the general case, we apply the Möbius transformation

$$g(z) = \frac{z-z_1}{1-\bar{z}_1 z}.$$

Then we have

$$g(z_1) = 0, \quad g(z_2) = \frac{z_2-z_1}{1-\bar{z}_1 z_2} = \left|\frac{z_1-z_2}{1-\bar{z_1}z_2}\right| e^{i\theta}.$$

So

$$\rho(z_1,z_2) = \rho(g(z_1), g(z_2)) = 2\tanh^{-1}\left|\frac{z_1-z_2}{1-\bar{z_1}z_2}\right|.$$ □

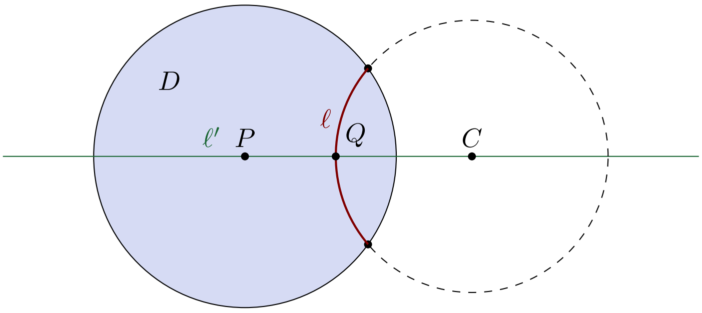
Again, we exploited the idea of performing the calculation in an easy case, and then using isometries to move everything else to the easy case. In general, when we have a "distinguished" point in the hyperbolic plane, it is often convenient to use the disk model, move it to 0 by an isometry.

Proposition. For every point P and hyperbolic line ℓ, with $P \not\in \ell$, there is a unique line ℓ' with $P \in \ell'$ such that ℓ' meets ℓ orthogonally, say $\ell \cap \ell' = Q$, and $\rho(P, Q) \leq \rho(P, \tilde{Q})$ for all $\tilde{Q} \in \ell$.

This is a familiar fact from Euclidean geometry. To prove this, we again apply the trick of letting $P = 0$.

Proof. wlog, assume $P = 0 \in D$. Note that a line in D (that is not a diameter) is a Euclidean circle. So it has a center, say C.

Since any line through P is a diameter, there is clearly only one line that intersects ℓ perpendicularly (recall angles in D is the same as the Euclidean angle).



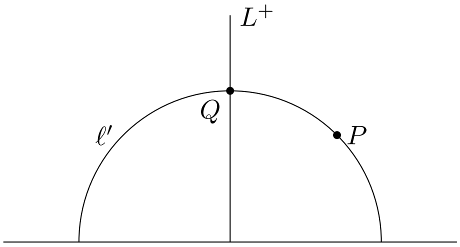
It is also clear that PQ minimizes the *Euclidean* distance between P and ℓ. While this is not the same as the hyperbolic distance, since hyperbolic lines through P are diameters, having a larger hyperbolic distance is equivalent to having a higher Euclidean distance. So this indeed minimizes the distance. □

How does reflection in hyperbolic lines work? This time, we work in the upper half-plane model, since we have a favorite line L^+.

Lemma (Hyperbolic reflection)**.** Suppose g is an isometry of the hyperbolic half-plane H and g fixes every point in $L^+ = \{iy : y \in \mathbb{R}^+\}$. Then G is either the identity or $g(z) = -\bar{z}$, i.e. it is a reflection in the vertical axis L^+.

Observe we have already proved a similar result in Euclidean geometry, and the spherical version was proven in the first example sheet.

Proof. For every $P \in H \setminus L^+$, there is a unique line ℓ' containing P such that $\ell' \perp L^+$. Let $Q = L^+ \cap \ell'$.



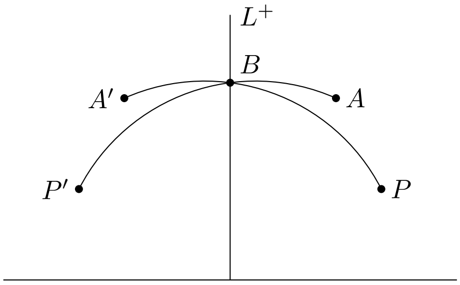
We see ℓ' is a semicircle, and by definition of isometry, we must have

$$\rho(P, Q) = \rho(g(P), Q).$$

Now note that $g(\ell')$ is also a line meeting L^+ perpendicularly at Q, since g fixes L^+ and preserves angles. So we must have $g(\ell') = \ell'$. Then in particular $g(P) \in \ell'$. So we must have $g(P) = P$ or $g(P) = P'$, where P' is the image under reflection in L^+.

Now it suffices to prove that if $g(P) = P$ for any one P, then $g(P)$ must be the identity (if $g(P) = P'$ for all P, then g must be given by $g(z) = -\bar{z}$).

Now suppose $g(P) = P$, and let $A \in H^+$, where $H^+ = \{z \in H : \operatorname{Re} z > 0\}$.



Now if $g(A) \neq A$, then $g(A) = A'$. Then $\rho(A', P) = \rho(A, P)$. But

$$\rho(A', P) = \rho(A', B) + \rho(B, P) = \rho(A, B) + \rho(B, P) > \rho(A, P),$$

by the triangle inequality, noting that $B \not\in (AP)$. This is a contradiction. So g must fix everything. □

Definition (Hyperbolic reflection)**.** The map $R : z \in H \mapsto -\bar{z} \in H$ is the *(hyperbolic) reflection in* L^+. More generally, given any hyperbolic line ℓ, let T be the isometry that sends ℓ to L^+. Then the *(hyperbolic) reflection in* ℓ is

$$R_\ell = T^{-1} R T$$

Again, we already know how to reflect in L^+. So to reflect in another line ℓ, we move our plane such that ℓ becomes L^+, do the reflection, and move back.

By the previous proposition, R_ℓ is the unique isometry that is not identity and fixes ℓ.

4.5 Hyperbolic triangles

Definition (Hyperbolic triangle)**.** A *hyperbolic triangle* ABC is the region determined by three hyperbolic line segments AB, BC and CA, including extreme cases where some vertices A, B, C are allowed to be "at infinity". More precisely, in the half-plane model, we allow them to lie in $\mathbb{R} \cup \{\infty\}$; in the disk model we allow them to lie on the unit circle $|z| = 1$.

We see that if A is "at infinity", then the angle at A must be zero.

Recall for a region $R \subseteq H$, we can compute the area of R as

$$\operatorname{area}(R) = \iint_R \frac{\mathrm{d}x\,\mathrm{d}y}{y^2}.$$

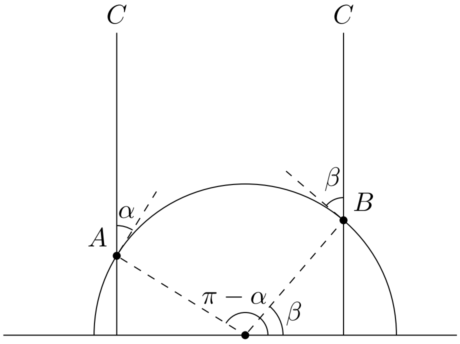
Similar to the sphere, we have

Theorem (Gauss-Bonnet theorem for hyperbolic triangles)**.** For each hyperbolic triangle Δ, say, ABC, with angles $\alpha, \beta, \gamma \geq 0$ (note that zero angle is possible), we have

$$\operatorname{area}(\Delta) = \pi - (\alpha + \beta + \gamma).$$

Proof. First do the case where $\gamma = 0$, so C is "at infinity". Recall that we like to use the disk model if we have a distinguished point in the hyperbolic plane. If we have a distinguished point at *infinity*, it is often advantageous to use the upper half plane model, since ∞ is a distinguished point at infinity.

So we use the upper-half plane model, and wlog $C = \infty$ (apply $\mathrm{PSL}(2, \mathbb{R})$) if necessary. Then AC and BC are vertical half-lines. So AB is the arc of a semi-circle. So AB is an arc of a semicircle.



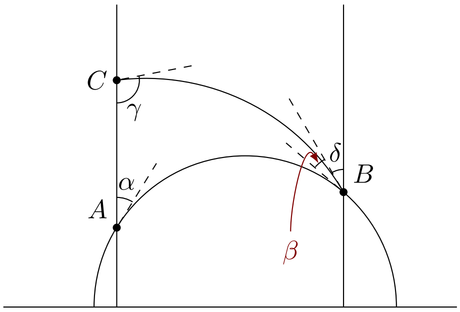
We use the transformation $z \mapsto z + a$ (with $a \in \mathbb{R}$) to center the semi-circle at 0. We then apply $z \mapsto bz$ (with $b > 0$) to make the circle have radius 1. Thus wlog $AB \subseteq \{x^2 + y^2 = 1\}$.

Now we have

$$\begin{aligned}
\operatorname{area}(T) &= \int_{\cos(\pi - \alpha)}^{\cos \beta} \int_{\sqrt{1 - x^2}}^{\infty} \frac{1}{y^2}\;\mathrm{d}y\;\mathrm{d}x\\
&= \int_{\cos(\pi - \alpha)}^{\cos \beta} \frac{1}{\sqrt{1 - x^2}}\;\mathrm{d}x\\
&= [-\cos^{-1}(x)]_{\cos(\pi - \alpha)}^{\cos \beta}\\
&= \pi - \alpha - \beta,
\end{aligned}$$

as required.

In general, we use H again, and we can arrange AC in a vertical half-line. Also, we can move AB to $x^2 + y^2 = 1$, noting that this transformation keeps AC vertical.



We consider $\Delta_1 = AB\infty$ and $\Delta_2 = CB\infty$. Then we can immediately write

$$\begin{aligned}\mathrm{area}(\Delta_1) &= \pi - \alpha - (\beta + \delta)\\ \mathrm{area}(\Delta_2) &= \pi - \delta - (\pi - \gamma) = \gamma - \delta.\end{aligned}$$

So we have

$$\mathrm{area}(T) = \mathrm{area}(\Delta_2) - \mathrm{area}(\Delta_1) = \pi - \alpha - \beta - \gamma,$$

as required. □

Similar to the spherical case, we have some hyperbolic sine and cosine rules. For example, we have

Theorem (Hyperbolic cosine rule)**.** In a triangle with sides a, b, c and angles α, β, γ, we have

$$\cosh c = \cosh a \cosh b - \sinh a \sinh b \cos \gamma.$$

Proof. See example sheet 2. □

Recall that in S^2, any two lines meet (in two points). In the Euclidean plane $\mathbb{R}^2$, any two lines meet (in one point) iff they are not parallel. Before we move on to the hyperbolic case, we first make a definition.

Definition (Parallel lines)**.** We use the disk model of the hyperbolic plane. Two hyperbolic lines are *parallel* iff they meet only at the boundary of the disk (at $|z| = 1$).

Definition (Ultraparallel lines)**.** Two hyperbolic lines are *ultraparallel* if they don't meet anywhere in $\{|z| \leq 1\}$.

In the Euclidean plane, we have the parallel axiom: given a line ℓ and $P \notin \ell$, there exists a unique line ℓ' containing P with $\ell \cap \ell' = \emptyset$. This fails in both S^2 and the hyperbolic plane — but for very different reasons! In S^2, there are no such parallel lines. In the hyperbolic plane, there are *many* parallel lines. There is a more deep reason for why this is the case, which we will come to at the very end of the course.

4.6 Hyperboloid model

Recall we said there is no way to view the hyperbolic plane as a subset of $\mathbb{R}^3$, and hence we need to mess with Riemannian metrics. However, it turns out we can indeed embed the hyperbolic plane in $\mathbb{R}^3$, if we give $\mathbb{R}^3$ a different metric!

Definition (Lorentzian inner product)**.** The *Lorentzian inner product* on $\mathbb{R}^3$ has the matrix

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

This is less arbitrary as it seems. Recall from IB Linear Algebra that we can always pick a basis where a non-degenerate symmetric bilinear form has diagonal made of 1 and -1. If we further identify A and $-A$ as the "same" symmetric bilinear form, then this is the only other possibility left.

Thus, we obtain the quadratic form given by

$$q(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x}\rangle = x^2 + y^2 - z^2.$$

We now define the 2-sheet hyperboloid as

$$\{\mathbf{x} \in \mathbb{R}^2 : q(\mathbf{x}) = -1\}.$$

This is given explicitly by the formula

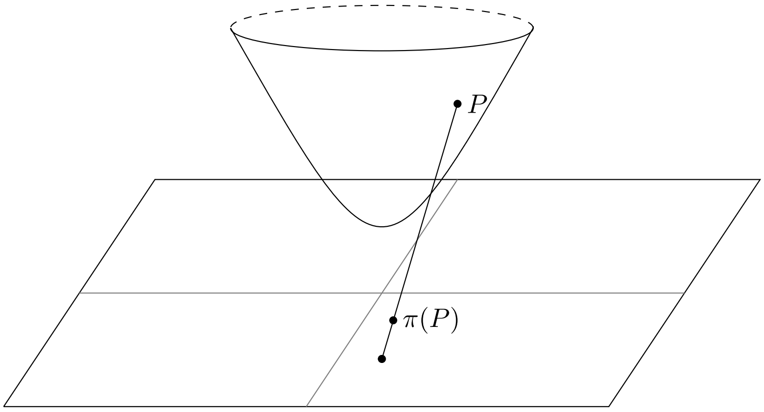
$$x^2 + y^2 = z^2 - 1.$$

We don't actually need to two sheets. So we define

$$S^+ = S \cap \{z > 0\}.$$

We let $\pi : S^+ \to D \subseteq \mathbb{C} = \mathbb{R}^2$ be the stereographic projection from (0, 0, -1) by

$$\pi(x, y, z) = \frac{x + iy}{1 + z} = u + iv.$$



We put $r^2 = u^2 + v^2$. Doing some calculations, we show that

(i) We always have $r < 1$, as promised.

(ii) The stereographic projection π is invertible with

$$\sigma(u, v) = \pi^{-1}(u, v) = \frac{1}{1 - r^2}(2u, 2v, 1 + r^2) \in S^+.$$

(iii) The tangent plane to S^+ at P is spanned by

$$\sigma_u = \frac{\partial \sigma}{\sigma u}, \quad \sigma_v = \frac{\partial \sigma}{\partial v}.$$

We can explicitly compute these to be

$$\begin{aligned}\sigma_u &= \frac{2}{(1-r^2)^2}(1+u^2-v^2, 2uv, 2u),\\ \sigma_v &= \frac{2}{(1-r^2)^2}(2uv, 1+v^2-u^2, 2v).\end{aligned}$$

We restrict the inner product $\langle\,\cdot\,,\,\cdot\,\rangle$ to the span of σ_u, σ_v, and we get a symmetric bilinear form assigned to each $u, v \in D$ given by

$$E\;\mathrm{d}u^2 + 2F\;\mathrm{d}u\;\mathrm{d}v + G\;\mathrm{d}v^2,$$

where

$$\begin{aligned}E &= \langle \sigma_u, \sigma_u\rangle = \frac{4}{(1-r^2)^2},\\ F &= \langle \sigma_u, \sigma_v\rangle = 0,\\ G &= \langle \sigma_v, \sigma_v\rangle = \frac{4}{(1-r^2)^2}.\end{aligned}$$

We have thus recovered the Poincare disk model of the hyperbolic plane.

5 Smooth embedded surfaces (in $\mathbb{R}^3$)

5.1 Smooth embedded surfaces

So far, we have been studying some specific geometries, namely Euclidean, spherical and hyperbolic geometry. From now on, we go towards greater generality, and study arbitrary surfaces. We will mostly work with surfaces that are smoothly embedded as subsets of $\mathbb{R}^3$, we can develop notions parallel to those we have had before, such as Riemannian metrics and lengths. At the very end of the course, we will move away from the needless restriction restriction that the surface is embedded in $\mathbb{R}^3$, and study surfaces that *just are*.

Definition (Smooth embedded surface)**.** A set $S \subseteq \mathbb{R}^3$ is a *(parametrized) smooth embedded surface* if every point $P \in S$ has an open neighbourhood $U \subseteq S$ (with the subspace topology on $S \subseteq \mathbb{R}^3$) and a map $\sigma : V \to U$ from an open $V \subseteq \mathbb{R}^2$ to U such that if we write $\sigma(u, v) = (x(u, v), y(u, v), z(u, v))$, then

(i) σ is a homeomorphism (i.e. a bijection with continuous inverse)

(ii) σ is C^∞ (smooth) on V (i.e. has continuous partial derivatives of all orders).

(iii) For all $Q \in V$, the partial derivatives $\sigma_u(Q)$ and $\sigma_v(Q)$ are linearly independent.

Recall that

$$\sigma_u(Q) = \frac{\partial \sigma}{\partial u}(Q) = \begin{pmatrix} \frac{\partial x}{\partial u} \\ \frac{\partial y}{\partial u} \\ \frac{\partial z}{\partial u} \end{pmatrix}(Q) = \mathrm{d}\sigma_Q(\mathbf{e}_1),$$

where $\mathbf{e}_1, \mathbf{e}_2$ is the standard basis of $\mathbb{R}^2$. Similarly, we have

$$\sigma_v(Q) = \mathrm{d}\sigma_Q(\mathbf{e}_2).$$

We define some terminology.

Definition (Smooth coordinates)**.** We say (u, v) are *smooth coordinates* on $U \subseteq S$.

Definition (Tangent space)**.** The subspace of $\mathbb{R}^3$ spanned by $\sigma_u(Q), \sigma_v(Q)$ is the *tangent space* $T_P S$ to S at $P = \sigma(Q)$.

Definition (Smooth parametrisation)**.** The function σ is a *smooth parametrisation* of $U \subseteq S$.

Definition (Chart)**.** The function $\sigma^{-1} : U \to V$ is a *chart* of U.

Proposition. Let $\sigma : V \to U$ and $\tilde{\sigma} : \tilde{V} \to U$ be two C^∞ parametrisations of a surface. Then the homeomorphism

$$\varphi = \sigma^{-1} \circ \tilde{\sigma} : \tilde{V} \to V$$

is in fact a diffeomorphism.

This proposition says any two parametrizations of the same surface are compatible.

Proof. Since differentiability is a local property, it suffices to consider φ on some small neighbourhood of a point in V. Pick our favorite point $(v_0, u_0) \in \tilde{V}$. We know $\sigma = \sigma(u, v)$ is differentiable. So it has a Jacobian matrix

$$\begin{pmatrix} x_u & x_v \\ y_u & y_v \\ z_y & z_v \end{pmatrix}.$$

By definition, this matrix has rank two at each point. wlog, we assume the first two rows are linearly independent. So

$$\det \begin{pmatrix} x_u & x_v \\ y_u & y_v \end{pmatrix} \neq 0$$

at $(v_0, u_0) \in \tilde{V}$. We define a new function

$$F(u, v) = \begin{pmatrix} x(u, v) \\ y(u, v) \end{pmatrix}.$$

Now the inverse function theorem applies. So F has a local C^∞ inverse, i.e. there are two open neighbourhoods $(u_0, v_0) \in N$ and $F(u_0, v_0) \in N' \subseteq \mathbb{R}^2$ such that $f : N \to N'$ is a diffeomorphism.

Writing $\pi : \tilde{\sigma} \to N'$ for the projection $\pi(x, y, z) = (x, y)$ we can put these things in a commutative diagram:

$$\begin{tikzcd} & \sigma(N) \arrow[d, "\pi"] \\ N \arrow[ur, "\sigma"] \arrow[r, "F"'] & N' \end{tikzcd}.$$

We now let $\tilde{N} = \tilde{\sigma}^{-1}(\sigma(N))$ and $\tilde{F} = \pi \circ \tilde{\sigma}$, which is yet again smooth. Then we have the following larger commutative diagram.

$$\begin{tikzcd} & \sigma(N) \arrow[d, "\pi"] & \\ N \arrow[ur, "\sigma"] \arrow[r, "F"'] & N' & \tilde{N} \arrow[ul, "\tilde{\sigma}"'] \arrow[l, "\tilde{F}"] \end{tikzcd}.$$

Then we have

$$\varphi = \sigma^{-1} \circ \tilde{\sigma} = \sigma^{-1} \circ \pi^{-1} \circ \pi \circ \tilde{\sigma} = F^{-1} \circ \tilde{F},$$

which is smooth, since F^{-1} and $\tilde{F}$ are. Hence φ is smooth everywhere. By symmetry of the argument, φ^{-1} is smooth as well. So this is a diffeomorphism. □

A more practical result is the following:

Corollary. The tangent plane $T_Q S$ is independent of parametrization.

Proof. We know

$$\tilde{\sigma}(\tilde{u}, \tilde{v}) = \sigma(\varphi_1(\tilde{u}, \tilde{v}), \varphi_2(\tilde{u}, \tilde{v})).$$

We can then compute the partial derivatives as

$$\begin{aligned}\tilde{\sigma}_{\tilde{u}} &= \varphi_{1,\tilde{u}}\sigma_u + \varphi_{2,\tilde{u}}\sigma_v\\ \tilde{\sigma}_{\tilde{v}} &= \varphi_{1,\tilde{v}}\sigma_u + \varphi_{2,\tilde{v}}\sigma_v\end{aligned}$$

Here the transformation is related by the Jacobian matrix

$$\begin{pmatrix}\varphi_{1,\tilde{u}} & \varphi_{1,\tilde{v}}\\ \varphi_{2,\tilde{u}} & \varphi_{2,\tilde{v}}\end{pmatrix} = J(\varphi).$$

This is invertible since φ is a diffeomorphism. So $(\sigma_{\tilde{u}}, \sigma_{\tilde{v}})$ and (σ_u, σ_v) are different basis of the same two-dimensional vector space. So done. □

Note that we have

$$\tilde{\sigma}_{\tilde{u}} \times \tilde{\sigma}_{\tilde{v}} = \det(J(\varphi))\sigma_u \times \sigma_v.$$

So we can define

Definition (Unit normal)**.** The *unit normal* to S at $Q \in S$ is

$$N = N_Q = \frac{\sigma_u \times \sigma_v}{\|\sigma_u \times \sigma_v\|},$$

which is well-defined up to a sign.

Often, instead of a parametrization $\sigma : V \subseteq \mathbb{R}^2 \to U \subseteq S$, we want the function the other way round. We call this a chart.

Definition (Chart)**.** Let $S \subseteq \mathbb{R}^3$ be an embedded surface. The map $\theta = \sigma^{-1} : U \subseteq S \to V \subseteq \mathbb{R}^2$ is a *chart.*

Example. Let $S^2 \subseteq \mathbb{R}^3$ be a sphere. The two stereographic projections from $\pm\mathbf{e}_3$ give two charts, whose domain together cover S^2.

Similar to what we did to the sphere, given a chart $\theta : U \to V \subseteq \mathbb{R}^2$, we can induce a Riemannian metric on V. We first get an inner product on the tangent space as follows:

Definition (First fundamental form)**.** If $S \subseteq \mathbb{R}^3$ is an embedded surface, then each $T_Q S$ for $Q \in S$ has an inner product from $\mathbb{R}^3$, i.e. we have a family of inner products, one for each point. We call this family the *first fundamental form.*

This is a theoretical entity, and is more easily worked with when we have a chart. Suppose we have a parametrization $\sigma : V \to U \subseteq S$, $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$, and $P \in V$. We can then define

$$\langle \mathbf{a}, \mathbf{b}\rangle_P = \langle \mathrm{d}\sigma_P(\mathbf{a}), \mathrm{d}\sigma_P(\mathbf{b})\rangle_{\mathbb{R}^3}.$$

With respect to the standard basis $\mathbf{e}_1, \mathbf{e}_2 \in \mathbb{R}^2$, we can write the first fundamental form as

$$E\ \mathrm{d}u^2 + 2F\ \mathrm{d}u\ \mathrm{d}v + G\ \mathrm{d}v^2,$$

where

$$\begin{aligned}E &= \langle \sigma_u, \sigma_u\rangle = \langle \mathbf{e}_1, \mathbf{e}_1\rangle_P\\ F &= \langle \sigma_u, \sigma_v\rangle = \langle \mathbf{e}_1, \mathbf{e}_2\rangle_P\\ G &= \langle \sigma_v, \sigma_v\rangle = \langle \mathbf{e}_2, \mathbf{e}_2\rangle_P.\end{aligned}$$

Thus, this induces a Riemannian metric on V. This is also called the first fundamental form corresponding to σ. This is what we do in practical examples.

We will assume the following property, which we are not bothered to prove.

Proposition. If we have two parametrizations related by $\tilde{\sigma} = \sigma \circ \varphi : \tilde{V} \to U$, then $\varphi : \tilde{V} \to V$ is an isometry of Riemannian metrics (on V and $\tilde{V}$).

Definition (Length and energy of curve)**.** Given a smooth curve $\Gamma : [a, b] \to S \subseteq \mathbb{R}^3$, the *length* of γ is

$$\text{length}(\Gamma) = \int_a^b \|\Gamma'(t)\| \;\mathrm{d}t.$$

The *energy* of the curve is

$$\text{energy}(\Gamma) = \int_a^b \|\Gamma'(t)\|^2 \;\mathrm{d}t.$$

We can think of the energy as something like the kinetic energy of a particle along the path, except that we are missing the factor of $\frac{1}{2}m$, because they are annoying.

How does this work with parametrizations? For the sake of simplicity, we assume $\Gamma([a, b]) \subseteq U$ for some parametrization $\sigma : V \to U$. Then we define the new curve

$$\gamma = \sigma^{-1} \circ \Gamma : [a, b] \to V.$$

This curve has two components, say $\gamma = (\gamma_1, \gamma_2)$. Then we have

$$\Gamma'(t) = (\mathrm{d}\sigma)_{\gamma(t)}(\dot{\gamma}_1(t)\mathbf{e}_1 + \dot{\gamma}_2(t)\mathbf{e}_2) = \dot{\gamma}_1\sigma_u + \dot{\gamma}_2\sigma_v,$$

and thus

$$\|\Gamma'(t)\| = \langle \dot{\gamma}, \dot{\gamma} \rangle_P^{\frac{1}{2}} = (E\dot{\gamma}_1^2 + 2F\dot{\gamma}_1\dot{\gamma}_2 + G\dot{\gamma}_2^2)^{\frac{1}{2}}.$$

So we get

$$\text{length}\,\Gamma = \int_a^b (E\dot{\gamma}_1^2 + 2F\dot{\gamma}_1\dot{\gamma}_2 + G\dot{\gamma}_2^2)^{\frac{1}{2}} \;\mathrm{d}t.$$

Similarly, the energy is given by

$$\text{energy}\,\Gamma = \int_a^b (E\dot{\gamma}_1^2 + 2F\dot{\gamma}_1\dot{\gamma}_2 + G\dot{\gamma}_2^2) \;\mathrm{d}t.$$

This agrees with what we've had for Riemannian metrics.

Definition (Area)**.** Given a smooth C^∞ parametrization $\sigma : V \to U \subseteq S \subseteq \mathbb{R}^3$, and a region $T \subseteq U$, we define the *area* of T to be

$$\text{area}(T) = \int_{\theta(T)} \sqrt{EG - F^2} \;\mathrm{d}u \;\mathrm{d}v,$$

whenever the integral exists (where $\theta = \sigma^{-1}$ is a chart).

Proposition. The area of T is independent of the choice of parametrization. So it extends to more general subsets $T \subseteq S$, not necessarily living in the image of a parametrization.

Proof. Exercise! □

Note that in examples, $\sigma(V) = U$ often is a dense set in S. For example, if we work with the sphere, we can easily parametrize everything but the poles. In that case, it suffices to use just one parametrization σ for area(S).

Note also that areas are invariant under isometries.

5.2 Geodesics

We now come to the important idea of a *geodesic*. We will first define these for Riemannian metrics, and then generalize it to general embedded surfaces.

Definition (Geodesic)**.** Let $V \subseteq \mathbb{R}^2_{u,v}$ be open, and $E\ \mathrm{d}u^2 + 2F\ \mathrm{d}u\ \mathrm{d}v + G\ \mathrm{d}v^2$ be a Riemannian metric on V. We let

$$\gamma = (\gamma_1, \gamma_2) : [a, b] \to V$$

be a smooth curve. We say γ is a *geodesic* with respect to the Riemannian metric if it satisfies

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(E\dot{\gamma}_1 + F\dot{\gamma}_2) &= \frac{1}{2}(E_u\dot{\gamma}_1^2 + 2F_u\dot{\gamma}_1\dot{\gamma}_2 + G_u\dot{\gamma}_2^2)\\ \frac{\mathrm{d}}{\mathrm{d}t}(F\dot{\gamma}_1 + G\dot{\gamma}_2) &= \frac{1}{2}(E_v\dot{\gamma}_1^2 + 2F_v\dot{\gamma}_1\dot{\gamma}_2 + G_v\dot{\gamma}_2^2)\end{aligned}$$

for all $t \in [a, b]$. These equations are known as the *geodesic ODEs*.

What exactly do these equations mean? We will soon show that these are curves that minimize (more precisely, are stationary points of) energy. To do so, we need to come up with a way of describing what it means for γ to minimize energy among all possible curves.

Definition (Proper variation)**.** Let $\gamma : [a, b] \to V$ be a smooth curve, and let $\gamma(a) = p$ and $\gamma(b) = q$. A *proper variation* of γ is a C^∞ map

$$h : [a, b] \times (-\varepsilon, \varepsilon) \subseteq \mathbb{R}^2 \to V$$

such that

$$h(t, 0) = \gamma(t) \text{ for all } t \in [a, b],$$

and

$$h(a, \tau) = p, \quad h(b, \tau) = q \text{ for all } |\tau| < \varepsilon,$$

and that

$$\gamma_\tau = h(\,\cdot\,, \tau) : [a, b] \to V$$

is a C^∞ curve for all fixed $\tau \in (-\varepsilon, \varepsilon)$.

Proposition. A smooth curve γ satisfies the geodesic ODEs if and only if γ is a stationary point of the energy function for all proper variation, i.e. if we define the function

$$E(\tau) = \text{energy}(\gamma_\tau) : (-\varepsilon, \varepsilon) \to \mathbb{R},$$

then

$$\left.\frac{\mathrm{d}E}{\mathrm{d}\tau}\right|_{\tau=0} = 0.$$

Proof. We let $\gamma(t) = (u(t), v(t))$. Then we have

$$\text{energy}(\gamma) = \int_a^b (E(u,v)\dot{u}^2 + 2F(u,v)\dot{u}\dot{v} + G(u,v)\dot{v}^2)\;\mathrm{d}t = \int_a^b I(u, v, \dot{u}, \dot{v})\;\mathrm{d}t.$$

We consider this as a function of four variables $u, \dot{u}, v, \dot{v}$, which are not necessarily related to one another. From the calculus of variations, we know γ is stationary if and only if

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial I}{\partial \dot{u}}\right) = \frac{\partial I}{\partial u},\quad \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial I}{\partial \dot{v}}\right) = \frac{\partial I}{\partial v}.$$

The first equation gives us

$$\frac{\mathrm{d}}{\mathrm{d}t}(2(E\dot{u} + F\dot{v})) = E_u \dot{u}^2 + 2F_u \dot{u}\dot{v} + G_u \dot{v}^2,$$

which is exactly the geodesic ODE. Similarly, the second equation gives the other geodesic ODE. So done. □

Since the definition of a geodesic involves the derivative only, which is a local property, we can easily generalize the definition to arbitrary embedded surfaces.

Definition (Geodesic on smooth embedded surface)**.** Let $S \subseteq \mathbb{R}^3$ be an embedded surface. Let $\Gamma: [a, b] \to S$ be a smooth curve in S, and suppose there is a parametrization $\sigma: V \to U \subseteq S$ such that $\operatorname{im}\Gamma \subseteq U$. We let $\theta = \sigma^{-1}$ be the corresponding chart.

We define a new curve in V by

$$\gamma = \theta \circ \Gamma : [a, b] \to V.$$

Then we say Γ is a *geodesic* on S if and only if γ is a geodesic with respect to the induced Riemannian metric.

For a general $\Gamma: [a, b] \to V$, we say Γ is a *geodesic* if for each point $t_0 \in [a, b]$, there is a neighbourhood $\tilde{V}$ of t_0 such that $\operatorname{im}\Gamma|_{\tilde{V}}$ lies in the domain of some chart, and $\Gamma|_{\tilde{V}}$ is a geodesic in the previous sense.

Corollary. If a curve Γ minimizes the energy among all curves from $P = \Gamma(a)$ to $Q = \Gamma(b)$, then Γ is a geodesic.

Proof. For any a_1, a_2 such that $a \leq a_1 \leq b_1 \leq b$, we let $\Gamma_1 = \Gamma|_{[a_1, b_1]}$. Then Γ_1 also minimizes the energy between a_1 and b_1 for all curves between $\Gamma(a_1)$ and $\Gamma(b_1)$.

If we picked a_1, b_1 such that $\Gamma([a_1, b_1]) \subseteq U$ for some parametrized neighbourhood U, then Γ_1 is a geodesic by the previous proposition. Since the parametrized neighbourhoods cover S, at each point $t_0 \in [a, b]$, we can find a_1, b_1 such that $\Gamma([a_1, b_1]) \subseteq U$. So done. □

This is good, but we can do better. To do so, we need a lemma.

Lemma. Let $V \subseteq \mathbb{R}^2$ be an open set with a Riemannian metric, and let $P, Q \in V$. Consider C^∞ curves $\gamma: [a, b] \to V$ such that $\gamma(0) = P, \gamma(1) = Q$. Then such a γ will minimize the energy (and therefore is a geodesic) if and only if γ minimizes the length *and* has constant speed.

This means being a geodesic is *almost* the same as minimizing length. It's just that to be a geodesic, we have to parametrize it carefully.

Proof. Recall the Cauchy-Schwartz inequality for continuous functions $f, g \in C[0, 1]$, which says

$$\left(\int_0^1 f(x)g(x)\;\mathrm{d}x\right)^2 \leq \left(\int_0^1 f(x)^2\;\mathrm{d}x\right)\left(\int_0^1 g(x)^2\;\mathrm{d}x\right),$$

with equality iff $g = \lambda f$ for some $\lambda \in \mathbb{R}$, or $f = 0$, i.e. g and f are linearly dependent.

We now put $f = 1$ and $g = \|\dot{\gamma}\|$. Then Cauchy-Schwartz says

$$(\operatorname{length}\gamma)^2 \leq \operatorname{energy}(\gamma),$$

with equality if and only if $\dot{\gamma}$ is constant.

From this, we see that a curve of minimal energy must have constant speed. Then it follows that minimizing energy is the same as minimizing length if we move at constant speed. □

Is the converse true? Are all geodesics length minimizing? The answer is "almost". We have to be careful with our conditions in order for it to be true.

Proposition. A curve Γ is a geodesic iff and only if it minimizes the energy *locally*, and this happens if it minimizes the length locally and has constant speed.

Here minimizing a quantity locally means for every $t \in [a, b]$, there is some $\varepsilon > 0$ such that $\Gamma|_{[t-\varepsilon, t+\varepsilon]}$ minimizes the quantity.

We will not prove this. Local minimization is the best we can hope for, since the definition of a geodesic involves differentiation, and derivatives are local properties.

Proposition. In fact, the geodesic ODEs imply $\|\Gamma'(t)\|$ is constant.

We will also not prove this, but in the special case of the hyperbolic plane, we can check this directly. This is an exercise on the third example sheet.

A natural question to ask is that if we pick a point P and a tangent direction $\mathbf{a}$, can we find a geodesic through P whose tangent vector at P is $\mathbf{a}$?

In the geodesic equations, if we expand out the derivative, we can write the equation as

$$\begin{pmatrix} E & F \\ F & G \end{pmatrix}\begin{pmatrix} \ddot{\gamma}_1 \\ \ddot{\gamma}_2 \end{pmatrix} = \text{something}.$$

Since the Riemannian metric is positive definite, we can invert the matrix and get an equation of the form

$$\begin{pmatrix} \ddot{\gamma}_1 \\ \ddot{\gamma}_2 \end{pmatrix} = H(\gamma_1, \gamma_2, \dot{\gamma}_1, \dot{\gamma_2})$$

for some function H. From the general theory of ODE's in IB Analysis II, subject to some sensible conditions, given any $P = (u_0, v_0) \in V$ and $\mathbf{a} = (p_0, q_0) \in \mathbb{R}^2$, there is a *unique* geodesic curve $\gamma(t)$ defined for $|t| < \varepsilon$ with $\gamma(0) = P$ and

$\dot{\gamma}(0) = \mathbf{a}$. In other words, we can choose a point, and a direction, and then there is a geodesic going that way.

Note that we need the restriction that γ is defined only for $|t| < \varepsilon$ since we might run off to the boundary in finite time. So we need not be able to define it for all $t \in \mathbb{R}$.

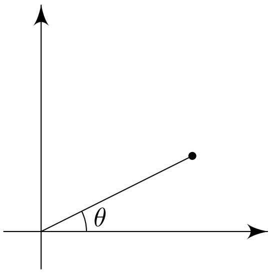
How is this result useful? We can use the uniqueness part to find geodesics. We can try to find some family of curves $\mathcal{C}$ that are length-minimizing. To prove that we have found *all* of them, we can show that given any point $P \in V$ and direction $\mathbf{a}$, there is some curve in $\mathcal{C}$ through P with direction $\mathbf{a}$.

Example. Consider the sphere S^2. Recall that arcs of great circles are length-minimizing, at least locally. So these are indeed geodesics. Are these all? We know for any $P \in S^2$ and any tangent direction, there exists a unique great circle through P in this direction. So there cannot be any other geodesics on S^2, by uniqueness.

Similarly, we find that hyperbolic line are precisely all the geodesics on a hyperbolic plane.

We have defined these geodesics as solutions of certain ODEs. It is possible to show that the solutions of these ODEs depend C^∞-smoothly on the initial conditions. We shall use this to construct around each point $P \in S$ in a surface *geodesic polar coordinates*. The idea is that to specify a point near P, we can just say "go in direction θ, and then move along the corresponding geodesic for time r".

We can make this (slightly) more precise, and provide a quick sketch of how we can do this formally. We let $\psi : U \to V$ be some chart with $P \in U \subseteq S$. We wlog $\psi(P) = 0 \in V \subseteq \mathbb{R}^2$. We denote by θ the polar angle (coordinate), defined on $V \setminus \{0\}$.



Then for any given θ, there is a unique geodesic $\gamma^\theta : (-\varepsilon, \varepsilon) \to V$ such that $\gamma^\theta(0) = 0$, and $\dot{\gamma}^\theta(0)$ is the unit vector in the θ direction.

We define

$$\sigma(r, \theta) = \gamma^\theta(r)$$

whenever this is defined. It is possible to check that σ is C^∞-smooth. While we would like to say that σ gives us a parametrization, this is not exactly true, since we cannot define θ continuously. Instead, for each θ_0, we define the region

$$W_{\theta_0} = \{(r, \theta) : 0 < r < \varepsilon, \theta_0 < \theta < \theta_0 + 2\pi\} \subseteq \mathbb{R}^2.$$

Writing V_0 for the image of W_{θ_0} under σ, the composition

$$W_{\theta_0} \xrightarrow{\sigma} V_0 \xrightarrow{\psi^{-1}} U_0 \subseteq S$$

is a valid parametrization. Thus $\sigma^{-1} \circ \psi$ is a valid chart.

The image (r, θ) of this chart are the *geodesic polar coordinates*. We have the following lemma:

Lemma (Gauss' lemma)**.** The geodesic circles $\{r = r_0\} \subseteq W$ are orthogonal to their radii, i.e. to γ^θ, and the Riemannian metric (first fundamental form) on W is

$$\mathrm{d}r^2 + G(r, \theta)\;\mathrm{d}\theta^2.$$

This is why we like geodesic polar coordinates. Using these, we can put the Riemannian metric into a very simple form.

Of course, this is just a sketch of what really happens, and there are many holes to fill in. For more details, go to IID Differential Geometry.

Definition (Atlas)**.** An *atlas* is a collection of charts covering the whole surface.

The collection of all geodesic polars about all points give us an example. Other interesting atlases are left as an exercise on example sheet 3.

5.3 Surfaces of revolution

So far, we do not have many examples of surfaces. We now describe a nice way of obtaining surfaces — we obtain a surface S by rotating a plane curve η around a line ℓ. We may wlog assume that coordinates a chosen so that ℓ is the z-axis, and η lies in the $x - z$ plane.

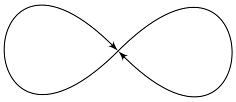
More precisely, we let $\eta : (a, b) \to \mathbb{R}^3$, and write

$$\eta(u) = (f(u), 0, g(u)).$$

Note that it is possible that $a = -\infty$ and/or $b = \infty$.

We require $\|\eta'(u)\| = 1$ for all u. This is sometimes known as *parametrization by arclength*. We also require $f(u) > 0$ for all $u > 0$, or else things won't make sense.

Finally, we require that η is a homeomorphism to its image. This is more than requiring η to be injective. This is to eliminate things like



Then S is the image of the following map:

$$\sigma(u, v) = (f(u) \cos v, f(u) \sin v, g(u))$$

for $a < u < b$ and $0 \leq v \leq 2\pi$. This is not exactly a parametrization, since it is not injective ($v = 0$ and $v = 2\pi$ give the same points). To rectify this, for each $\alpha \in \mathbb{R}$, we define

$$\sigma^\alpha : (a, b) \times (\alpha, \alpha + 2\pi) \to S,$$

given by the same formula, and this is a homeomorphism onto the image. The proof of this is left as an exercise for the reader.

Assuming this, we now show that this is indeed a parametrization. It is evidently smooth, since f and g both are. To show this is a parametrization, we need to show that the partial derivatives are linearly independent. We can compute the partial derivatives and show that they are non-zero. We have

$$\begin{aligned}\sigma_u &= (f'\cos v, f'\sin v, g')\\ \sigma_v &= (-f\sin v, f\cos v, 0).\end{aligned}$$

We then compute the cross product as

$$\sigma_u \times \sigma_v = (-fg'\cos v, -fg'\sin v, ff').$$

So we have

$$\|\sigma_u \times \sigma_v\| = f^2(g'^2 + f'^2) = f^2 \neq 0.$$

Thus every σ^α is a valid parametrization, and S is a valid embedded surface.

More generally, we can allow S to be covered by several families of parametrizations of type σ^α, i.e. we can consider more than one curve or more than one axis of rotation. This allows us to obtain, say, S^2 or the embedded torus (in the old sense, we cannot view S^2 as a surface of revolution in the obvious way, since we will be missing the poles).

Definition (Parallels)**.** On a surface of revolution, *parallels* are curves of the form

$$\gamma(t) = \sigma(u_0, t) \text{ for fixed } u_0.$$

Meridians are curves of the form

$$\gamma(t) = \sigma(t, v_0) \text{ for fixed } v_0.$$

These are generalizations of the notions of longitude and latitude (in some order) on Earth.

In a general surface of revolution, we can compute the first fundamental form with respect to σ as

$$\begin{aligned}E &= \|\sigma_u\|^2 = f'^2 + g'^2 = 1,\\ F &= \sigma_u \cdot \sigma_v = 0\\ G &= \|\sigma_v\|^2 = f^2.\end{aligned}$$

So its first fundamental form is also of the simple form, like the geodesic polar coordinates.

Putting these explicit expressions into the geodesic formula, we find that the geodesic equations are

$$\begin{aligned}\ddot{u} &= f\frac{\mathrm{d}f}{\mathrm{d}u}\dot{v}^2\\ \frac{\mathrm{d}}{\mathrm{d}t}(f^2\dot{v}) &= 0.\end{aligned}$$

Proposition. We assume $\|\dot{\gamma}\| = 1$, i.e. $\dot{u}^2 + f^2(u)\dot{v}^2 = 1$.

(i) Every unit speed meridians is a geodesic.

(ii) A (unit speed) parallel will be a geodesic if and only if

$$\frac{\mathrm{d}f}{\mathrm{d}u}(u_0) = 0,$$

i.e. u_0 is a critical point for f.

Proof.

(i) In a meridian, $v = v_0$ is constant. So the second equation holds. Also, we know $\|\dot{\gamma}\| = |\dot{u}| = 1$. So $\ddot{u} = 0$. So the first geodesic equation is satisfied.

(ii) Since $o = o_u$, we know $f(u_0)^2\dot{v}^2 = 1$. So

$$\dot{v} = \pm\frac{1}{f(u_0)}.$$

So the second equation holds. Since $\dot{v}$ and f are non-zero, the first equation is satisfied if and only if $\frac{\mathrm{d}f}{\mathrm{d}u} = 0$. □

5.4 Gaussian curvature

We will next consider the notion of curvature. Intuitively, Euclidean space is "flat", while the sphere is "curved". In this section, we will define a quantity known as the *curvature* that characterizes how curved a surface is.

The definition itself is not too intuitive. So what we will do is that we first study the curvature of curves, which is something we already know from, say, IA Vector Calculus. Afterwards, we will make an analogous definition for surfaces.

Definition (Curvature of curve)**.** We let $\eta : [0, \ell] \to R^2$ be a curve parametrized with unit speed, i.e. $\|\eta'\| = 1$. The *curvature* κ at the point $\eta(s)$ is determined by

$$\eta'' = \kappa\mathbf{n},$$

where $\mathbf{n}$ is the unit normal, chosen so that κ is non-negative.

If $f : [c, d] \to [0, \ell]$ is a smooth function and $f'(t) > 0$ for all t, then we can reparametrize our curve to get

$$\gamma(t) = \eta(f(t)).$$

We can then find

$$\dot{\gamma}(t) = \frac{\mathrm{d}f}{\mathrm{d}t}\eta'(f(t)).$$

So we have

$$\|\dot{\gamma}\|^2 = \left(\frac{\mathrm{d}f}{\mathrm{d}t}\right)^2.$$

We also have by definition

$$\eta''(f(t)) = \kappa\mathbf{n},$$

where κ is the curvature at $\gamma(t)$.

On the other hand, Taylor's theorem tells us

$$\begin{aligned}\gamma(t + \Delta t) - \gamma(t) &= \left(\frac{\mathrm{d}f}{\mathrm{d}t}\right)\eta'(f(t))\Delta t\\ &\quad + \frac{1}{2}\left[\left(\frac{\mathrm{d}^2 f}{\mathrm{d}t^2}\right)\eta'(f(t)) + \left(\frac{\mathrm{d}f}{\mathrm{d}t}\right)^2\eta''(f(t))\right] + \text{higher order terms.}\end{aligned}$$

Now we know by assumption that

$$\eta' \cdot \eta' = 1.$$

Differentiating thus give s

$$\eta' \cdot \eta'' = 0.$$

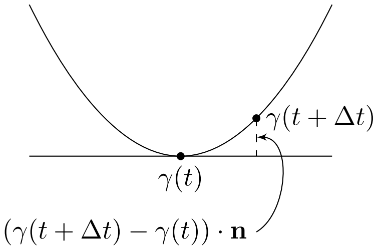
Hence we get

$$\eta' \cdot \mathbf{n} = 0.$$

We now take the dot product of the Taylor expansion with $\mathbf{n}$, killing off all the η' terms. Then we get

$$(\gamma(t + \Delta t) - \gamma(t)) \cdot \mathbf{n} = \frac{1}{2}\kappa \|\dot\gamma\|^2 (\Delta t)^2 + \cdots, \tag{$*$}$$

where κ is the curvature. This is the distance denoted below:



We can also compute

$$\|\gamma(t + \Delta t) - \gamma(t)\|^2 = \|\dot\gamma\|^2 (\Delta t)^2. \tag{$\dagger$}$$

So we find that $\frac{1}{2}\kappa$ is the ratio of the leading (quadratic) terms of $(*)$ and $(\dagger)$, and is independent of the choice of parametrization.

We now try to apply this thinking to embedded surfaces. We let $\sigma : V \to U \subseteq S$ be a parametrization of a surface S (with $V \subseteq \mathbb{R}^2$ open). We apply Taylor's theorem to σ to get

$$\begin{aligned}\sigma(u + \Delta u, v + \Delta v) - \sigma(u, v) &= \sigma_u \Delta u + \sigma_v \Delta v\\ &\quad + \frac{1}{2}(\sigma_{uu}(\Delta u^2) + 2\sigma_{uv}\Delta u \Delta v + \sigma_{vv}(\Delta v)^2) + \cdots.\end{aligned}$$

We now measure the deviation from the tangent plane, i.e.

$$(\sigma(u + \Delta u, v + \Delta v) - \sigma(u, v)) \cdot \mathbf{N} = \frac{1}{2}(L(\Delta u)^2 + 2M\Delta u \Delta v + N(\Delta v)^2) + \cdots,$$

where

$$\begin{aligned}L &= \sigma_{uu} \cdot \mathbf{N},\\ M &= \sigma_{uv} \cdot \mathbf{N},\\ N &= \sigma_{vv} \cdot \mathbf{N}.\end{aligned}$$

Note that $\mathbf{N}$ and N are different things. $\mathbf{N}$ is the unit normal, while N is the expression given above.

We can also compute

$$\|\sigma(u + \Delta u, v + \Delta v) - \sigma(u, v)\|^2 = E(\Delta u)^2 + 2F\Delta u \Delta v + G(\Delta v)^2 + \cdots.$$

We now define the second fundamental form as

Definition (Second fundamental form)**.** The *second fundamental form* on V with $\sigma : V \to U \subseteq S$ for S is

$$L\;\mathrm{d}u^2 + 2M\;\mathrm{d}u\;\mathrm{d}v + N\;\mathrm{d}v^2,$$

where

$$\begin{aligned} L &= \sigma_{uu} \cdot \mathbf{N} \\ M &= \sigma_{uv} \cdot \mathbf{N} \\ N &= \sigma_{vv} \cdot \mathbf{N}. \end{aligned}$$

Definition (Gaussian curvature)**.** The *Gaussian curvature* K of a surface of S at $P \in S$ is the ratio of the determinants of the two fundamental forms, i.e.

$$K = \frac{LN - M^2}{EG - F^2}.$$

This is valid since the first fundamental form is positive-definite and in particular has non-zero derivative.

We can imagine that K is a capital κ, but it looks just like a normal capital K.

Note that $K > 0$ means the second fundamental form is definite (i.e. either positive definite or negative definite). If $K < 0$, then the second fundamental form is indefinite. If $K = 0$, then the second fundamental form is semi-definite (but not definite).

Example. Consider the unit sphere $S^2 \subseteq \mathbb{R}^3$. This has $K > 0$ at each point. We can compute this directly, or we can, for the moment, pretend that $M = 0$. Then by symmetry, $N = M$. So $K > 0$.

On the other hand, we can imagine a Pringle crisp (also known as a hyperbolic paraboloid), and this has $K < 0$. More examples are left on the third example sheet. For example we will see that the embedded torus in $\mathbb{R}^3$ has points at which $K > 0$, some where $K < 0$, and others where $K = 0$.

It can be deduced, similar to the curves, that K is independent of parametrization.

Recall that around each point, we can get some nice coordinates where the first fundamental form looks simple. We might expect the second fundamental form to look simple as well. That is indeed true, but we need to do some preparation first.

Proposition. We let

$$\mathbf{N} = \frac{\sigma_u \times \sigma_v}{\|\sigma_u \times \sigma_v\|}$$

be our unit normal for a surface patch. Then at each point, we have

$$\begin{aligned} \mathbf{N}_u &= a\sigma_u + b\sigma_v, \\ \mathbf{N}_v &= c\sigma_u + d\sigma_v, \end{aligned}$$

where

$$-\begin{pmatrix} L & M \\ M & N \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} E & F \\ F & G \end{pmatrix}.$$

In particular,

$$K = ad - bc.$$

Proof. Note that

$$\mathbf{N} \cdot \mathbf{N} = 1.$$

Differentiating gives

$$\mathbf{N} \cdot \mathbf{N}_u = 0 = \mathbf{N} \cdot \mathbf{N}_v.$$

Since σ_u, σ_v and $\mathbf{N}$ for an orthogonal basis, at least there are some a, b, c, d such that

$$\begin{aligned} \mathbf{N}_u &= a\sigma_u + b\sigma_v \\ \mathbf{N}_v &= c\sigma_u + d\sigma_v. \end{aligned}$$

By definition of σ_u, we have

$$\mathbf{N} \cdot \sigma_u = 0.$$

So differentiating gives

$$\mathbf{N}_u \cdot \sigma_u + \mathbf{N} \cdot \sigma_{uu} = 0.$$

So we know

$$\mathbf{N}_u \cdot \sigma_u = -L.$$

Similarly, we find

$$\mathbf{N}_u = \sigma_v = -M = N_v \cdot \sigma_u, \quad \mathbf{N}_v \cdot \sigma_v = -N.$$

We dot our original definition of $\mathbf{N}_u, \mathbf{N}_v$ in terms of a, b, c, d with σ_u and σ_v to obtain

$$\begin{aligned} -L &= aE + bF & -M &= aF + bG \\ -M &= cE + dF & -N &= cF + dG. \end{aligned}$$

Taking determinants, we get the formula for the curvature. □

If we have nice coordinates on S, then we get a nice formula for the Gaussian curvature K.

Theorem. Suppose for a parametrization $\sigma : V \to U \subseteq S \subseteq \mathbb{R}^3$, the first fundamental form is given by

$$\mathrm{d}u^2 + G(u, v)\,\mathrm{d}v^2$$

for some $G \in C^\infty(V)$. Then the Gaussian curvature is given by

$$K = \frac{-(\sqrt{G})_{uu}}{\sqrt{G}}.$$

In particular, we do not need to compute the second fundamental form of the surface.

This is purely a technical result.

Proof. We set

$$\mathbf{e} = \sigma_u, \quad \mathbf{f} = \frac{\sigma_v}{\sqrt{G}}.$$

Then $\mathbf{e}$ and $\mathbf{f}$ are unit and orthogonal. We also let $\mathbf{N} = \mathbf{e} \times \mathbf{f}$ be a third unit vector orthogonal to $\mathbf{e}$ and $\mathbf{f}$ so that they form a basis of $\mathbb{R}^3$.

Using the notation of the previous proposition, we have

$$\begin{aligned}\mathbf{N}_u \times \mathbf{N}_v &= (a\sigma_u + b\sigma_v) \times (c\sigma_u + d\sigma_v)\\ &= (ad - bc)\sigma_u \times \sigma_v\\ &= K\sigma_u \times \sigma_v\\ &= K\sqrt{G}\mathbf{e} \times \mathbf{f}\\ &= K\sqrt{G}\mathbf{N}.\end{aligned}$$

Thus we know

$$\begin{aligned}K\sqrt{G} &= (\mathbf{N}_u \times \mathbf{N}_v) \cdot \mathbf{N}\\ &= (\mathbf{N}_u \times \mathbf{N}_v) \cdot (\mathbf{e} \times \mathbf{f})\\ &= (\mathbf{N}_u \cdot \mathbf{e})(\mathbf{N}_v \cdot \mathbf{f}) - (\mathbf{N}_u \cdot \mathbf{f})(\mathbf{N}_v \cdot \mathbf{e}).\end{aligned}$$

Since $\mathbf{N} \cdot \mathbf{e} = 0$, we know

$$N_u \cdot \mathbf{e} + \mathbf{N} \cdot \mathbf{e}_u = 0.$$

Hence to evaluate the expression above, it suffices to compute $\mathbf{N} \cdot \mathbf{e}_u$ instead of $\mathbf{N}_u \cdot \mathbf{e}$.

Since $\mathbf{e} \cdot \mathbf{e} = 1$, we know

$$\mathbf{e} \cdot \mathbf{e}_u = 0 = \mathbf{e} \cdot \mathbf{e}_v.$$

So we can write

$$\begin{aligned}\mathbf{e}_u &= \alpha\mathbf{f} + \lambda_1\mathbf{N}\\ \mathbf{e}_v &= \beta\mathbf{f} + \lambda_2\mathbf{N}.\end{aligned}$$

Similarly, we have

$$\begin{aligned}\mathbf{f}_u &= -\tilde{\alpha}\mathbf{e} + \mu_1\mathbf{N}\\ \mathbf{f}_v &= -\tilde{\beta}\mathbf{e} + \mu_2\mathbf{N}.\end{aligned}$$

Our objective now is to find the coefficients μ_i, λ_i, and then

$$K\sqrt{G} = \lambda_1\mu_2 - \lambda_2\mu_1.$$

Since we know $\mathbf{e} \cdot \mathbf{f} = 0$, differentiating gives

$$\begin{aligned}\mathbf{e}_u \cdot \mathbf{f} + \mathbf{e} \cdot \mathbf{f}_u &= 0\\ \mathbf{e}_v \cdot \mathbf{f} + \mathbf{e} \cdot \mathbf{f}_v &= 0.\end{aligned}$$

Thus we get

$$\tilde{\alpha} = \alpha, \quad \tilde{\beta} = \beta.$$

But we have

$$\alpha = \mathbf{e}_u \cdot \mathbf{f} = \sigma_{uu} \cdot \frac{\sigma_v}{\sqrt{G}} = \left((\sigma_u \cdot \sigma_v)_u - \frac{1}{2}(\sigma_u \cdot \sigma_u)_v\right)\frac{1}{\sqrt{G}} = 0,$$

since $\sigma_u \cdot \sigma_v = 0, \sigma_u \cdot \sigma_u = 1$. So α vanishes.

Also, we have

$$\beta = \mathbf{e}_v \cdot \mathbf{f} = \sigma_{uv} \cdot \frac{\sigma_v}{\sqrt{G}} = \frac{1}{2}\frac{G_u}{\sqrt{G}} = (\sqrt{G})_u.$$

Finally, we can use our equations again to find

$$\begin{aligned}\lambda_1\mu_1 - \lambda_2\mu_1 &= \mathbf{e}_u \cdot \mathbf{f}_v - \mathbf{e}_v \cdot \mathbf{f}_u\\ &= (\mathbf{e} \cdot \mathbf{f}_v)_u - (\mathbf{e} \cdot \mathbf{f}_u)_v\\ &= -\tilde{\beta}_u - (-\tilde{\alpha})_u\\ &= -(\sqrt{G})_{uu}.\end{aligned}$$

So we have

$$K\sqrt{G} = -(\sqrt{G})_{uu},$$

as required. Phew. □

Observe, for σ as in the previous theorem, K depends only on the first fundamental from, not on the second fundamental form. When Gauss discovered this, he was so impressed that he called it the *Theorema Egregium*, which means

Corollary (Theorema Egregium)**.** If S_1 and S_2 have locally isometric charts, then K is locally the same.

Proof. We know that this corollary is valid under the assumption of the previous theorem, i.e. the existence of a parametrization σ of the surface S such that the first fundamental form is

$$\mathrm{d}u^2 + G(u, v)\,\mathrm{d}v^2.$$

Suitable σ includes, for each point $P \in S$, the geodesic polars (ρ, θ). However, P itself is not in the chart, i.e. $P \not\in \sigma(U)$, and there is no guarantee that there will be some geodesic polar that covers P. To solve this problem, we notice that K is a C^∞ function of S, and in particular continuous. So we can determine the curvature at P as

$$K(P) = \lim_{\rho\to 0} K(\rho, \sigma).$$

So done.

Note also that every surface of revolution has such a suitable parametrization, as we have previously explicitly seen. □

6 Abstract smooth surfaces

While embedded surfaces are quite general surfaces, they do not include, say, the hyperbolic plane. We can generalize our notions by considering surfaces "without embedding in $\mathbb{R}^3$". These are known as abstract surfaces.

Definition (Abstract smooth surface)**.** An *abstract smooth surface* S is a metric space (or Hausdorff (and second-countable) topological space) equipped with homeomorphisms $\theta_i : \mathcal{U}_i \to V_i$, where $\mathcal{U}_i \subseteq S$ and $V_i \subseteq \mathbb{R}^2$ are open sets such that

(i) $S = \bigcup_i \mathcal{U}_i$

(ii) For any i, j, the transition map

$$\phi_{ij} = \theta_j \circ \theta_i^{-1} : \theta_j(\mathcal{U}_i \cap \mathcal{U}_j) \to \theta_i(\mathcal{U}_i \cap \mathcal{U}_j)$$

is a diffeomorphism. Note that $\theta_j(\mathcal{U}_i \cap \mathcal{U}_j)$ and $\theta_i(\mathcal{U}_i \cap \mathcal{U}_j)$ are open sets in $\mathbb{R}^2$. So it makes sense to talk about whether the function is a diffeomorphism.

Like for embedded surfaces, the maps θ_i are called *charts*, and the collection of θ_i's satisfying our conditions is an *atlas* etc.

Definition (Riemannian metric on abstract surface)**.** A *Riemannian metric* on an abstract surface is given by Riemannian metrics on each $V_i = \theta_i(\mathcal{U}_i)$ subject to the compatibility condition that for all i, j, the transition map ϕ_{ij} is an isometry, i.e.

$$\langle d\varphi_P(\mathbf{a}), \mathrm{d}\varphi_P(\mathbf{b})\rangle_{\varphi(P)} = \langle \mathbf{a}, \mathbf{b}\rangle_P$$

Note that on the left, we are computing the Riemannian metric on V_i, while on the left, we are computing it on V_j.

Then we can define lengths, areas, energies on an abstract surface S.

It is clear that every embedded surface is an abstract surface, by forgetting that it is embedded in $\mathbb{R}^3$.

Example. The three classical geometries are all abstract surfaces.

(i) The Euclidean space $\mathbb{R}^2$ with $\mathrm{d}x^2 + \mathrm{d}y^2$ is an abstract surface.

(ii) The sphere $S^2 \subseteq \mathbb{R}^2$, being an embedded surface, is an abstract surface with metric

$$\frac{4(\mathrm{d}x^2 + \mathrm{d}y^2)}{(1 + x^2 + y^2)^2}.$$

(iii) The hyperbolic disc $D \subseteq \mathbb{R}^2$ is an abstract surface with metric

$$\frac{4(\mathrm{d}x^2 + \mathrm{d}y^2)}{(1 - x^2 - y^2)^2}.$$

and this is isometric to the upper half plane H with metric

$$\frac{\mathrm{d}x^2 + \mathrm{d}y^2}{y^2}$$

Note that in the first and last example, it was sufficient to use just one chart to cover every point of the surface, but not for the sphere. Also, in the case of the hyperbolic plane, we can have many different charts, and they are compatible.

Finally, we notice that we really need the notion of abstract surface for the hyperbolic plane, since it cannot be realized as an embedded surface in $\mathbb{R}^3$. The proof is not obvious at all, and is a theorem of Hilbert.

One important thing we can do is to study the curvature of surfaces. Given a $P \in S$, the Riemannian metric (on a chart) around P determines a "reparametrization" by geodesics, similar to embedded surfaces. Then the metric takes the form

$$\mathrm{d}\rho^2 + G(\rho, \theta)\ \mathrm{d}\theta^2.$$

We then define the curvature as

$$K = \frac{-(\sqrt{G})_{\rho\rho}}{\sqrt{G}}.$$

Note that for embedded surfaces, we obtained this formula as a theorem. For abstract surfaces, we take this as a *definition.*

We can check how this works in some familiar examples.

Example.

(i) In $\mathbb{R}^2$, we use the usual polar coordinates (ρ, θ), and the metric becomes

$$\mathrm{d}\rho^2 + \rho^2\ \mathrm{d}\theta^2,$$

where $x = \rho\cos\theta$ and $y = \rho\sin\theta$. So the curvature is

$$\frac{-(\sqrt{G})_{\rho\rho}}{\sqrt{G}} = \frac{-(\rho)_{\rho\rho}}{\rho} = 0.$$

So the Euclidean space has zero curvature.

(ii) For the sphere S, we use the spherical coordinates, fixing the radius to be 1. So we specify each point by

$$\sigma(\rho, \theta) = (\sin\rho\cos\theta, \sin\rho\sin\theta, \cos\rho).$$

Note that ρ is not really the radius in spherical coordinates, but just one of the angle coordinates. We then have the metric

$$\mathrm{d}\rho^2 + \sin^2\rho\ \mathrm{d}\theta^2.$$

Then we get

$$\sqrt{G} = \sin\rho,$$

and $K = 1$.

(iii) For the hyperbolic plane, we use the disk model D, and we first express our original metric in polar coordinates of the Euclidean plane to get

$$\left(\frac{2}{1-r^2}\right)^2 (\mathrm{d}r^2 + r^2\ \mathrm{d}\theta^2).$$

This is not geodesic polar coordinates, since r is given by the Euclidean distance, not hyperbolic distance. We will need to put

$$\rho = 2\tanh^{-1} r, \quad \mathrm{d}\rho = \frac{2}{1-r^2}\,\mathrm{d}r.$$

Then we have

$$r = \tanh\frac{\rho}{2},$$

which gives

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

So we finally get

$$\sqrt{G} = \sinh\rho,$$

with

$$K = -1.$$

We see that the three classic geometries are characterized by having constant 0, 1 and -1 curvatures.

We are almost able to state the Gauss-Bonnet theorem. Before that, we need the notion of triangulations. We notice that our old definition makes sense for (compact) abstract surfaces S. So we just use the same definition. We then define the *Euler number* of an abstract surface as

$$e(S) = F - E + V,$$

as before. Assuming that the Euler number is independent of triangulations, we know that this is invariant under homeomorphisms.

Theorem (Gauss-Bonnet theorem)**.** If the sides of a triangle $ABC \subseteq S$ are geodesic segments, then

$$\int_{ABC} K\,\mathrm{d}A = (\alpha + \beta + \gamma) - \pi,$$

where α, β, γ are the angles of the triangle, and $\mathrm{d}A$ is the "area element" given by

$$\mathrm{d}A = \sqrt{EG - F^2}\,\mathrm{d}u\,\mathrm{d}v,$$

on each domain $\mathcal{U} \subseteq S$ of a chart, with E, F, G as in the respective first fundamental form.

Moreover, if S is a compact surface, then

$$\int_S K\,\mathrm{d}A = 2\pi e(S).$$

We will not prove this theorem, but we will make some remarks. Note that we can deduce the second part from the first part. The basic idea is to take a triangulation of S, and then use things like each edge belongs to two triangles and each triangle has three edges.

This is a genuine generalization of what we previously had for the sphere and hyperbolic plane, as one can easily see.

Using the Gauss-Bonnet theorem, we can define the curvature $K(P)$ for a point $P \in S$ alternatively by considering triangles containing P, and then taking the limit

$$\lim_{\text{area}\to 0} \frac{(\alpha+\beta+\gamma)-\pi}{\text{area}} = K(P).$$

Finally, we note how this relates to the problem of the parallel postulate we have mentioned previously. The parallel postulate, in some form, states that given a line and a point not on it, there is a unique line through the point and parallel to the line. This holds in Euclidean geometry, but not hyperbolic and spherical geometry.

It is a fact that this is equivalent to the axiom that the angles of a triangle sum to π. Thus, the Gauss-Bonnet theorem tells us the parallel postulate is captured by the fact that the curvature of the Euclidean plane is zero everywhere.